Project Number: K2033910

The California Evaluation Framework

Prepared for the California Public Utilities Commission and the Project Advisory Group

February 2004

TecMarket Works
And the project Team Members:

Megdal & Associates
Architectural Energy Corporation
RLW Analytics
Resource Insight
B & B Resources
Ken Keating and Associates
Ed Vine and Associates
American Council for an Energy Efficient Economy
Ralph Prahl and Associates
Innovologie

Project Number: K2033910

The California Evaluation Framework

DRAFT

Prepared for

Southern California Edison Company

To conduct a joint study supported by

Pacific Gas & Electric Company
San Diego Gas & Electric Company
Southern California Edison Company
Southern California Gas Company

As mandated by the

California Public Utilities Commission

February 2004

Submitted by

TecMarket Works Framework Team

Acknowledgements

The authors wish to acknowledge and express our appreciation to the many individuals who contributed to the development of the California Evaluation Framework. Without the support and assistance of these individuals this effort would not have been possible.

The Project Advisory Group provided general and topic-specific project guidance and reviews of all chapters of the Framework. These individuals and their affiliations are:

- Marian Brown, Project Manager, Advisory Group Chairperson, Southern California Edison
- Eli Kollman and Jay Luboff, Energy Division, California Public Utilities Commission
- Don Schultz and Christine Tam, Office of Ratepayer Advocates, California Public Utilities Commission
- Mike Messenger and Sylvia Bender, California Energy Commission
- Valerie Richardson, Chris Ann Dickerson, and Kenneth James, Pacific Gas and Electric Company
- Athena Besa and Rob Rubin, Sempra Energy

In addition to the oversight and support of the Project Advisory Group, several individuals provided ongoing comments, support and recommendations to the Framework as a whole and to individual chapters of the Framework. Over 90 individuals provided over 350 comments or recommendations to the Project Team. We wish to thank all those who provided comments. We have reviewed and discussed all of the comments received and have considered these comments in the development of the Framework. In many cases the comments were instrumental in guiding the development of the project's focus and the Framework chapters. Several comments required consultation with the Project Advisory Group to reach a consensus on the impact of the comment to the Framework. We wish to especially thank the following individuals who provided valuable comments that helped frame some of the discussions and improved the focus and structure of the Framework.

- Hayley Goodson, The Utility Reform Network
- Adrienne Kandel, California Energy Commission
- Patrick McCarthy and Bill Stiegelman, Aspen Systems
- Maureen McNamara, United States Environmental Protection Agency
- Robert Mowris, Robert Mowris and Associates
- Monica Nevius, Consortium for Energy Efficiency
- Dean Schiffman, Sempra Energy

- Phil Sissons, Sisson and Associates

We also want to thank the TecMarket Works Project Team that collaborated to design and develop the new California Evaluation Framework under the direction and oversight of the CPUC and the Project Advisory Group. These individuals include:

- Nick Hall, TecMarket Works
- Lori Megdal, Megdal & Associates
- Pete Jacobs and Stuart Waterbury, Architectural Energy Corporation
- Roger Wright, RLW Analytics
- Paul Chernick, Resource Insight Incorporated
- Ken Keating, Ken Keating and Associates
- Sharyn Barata, B&B Resources
- Ed Vine, Edward Vine and Associates
- Steve Nadel and Marty Kushler, American Council for an Energy Efficient Economy
- Ralph Prahl, Ralph Prahl and Associates
- John Reed, Innovologie

Table of Contents

ACKNOWLEDGEMENTS II
CHAPTER 1: INTRODUCTION AND PURPOSE OF FRAMEWORK 1
PURPOSE OF EVALUATION AND A STANDARD EVALUATION FRAMEWORK 1
PROJECT APPROACH 2
EVALUATION TYPES AND CONSIDERATIONS COVERED BY THE FRAMEWORK 3
Impact Evaluation, and Measurement and Verification Approaches (Chapters 6 and 7) 3
Process Evaluations (Chapter 8) 4
Information and Education Program Evaluation (Chapter 9) 4
Market Transformation Program Evaluation (Chapter 10) 5
Non-Energy Effects Evaluation (Chapter 11) 6
Uncertainty (Chapter 12) 6
Sampling (Chapter 13) 6
Evaluation and Cost-Effectiveness (Chapter 14) 7
EVALUATION RESEARCH NOT COVERED BY THE FRAMEWORK 7
THE USE OF EVALUATION ROADMAPS 9
CHAPTER 2: STUDY METHODOLOGY AND PUBLIC INPUT PROCESS 10
STUDY METHODOLOGY 10
Development of Work Plan and Literature Review 11
Development of Principles and Roadmap 16
Development of Draft and Final Project Reports 16
PUBLIC INPUT PROCESS 17
CHAPTER 3: USE OF THE EVALUATION FRAMEWORK BY DIFFERENT STAKEHOLDERS 18
CHAPTER 4: EVALUATION OVERVIEW AND ISSUES 23
PREFACE 23
CORE PURPOSE OF EVALUATION 23
THE TWO KEY FUNCTIONS OF EVALUATION – SUMMATIVE AND FORMATIVE EVALUATIONS 24
THE TWO BASIC TYPES OF EVALUATION 25
PROGRAM THEORY/LOGIC MODEL USE IN EVALUATION 26
Developing a Program Theory 29
Reasons for an Evaluator to Use and Refine the Program Theory 31
Program Theories are Especially Important to the Evaluation of Complex Programs 32
The Relationship Between Program Theory, Market Assessment and Market Theory 33
EVALUATION ETHICS 33
HISTORY OF ENERGY EFFICIENCY PROGRAM EVALUATION IN CALIFORNIA 34
1970s-1994: The Pre-Protocol Era 35
1994-1997: The Protocol Period 35
1998-2000: Energy Efficiency, Electric Industry Restructuring, and The CBEE Period 36
Post-2000: Transition Period and the Energy Crisis 37
POLICY USE OF EVALUATION RESULTS 38
Policy Based on Evaluation Results 38
Interpreting and Adjusting Results for Summative versus Formative Purposes 38
Making the Framework a Living Document 39
APPENDIX TO CHAPTER 4 41
AN EXAMPLE OF A PROGRAM THEORY 41
Program Activities 42
Program Assumptions 42

AN EXAMPLE OF A LOGIC MODEL 44
Complete Evaluation Ethics Document from the American Evaluation Association 45
CHAPTER 5: UMBRELLA FRAMEWORK FOR EVALUATION PLANNING, FUNDING AND RESEARCH PRIORITIZATION DECISIONS 49
INTRODUCTION 49
THE EVALUATION PLANNING CYCLE 50
COMPONENTS OF THE EVALUATION PLAN 52
REVIEWING AND APPROVING THE EVALUATION PLAN 53
EVALUATIONS TO FACILITATE AN INTEGRATED PLANNING PROCESS 54
Input to Policy Considerations and Decisions 55
Early Feedback to Program Implementers 55
Program Lifecycle Stage 55
Evaluation Data Time Lags 56
Portfolio Planning Requirements 56
Evaluation Planning Requirements and Regulatory Oversight 57
Program Design/Solicitation, Selection/Review and Implementation Preparation 57
The Value of Dispersed Timing for Evaluations 57
Contract Requirements for "Pay for Performance" Programs 57
Market Inertia 58
Timing Needs for Retention, Measure Life, and Technical Degradation Analyses 58
Oversight and Review 58
EXAMPLE OF AN INTEGRATED EVALUATION TIMING SYSTEM 58
Program Duration 61
Timing of Evaluation Activities 61
Multiple Offset Program Cycles 61
Impact Analysis for Final Contract Purposes 62
Program Planning Period 62
ROADMAPS AND THE EVALUATION PLANNING PROCESS 62
All Programs 63
Programs with Impact Goals 64
Information and Education Programs 65
Market Effects or Market Transformation Programs 66
Programs That Rely On Non-Energy Effects To Achieve Their Goals 67
OVERVIEW OF THE FRAMEWORK 68
DETERMINING EVALUATION PRIORITIES 69
RECOMMENDATIONS FOR SETTING EVALUATION BUDGETS 70
EVALUATION SPENDING PRIORITIES 72
CONSOLIDATING PROGRAMS FOR EVALUATION 73
APPENDIX TO CHAPTER 5 76
AN EXAMPLE OF AN INTEGRATED PLANNING CYCLE 76
Example 1 – Three Year Program Cycle with Integrated Planning, Solicitation and Evaluation Process. 76
Example 2 – Offset Program Cycles 83
CHAPTER 6: IMPACT EVALUATION 86
PREFACE 86
SKILLS REQUIRED FOR IMPACT EVALUATION 88
INTRODUCTION AND KEY ISSUES 89
Overall Impact Evaluation Objective 90
Units of Measure 92
Interactions Between Impact Evaluation and Planning Processes 94
Selecting Methods 94

THE BILLING ANALYSIS PATH 96
Simple Aggregate Pre-Post Comparisons *98*
Comparing Pre-Post Billing Data for Programs with Experimental Design *99*
Types of Billing Analysis Models *100*
Critical Billing Analysis Issues and Quality Control Efforts *108*
Examples from Energy-Efficiency Evaluation *113*
Summary Guidelines *115*
ENGINEERING ANALYSIS 115
Simple Engineering Models *118*
Building Energy Simulation Models *124*
NET-TO-GROSS REQUIREMENTS AND METHOD OPTIONS 128
Principles for Undertaking and Using Net-to-Gross Analysis *129*
Selecting an Analysis Method *130*
Survey-Based Methods *131*
Econometric-Based Methods *137*
Summary of NTGR Method Options and Roadmap Choices *140*
CHAPTER 7: MEASUREMENT AND VERIFICATION 141
PREFACE 141
M&V and Process Evaluation *141*
Skills Required for Measurement and Verification *142*
INTRODUCTION AND KEY ISSUES 142
The Role of the International Performance Measurement and Verification Protocol *142*
Statistical Context for M&V Studies *144*
UPPER LEVEL ROADMAP 144
Measurement and Verification Options *145*
Choosing the M&V Option *146*
M&V PLAN 147
Identify Goals and Objectives *148*
Specify Building Characteristics *148*
Specify Data Products and Project Output *150*
Specify M&V Option *151*
Specify Data Analysis Procedures and Algorithms *152*
Specify Field Monitoring Data Points *152*
Estimate Data Product Accuracy *152*
Specify Verification and Quality Assurance Procedures *152*
Specify Recording and Data Exchange Formats *153*
MEASURE INSTALLATION VERIFICATION 153
OPTION A 153
Data Resources *156*
Uncertainty Analysis *157*
Option A Example *157*
OPTION B 158
Option B Example: *160*
OPTION C 161
Billing Data Collection *163*
Comparison Models *163*
Adjustment factors *163*
Analysis Techniques *163*
Model Accuracy Criteria *167*
Baseline Model Selection Criteria *167*
OPTION D 168
Calibration Accuracy *172*
M&V ACCURACY REQUIREMENTS 173

METERING AND MONITORING 174
Examples *177*
Monitoring Duration and Schedule *178*
Instrumentation *180*
Measure Sampling *185*
Sensor Placement *186*
Measurement Uncertainty *187*
QUALITY CONTROL 187
DATA WAREHOUSE 187
Data Resources *189*
Data Review, Validation, and Analysis *190*
Data Formatting *190*
Database Update *190*
Data Access *190*
APPENDIX TO CHAPTER 7: UNCERTAINTY CALCULATION EXAMPLE . 191
CHAPTER 8: PROCESS EVALUATION 196
PREFACE 196
SKILLS REQUIRED FOR PROCESS EVALUATIONS 197
INTRODUCTION AND KEY ISSUES 198
Definition *198*
Why These Two Goals are in the Definition *198*
The Need for Process Evaluations *201*
Process Evaluations are In-Depth Studies *203*
Evaluation Timing *203*
Process Evaluation Activities *206*
Access to Program Materials, Resources and Personnel *210*
MAKING DECISIONS AND SELECTING METHODS: THE PROCESS EVALUATION ROADMAP 211
Steps in the Process Evaluation Roadmap *212*
PROCESS EVALUATION BUDGETS 219
CHAPTER 9: INFORMATION AND EDUCATION PROGRAM EVALUATION 221
PREFACE 221
INTRODUCTION AND KEY ISSUES 222
SKILLS REQUIRED FOR INFORMATION/EDUCATIONAL EVALUATIONS 225
IDENTIFYING THE ISSUES 226
MAKING DECISIONS AND SELECTING METHODS 226
Program Evaluation Design, Tools and Examples *230*
STEPS IN DECIDING IF AN INFORMATION/EDUCATION EFFECTS EVALUATION IS REQUIRED 235
CHAPTER 10: MARKET TRANSFORMATION PROGRAM EVALUATION. 239
PREFACE 239
SKILLS REQUIRED FOR MARKET EFFECTS EVALUATION 240
INTRODUCTION AND KEY ISSUES 243
MARKET IDENTIFICATION AND EVALUATION CONSOLIDATION 244
ASSESSING AND USING PROGRAM THEORY/LOGIC MODELS FOR MARKET EVALUATION 245
Market Characterization and Baseline Measurement *248*
Types of Market Indicators to be Studied *250*
Orienting Theoretical Perspectives *251*

Baseline Measurement: Some Practical Considerations ... *252*
MARKET INDICATORS AND MEASURING MARKET PROGRESS ... 253
The Need to Tell a Story ... *254*
The Need for Targeted Studies in the Early Years ... *254*
The Usefulness of a Consistent Framework and Institutional Process ... *255*
Dealing with Large Numbers of Market Indicators ... *256*
Don't Forget about Gross Unit Savings ... *256*
ASSESS SUSTAINABILITY AND PLAN TO MEASURE LONG-TERM MARKET EFFECTS AND POST-PROGRAM SUSTAINABILITY ... 260
CHAPTER 11: NON-ENERGY EFFECTS EVALUATION ... 263
PREFACE ... 263
INTRODUCTION AND KEY ISSUES ... 263
Definition ... *264*
Why Non-Energy Effects are Included in the Framework ... *265*
Evaluation Timing ... *266*
Skills Needed to Conduct Non-Energy Effects Research ... *267*
Evaluation Planning and Approach Tools ... *268*
RESEARCH METHODS AND APPROACHES ... 269
Partial Listing of Non-Energy Effects Presented in the Literature ... *271*
MAKING DECISIONS AND SELECTING METHODS ... 273
Steps in the Non-Energy Effects Evaluation Roadmap ... *273*
CHAPTER 12: UNCERTAINTY ... 280
INTRODUCTION AND KEY ISSUES ... 280
Skills Required for the Uncertainty Analysis Tasks ... *280*
BIAS AND STATISTICAL PRECISION ... 281
Basic Definitions ... *282*
How Bias Arises ... *285*
Special Techniques to Assess Bias ... *287*
Reporting Potential Bias ... *290*
Reducing the Standard Deviation of an Estimator ... *290*
INTEGRATING THE RESULTS FROM MULTIPLE EVALUATION STUDIES ... 291
Estimating the Total Savings of a Portfolio of Programs ... *292*
Contrasting the Results of Two Independent Studies ... *293*
Pooling Two Statistically Independent Estimators of the Same Parameter ... *294*
Chaining the Results of Two Evaluation Studies ... *296*
Converting a P-Value to a Relative Precision or Error Bound ... *297*
ALLOCATION OF RESOURCES TO EVALUATION ... 298
Bayesian Decision Theory and Power Analysis ... *299*
Propagation of Uncertainty ... *300*
The Fixed Relative Precision Criterion ... *301*
Optimal Allocation for the Overall Savings of the Portfolio ... *303*
CHAPTER 13: SAMPLING ... 307
INTRODUCTION AND KEY ISSUES ... 307
Skills Required for Sampling ... *308*
Key Issues ... *308*
General Steps in a Study ... *309*
Population Database ... *310*
SIMPLE RANDOM SAMPLING ... 310
Basic Definitions ... *311*
Sampling Distribution of the Sample Mean ... *312*
Expected Statistical Precision ... *313*
Choice of Sample Size ... *314*
Statistical Analysis ... *316*

Applicability to Impact Evaluation ... *317*
STRATIFIED RATIO ESTIMATION ... 319
Goals of the Section and Basic Definitions ... *321*
The Ratio Model ... *324*
Sampling Distributions ... *326*
Expected Statistical Precision and Choice of Sample Size ... *326*
Assessing the Error Ratio without a Prior Sample ... *327*
Estimating the Error Ratio from a Sample ... *328*
Model-Based Stratification ... *328*
The Expected Statistical Precision for Any Sample Design ... *329*
Applicability to Impact Evaluation ... *330*
A SAMPLING ROADMAP ... 332
The Population Database ... *332*
Sample Design ... *334*
Sample Selection ... *340*
Data Collection and M&V ... *342*
Statistical Analysis ... *344*
Reporting Guidelines ... *351*
APPENDIX TO CHAPTER 13 ... 353
EXAMPLE AND USE OF STRATIFIED RATIO ESTIMATION FOR SAMPLING ... 353
Choose the Sampling Approach ... *356*
Sample Selection ... *362*
Statistical Analysis ... *367*
Analysis of Potential Selection Bias ... *373*
Other Results ... *375*
CHAPTER 14: EVALUATION AND COST-EFFECTIVENESS ... 377
PREFACE ... 377
SKILLS NEEDED ... 378
COST-BENEFIT ANALYSIS FOR EVALUATORS ... 378
FACTORS INFLUENCING PROGRAM COST-EFFECTIVENESS ... 379
TREATMENT OF SPECIAL CATEGORIES OF COSTS ... 381
Participant Supported Upgrades ... *381*
Start-Up and Future Improvement Costs ... *381*
Evaluation Costs ... *382*
UNDERSTANDING TEMPORAL AND GEOGRAPHICAL VARIATIONS IN AVOIDED COSTS ... 383
PROVIDING USEFUL AVOIDED-COST ESTIMATION TO EVALUATORS ... 385
CHAPTER 15: OVERARCHING EVALUATION STUDIES ... 387
MEASURE SATURATION STUDIES ... 387
ENERGY-SAVINGS POTENTIAL STUDIES ... 388
PERSISTENCE AND RETENTION STUDIES ... 390
PORTFOLIO ANALYSES ... 392
BEST PRACTICE STUDIES ... 394
MARKET ANALYSES ... 394
Market Structure and Operations Studies ... *394*
Market Share Tracking Studies ... *395*
UPDATING OF KEY PARAMETERS ... 395
EVALUATION METHODOLOGY DEVELOPMENT ... 396
MANAGING AND FUNDING OVERARCHING STUDIES ... 396

Chapter 1: Introduction and Purpose of Framework

The California Evaluation Framework (Framework) provides a consistent, systemized, cyclic approach for planning and conducting evaluations of California's energy efficiency and resource acquisition programs. This document presents that Framework and provides valuable information concerning when evaluations should be conducted, the types of evaluation that can be conducted and a discussion of approaches for conducting those studies. The intended audience for various sections includes policy staff, program portfolio managers, program planners and implementers, evaluators, and other stakeholders. (Chapter 3 provides a review of the potential users of this document and guidance about which chapters and at what level of review they would be most beneficial.)

Purpose of Evaluation and a Standard Evaluation Framework

There are two primary purposes for conducting evaluations of energy programs in California. These are: 1) to reliably document program effects, and 2) to improve program designs and operations to be more cost-effective at obtaining energy resources. All program evaluation efforts associated with California's energy programs fall under one or both of these overall purposes for conducting evaluations.

To attain these goals and to allow policy makers to more completely understand the potential effects from programs and program portfolios, the California Public Utilities Commission (CPUC) mandated the construction of a statewide evaluation framework. This Framework provides a rigorous systems approach to conducting evaluations so that all programs are able to document their effects and be compared to other programs and supply options.

In order to initiate the construction of the Framework, the Project Advisory Group developed a request for proposals that was approved by the CPUC. Through a competitive bidding process, the group selected the TecMarket Works team to develop the Framework within a process allowing extensive public input. This document presents the results of that effort. The Framework provides important information on what types of program evaluations are sought and provides guidance on methods and approaches that can lead to high quality evaluation studies. The Framework is not a step-by-step evaluation procedures manual, but it is an evaluation guidance document that contains information relating to the focus and implementation of energy program evaluation studies in California.

There are several CPUC evaluation goals incorporated into this project that served to focus the efforts of the CPUC staff, the Project Advisory Group and the TecMarket Works team. These goals serve to provide an evaluation framework that when implemented in California will:

1. Provide reliable evaluation results to support energy policy and supply decisions,

2. Allow programs to be equably compared according to their energy impacts,
3. Help understand and verify program energy and peak savings,
4. Help identify and quantify market and non-energy effects,
5. Provide information needed to estimate program cost-effectiveness, and
6. Provide recommendations for program changes that help improve cost-effectiveness.

In addition to accomplishing the above high-level goals, the Framework was designed to accomplish or support the following more specific objectives:

1. Increase the level of reliability of program savings impact estimates for use in resource planning forums where the uncertainty of these estimates needs to be compared against the uncertainty of other key components of the resource plan.
2. Increase the quality of feedback to program administrators from evaluation projects to both improve program designs and increase the net savings from their programs.
3. Provide guidance to program administrators on what types of evaluation are recommended and are likely to be most beneficial for documenting operations and objective accomplishments.
4. Provide guidance to program administrators on the methodological approaches and study focus needed to perform specific types of evaluations.
5. Provide a Framework with flexibility that allows for the use of alternative evaluation approaches, especially when they can be shown to provide as reliable results as the methods presented in the Framework.

This document accomplishes these objectives and describes the recommended efforts needed to effectively evaluate different types of energy programs. The document also describes the skill sets necessary to accomplish these efforts. This document does not make recommendations on which organizations should be responsible for developing evaluation plans or completing the evaluation studies. The procedures for implementing this Framework are to be determined by the California Measurement Advisory Council (CALMAC) and the CPUC. Additionally, it is premature to settle the evaluation procedures until the overall framework for administering programs is determined by the CPUC.

Project Approach

In constructing this Framework, the TecMarket Works team relied to a substantial degree on past California evaluation protocols developed in the early 1990's and modified over the intervening years, and on other seminal evaluation publications found in the literature. These documents are presented in the bibliography section of this report. Readers are encouraged to reference these documents, as they are a critical part of the evolution of the

energy program evaluation field. The project team then began working with the CPUC staff, the Project Advisory Group and interested stakeholders through a project development process involving significant issue discussions with the Project Advisory Group and then incorporating extensive public input as the draft framework was developed. (See Chapter 2 for further information on the project methodology and public input process.)

Evaluation Types and Considerations Covered by the Framework

Nine key components of the Framework are incorporated into the Framework's design for performing program evaluations (See Chapter 5: Umbrella Roadmap). These nine components consist of the types of evaluations covered by the Framework and key concepts and considerations that go into planning and conducting evaluations in California. These nine components are summarized below and detailed in the chapters specifically dealing with each component.

Impact Evaluation, and Measurement and Verification Approaches (Chapters 6 and 7)[1]

Impact evaluations focus on estimating the gross and net effects from the implementation of one or more energy efficiency programs. Most program impact projections contain ex-ante estimates of savings. These estimates are what the program is expected to save as a result of its implementation efforts. These estimates are used for program planning and contracting purposes and for prioritizing program funding choices.

The impact evaluation focuses on identifying and estimating the amount of energy and demand the program actually provides. Estimates of actual savings are ex-post savings; program savings that can be documented after the program has made the changes that are to produce the savings. Savings induced by the program are called "net" savings, as they are beyond or in addition to what would have occurred without the program. Ex-post net savings are the savings estimates as measured/verified as being achieved by the program.

In approving programs, the CPUC bases its decisions on ex-ante savings, the savings that are expected to be delivered by the program. What the state receives from the program is ex-post net savings, the estimated energy savings that are actually caused by the expenditure of the program dollars. These savings may change over time. Ex-post net savings, documented via an impact evaluation, can be in a range of effects that are from substantially more than the projected ex-ante savings to substantially less, or somewhere

[1] Impact evaluations and their supporting measurement and verification (M&V) activities are discussed in two chapters of the Framework. The decision to place these related efforts within two chapters (rather than one chapter with a M&V sub-chapter) is because impact evaluations can be (in some limited cases) conducted without supportive M&V efforts. In addition, the authors agreed that these topics are substantially different and are important enough that each of these topics needs to be treated as a separate chapter allowing readers to focus on one topic. They are discussed together here to indicate their importance in the impact estimation process.

in between. Over the years, impact evaluations have helped to refine and define ex-ante savings estimates so that today the difference between ex-ante savings and ex-post savings is expected to be small, unless there is a difference in the markets being served by a program, the venders offering the program, the technologies promoted by a program, or other changes that can be expected to effect achieved savings. The impact evaluation documents in a rigorous estimation process the energy and demand impacts that the program provides. The Impact Evaluation chapter discusses the evaluation approaches suggested for program evaluators to document the amount of energy saved via the program's efforts. However, other methods and approaches can be used if they provide the same level of expected accuracy as those methods presented in the Framework.

The impact evaluation will often employ metering, monitoring and verification tools to help accurately estimate the ex-post program savings. These efforts are typically referred to as "M&V," meaning either *Measurement and Verification* or *Monitoring and Verification,* depending on the publications or reference used. The M&V approaches typically used in impact evaluation are discussed in the M&V chapter of this document (Chapter 7), but in summary, are some form of field measurements taken to help identify how much energy is used before the program actions are taken, how much energy is being used after the actions are taken, the use conditions associated with an installed technology, or a change in behaviors that is to produce the energy savings.

Process Evaluations (Chapter 8)

The process evaluation is a systematic assessment of an energy efficiency program for the purposes of documenting program operations at the time of the examination and identifying improvements that can be made to increase the program's efficiency or effectiveness for acquiring energy resources. In addition, a process evaluation can also help increase the effectiveness of other programs by providing other program planners and administrators with the evaluation results. These planners can then review the process evaluation results to determine if their programs can benefit from the evaluation's findings and recommendations.

The Process Evaluation chapter discusses the intent and focus of the process evaluation and the skill levels needed to conduct these evaluations. It also identifies the variety of investigative issues associated with the process evaluation and the tools typically used in these studies. Additionally, the chapter discusses the need for process evaluations to be conducted in time for programs to benefit from the evaluation findings, and the need to establish early information feedback systems with the program administrators to allow for evaluation results to feed the program redesign process to obtain the maximum level of energy resources within the program delivery period. One of the best methods for identifying "best practices" is through the process evaluation.

Information and Education Program Evaluation (Chapter 9)

Information and education program evaluations focus on assessing the degree to which program goals are accomplished and estimating the effects of the program's activities on

their target markets. They can also serve as an information source for assessing the cost-effectiveness of the program. Evaluations of information and education programs in California typically have a different research goal than the evaluations of programs that have energy impact goals. Most of California's information and education programs do not have energy impact goals and are not expected to be cost-effective from an energy acquisition perspective. Instead these programs are designed to influence the ability of other programs to achieve their energy impact goals or they are focused on trying to influence the short-term or long-term decision processes associated with acquisition and use of energy-consuming technologies or behaviors. The evaluation of information and education programs within the Framework focuses on documenting the effects of the program at reaching their information transfer or educational goals.

Market Transformation Program Evaluation (Chapter 10)

The Market Transformation (MT) Program Evaluation chapter of the Framework focuses on the evaluation of program-induced market effects when the program being evaluated has a goal of making longer-term or lasting changes in the way a market operates. These evaluations examine changes within a market that are caused, at least in part, by the energy efficiency programs attempting to change that market. These evaluations are challenging, as markets are constantly in a state of change as new and competing technologies are offered or as other non-program market transformation efforts compete with the program's efforts.

Two other forms of market evaluations (market baseline studies and market operations studies) are also briefly discussed in this chapter, as their development and use may be required to support a MT program evaluation.

Market baseline studies focus on documenting the status of the market that the program is attempting to change, so that changes caused by the program can be compared to the baseline conditions. The Framework discusses the challenges associated with these studies and the need to identify full or partial causal effects to the observed change in order to know which changes should be considered as normal baseline changes.

A market operations evaluation is a study that documents how a market operates relative to technology flows and information exchange within the targeted market. A primary purpose of these studies is to guide program theories and designs with factual information about how a market operates. This allows the program design to be integrated into the market in a way that supports accomplishing program goals. These studies also help in developing recommendations for program changes to make the program more successful at reaching its goals. From this perspective, it is easy to see why a program operations evaluation may be part of or integrated into a program process evaluation. Like the process evaluation, the market operations evaluation has the ability to help all programs that have impact goals within the market being assessed.

Non-Energy Effects Evaluation (Chapter 11)

Non-energy effects evaluations look at the intended or unintended effects that occur in addition to the energy impacts associated with a program. The Public Goods Charge (PGC) funds in California are obtained from ratepayers to support energy efficiency investments in lieu of supply-side investments for energy. Consequently, it is CPUC policy (as expressed during by CPUC staff during this project) that calculations of program cost-effectiveness should be based only on the value of energy savings created by the program, and not on other benefits. Given this approach, non-energy effects evaluations conducted on programs funded through the Public Goods Charge are limited in scope. In general, non-energy effects evaluations can be conducted when the evaluation focuses on providing information that will help increase the energy impacts of a program or programs, or when the CPUC explicitly approves an evaluation in order to document specific non-energy effects that are of interest to the CPUC. This chapter identifies the conditions for conducting a non-energy effects evaluation and provides a discussion on the variety of types of non-energy effects that have been studied, and the methodologies that have been used.

Uncertainty (Chapter 12)

As a result of CPUC policies (expressed during this project) associated with this Framework, it is the explicit intent of this Framework that evaluations conducted on California energy efficiency programs be conducted in a way that provides reliable technology-specific, or in some cases, program-specific ex-post net energy impact findings.

The chapter describes how evaluations should assess and report the level of uncertainty and potential sources of bias associated with the evaluation findings and explain what actions are taken to mitigate these. Evaluation users should to be able to determine the reliability of the study results and to determine if the results can be used for supply decisions, for public program policy making, or for updating deemed energy factors. The information is also designed for use in a summative fashion, where the propagation of uncertainty could be calculated for a group of programs, for the PGC-funded portfolio, or the overall portfolio of energy efficiency programs.

Sampling (Chapter 13)

This chapter presents and discusses the relationship between sample size and sample selection methods and the ability to assess ex-post effects that fairly represent the impacts that a program has achieved. Methods associated with calculating sample sizes and for allocating samples for studies are outlined in this chapter. The chapter also provides information for the calculation of relative precision, factors needed to create efficient sample designs, and how to use these within an evaluation's analysis phase.

Evaluation and Cost-Effectiveness (Chapter 14)

This chapter is targeted to three audiences in the area where issues overlap between program evaluations, cost-effectiveness analysis, and their uses and interpretations. First, it helps evaluators see how the results from their evaluations will be used in cost-effectiveness analysis. Second, it demonstrates to program staff and administrators who calculate or use cost-effectiveness analysis how evaluation and cost-effectiveness work together. Third, it helps policy makers understand some of the key issues involved in using evaluation results to estimate cost-effectiveness, since these tests are often used to inform a policy decision about whether to continue to invest in a program.

The chapter does not establish methods for calculating avoided costs or for conducting cost-effectiveness tests. The California *Standard Practice Manual* and the *Avoided Cost and Cost Effectiveness Study* address these topics.[2, 3]

Evaluation Research Not Covered by the Framework

While the Framework covers a wide range of types of evaluation research, it is not designed to cover three types of tangential research that may be associated with designing, planning or conducting evaluations of energy programs. The Framework does not address the following types of research efforts:

- Program-specific market research or market operations evaluations for purposes other than conducting independent evaluation assessments;
- Research, development and deployment (RD&D) program evaluations;
- Low-income program evaluations; and
- Overarching evaluation studies.

A brief explanation of these exclusions is presented below.

Program-specific market research or market operations evaluations for purposes other than conducting independent evaluation assessments
It is often desirable for program administrators or others to conduct market research for the purposes of planning, designing, administering, or implementing one or more energy programs. When these studies are conducted for reasons other than to evaluate the performance of an energy program, the recommended approaches presented in the Framework do not apply. This is not meant to imply that this research is not important. It can be critical for the design of cost-effective programs or to optimize program design. As such, allocating program resources dollars for this type of research can be well-justified.

---

2 *Standard Practice Manual (SPM) Economic Analysis of Demand-Side Management Programs*. (* California State Governor's Office 2001)

3 *California Avoided Cost and Cost Effectiveness Study* currently being conducted.

Research, development and deployment (RD&D) program evaluations
This Framework does not cover evaluations of energy technology research, development and deployment programs. These programs typically focus on the development and commercialization of new energy products or services. They are not specifically designed to acquire near-term energy savings. Because of this, this Framework does not cover them. Excluding evaluation of RD&D efforts from this Framework is not meant to convey that RD&D programs or evaluating RD&D programs is not important to the resource acquisition process in California. They are an important part of the long-term chain of events leading to the ability of energy programs to acquire energy resources. California has a long history of providing energy efficiency program funding for emerging technology programs. These programs are designed to speed market acceptance of emerging technologies that have successfully reached the commercialization phase but are not widely used due to customers' lack of knowledge, concerns, or questions about them. As a result, these programs can be evaluated using approaches similar to those for conducting evaluations of information and education programs.

Low-income program evaluations
California's low-income program evaluations are also not covered by this Framework. In California these programs have their own evaluation framework and evaluation decision policies specific to these equity programs. The offering of low-income programs is often influenced more by policies based on the need for the program's services than on the amount of energy they save or the cost-effectiveness of the programs.

Overarching evaluation studies
There are a number of other types of evaluation studies that are not program-specific in their focus or in the researchable issues they address that are not covered by this Framework. These studies are discussed in Chapter 15 as there are significant contributors to the overall evaluation effort for efficiency program and can benefit multiple programs or program portfolios. The Framework does not attempt to guide the evaluation approaches of these types of evaluations. Evaluations that fall under the overarching classification include:

- Measure saturation studies,
- Energy-savings potential studies (technical, economic, achievable-market),
- Portfolio analyses (including "best practices" and "lessons learned" studies),
- Market and market operations analysis (beyond program level),
- Studies that update key parameters that influence multiple programs (e.g., measure lifetime, avoided costs), and
- Development of improved methodologies for evaluating programs.

However, when program-specific evaluations are consolidated, so that the evaluation deals with a group of programs, the Framework can guide and assist these evaluations.

The Use of Evaluation Roadmaps

Within each of the chapters dealing with a specific type of evaluation is an evaluation decision "roadmap." These roadmaps are designed to assist the program administrator with their evaluation planning process and the related program-specific evaluation decisions. The administrators of California's energy efficiency and resource acquisition programs are encouraged to use these roadmaps in their evaluation planning processes. The roadmaps consist of a set of decision trees or decision flow diagrams that walk through the process of determining if an evaluation is needed and what type of evaluations, methods, or decisions are expected.

Chapter 2: Study Methodology and Public Input Process

Study Methodology

The California Evaluation Framework (Framework) was developed to help program administrators plan, prioritize and conduct their program evaluation, measurement and verification (EM&V) activities. A key project goal incorporated into the development of the Framework was to produce a single "user-friendly" document that could provide evaluation guidance for a comprehensive set of program-specific evaluation efforts. This effort was to incorporate a wide range of public input into that process. The final Framework document is designed to serve as an evaluation roadmap, providing practical guidance that program administrators can follow in order to plan their evaluation efforts to meet the needs of the CPUC and other program or energy supply stakeholders.

The previous California framework study, *A Framework for Planning and Assessing Publicly Funded Energy Efficiency,* (hereafter referred to as the *2001 Framework Study*[4]) can be viewed as an authoritative text on market transformation programs and the evaluation of those programs, and should be considered a key supportive document to this Framework. The *2001 Framework Study* describes how market transformation programs can be evaluated and what tools can be used in those evaluations. It also provides a theoretical foundation for placing market interventions (whether from a market transformation or resource acquisition perspective) into energy efficiency markets.

However, this project is built upon an Advisory Group and California Public Utility Commission (CPUC) vision that was established early in the project, that program administrators need an evaluation road map that serves as a decision-guidance system for determining what research should be conducted for specific types of energy programs. This Framework also provides advice on how these evaluations should be conducted. The Framework is also designed as a reference manual for evaluators, providing guidance on planning, budgeting, and implementing program-specific evaluations. Additionally, the Framework can be used as an evaluation policy support document, providing policy makers with an information source from which evaluation policies and budgets can be guided. The Framework provides information to help guide the selection of evaluation approaches, including the measurement and verification approaches that must inform the impact estimation process. Likewise, the Framework reflects the needs of program administrators, designers and policy makers to have early feedback from the evaluation process to support the evolution of program services to be increasingly cost-effective. At the same time, the Framework recognizes that it is equally important that the EM&V efforts should themselves be cost-effective.

---

4 *A Framework for Planning and Assessing Publicly Funded Energy Efficiency* (* Sebold et al. 2001). Primary investigators included Drs.: Frederick D. Sebold and Alan Fields, RER; Lisa Skumatz, Skumatz Economic Research Associates, Inc.; Shel Feldman, Shel Feldman Management Consulting, Miriam Goldberg, Xenergy, Inc.; Ken Keating, Consultant; and Jane Peters, Research Into Action, Inc.

Development of Work Plan and Literature Review

Project Initiation, CALMAC Kick-off, and Work Plan Development

The first steps in the development of the new Framework consisted of a project initiation meeting and the development of a detailed Draft Work Plan. These efforts began with a project initiation teleconference with the Project Advisory Group. This first meeting began to set the stage for the development of the Framework and to identify the evaluation subjects to be addressed in the Framework. Following the initiation meeting the project team conducted a series of meetings in which the draft work plans were developed, reviewed and modified into the final project work plan.

To begin the Framework's public input process, a project kick-off public meeting was held in San Francisco shortly after the draft program plans were developed. An announcement for this meeting was publicly posted on the California Measurement Advisory Council (CALMAC) service list. The basic design of the project, the scope of the work, and the anticipated project efforts were presented at the public meeting and on the CALMAC web site. Public input on these presentations was received during the CALMAC meeting and via telephone and e-mail correspondence. These comments focused on the project's scope of work, issues to be addressed, the needs and potential uses of the project, and questions concerning the project's processes and the work to be performed. Over 150 comments were received during the first meeting. Minutes from this CALMAC meeting, along with the material used at the meeting, were developed, edited for completeness, and placed on the CALMAC website for public consultation.

Following this meeting the draft work plan was finalized based on the comments and feedback received from the Advisory Group and from the public. A final Work Plan was then submitted for review to the Project Advisory Group. Comments were incorporated to produce a final Work Plan that was used to guide the project.

The Project Advisory Group provided general and topic-specific project guidance and reviewed all chapters of the Framework as they were developed. The Advisory Group also provided consistent, ongoing guidance to the effort and provided valuable feedback as the chapters were developed. However, that does not mean that a consensus was achieved on all decisions regarding the Framework effort. Several rounds of Advisory Group and public input were provided in order for the scope and focus of the Framework to be finalized. Nevertheless, this document is the sole responsibility of the TecMarket Works Team that produced it. (Therefore, responsibility for the contents of this document needs to be borne most heavily by the leaders of this team.)

The Project Advisory Group members and their affiliations are:

- Marian Brown, Project Manager, Advisory Group Chairperson, Southern California Edison
- Eli Kollman and Jay Luboff, Energy Division, California Public Utilities Commission

- Don Schultz and Christine Tam, Office of Ratepayer Advocates, California Public Utilities Commission
- Mike Messenger and Sylvia Bender, California Energy Commission
- Valerie Richardson, Chris Ann Dickerson, and Kenneth James, Pacific Gas and Electric Company
- Athena Besa and Rob Rubin, Sempra Energy

The TecMarket Works Framework Team consists of:

- Nick Hall, TecMarket Works
- Lori Megdal, Megdal & Associates
- Pete Jacobs and Stuart Waterbury, Architectural Energy Corporation
- Roger Wright, RLW Analytics
- Paul Chernick, Resource Insight Incorporated
- Ken Keating, Ken Keating and Associates
- Sharyn Barata, B&B Resources
- Ed Vine, Ed Vine and Associates
- Steve Nadel and Marty Kushler, American Council for an Energy Efficient Economy
- Ralph Prahl, Ralph Prahl and Associates
- John Reed, Innovologie

The TecMarket Works Framework Team was organized into several expert teams based on specific evaluation type and development work areas. These teams are shown in Figure 2.1.

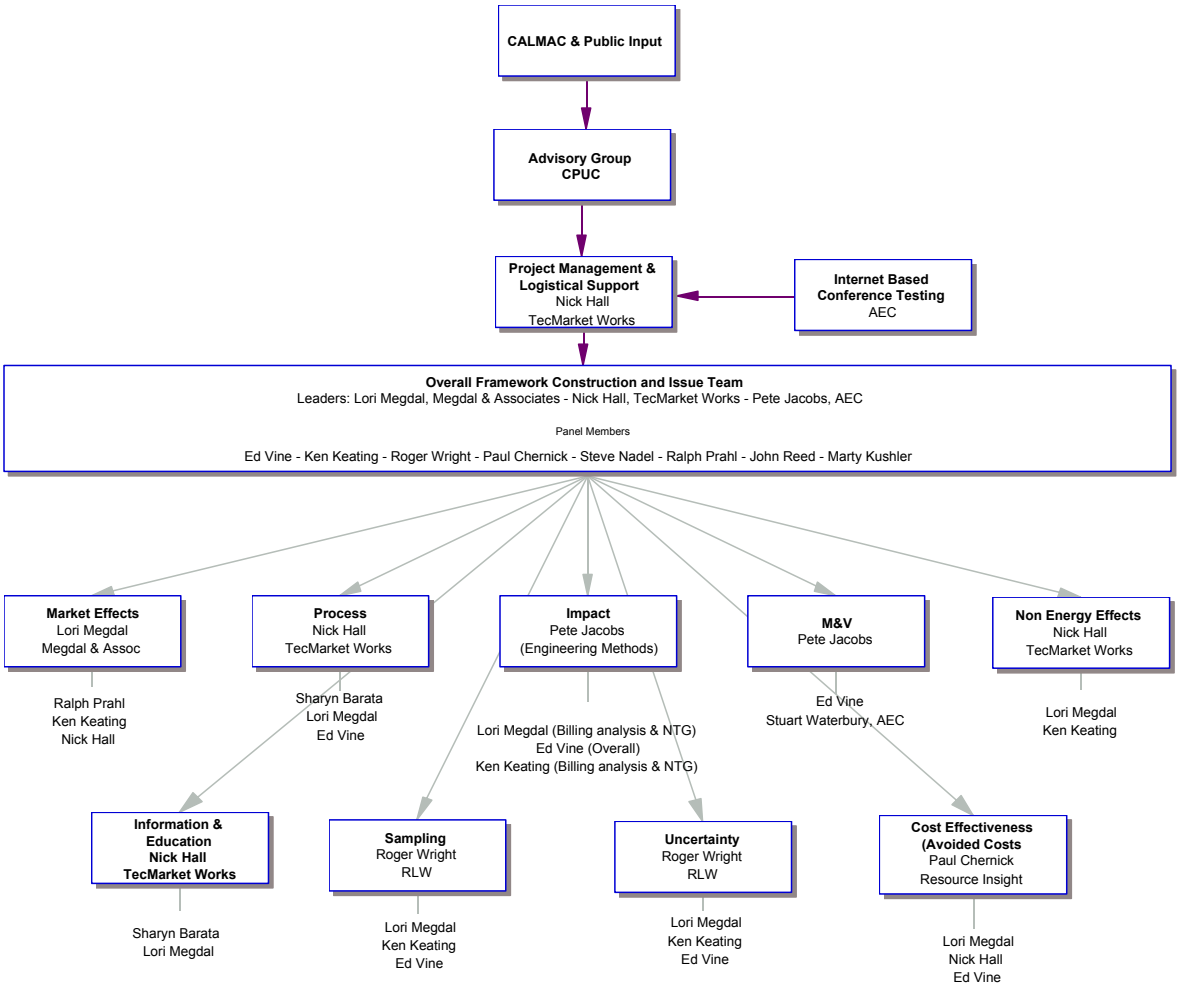

Figure 2.1: Project Team Assignments and Organization

Literature Review

One of the first project tasks was a literature review to support the development of the Framework. The primary purpose of the literature review was to not have the Framework re-invent the evaluation wheel in California, but to use currently available literature to guide the Framework development process.

The individual chapters of the Framework correspond to specific types of evaluation research and provide valuable information on determining which evaluations to conduct for the different types of energy efficiency and procurement programs offered in California. These chapters focus on the following type of research:

- Impact evaluation,
- Measurement and verification (M&V),
- Process Evaluation,
- Information and Education Program Evaluation,
- Market Transformation (MT) Program Evaluation, and
- Non-Energy Effects (NEE) Evaluation.

In addition to these evaluation-focused chapters there are two additional chapters that cut across all of these types of evaluation that are incorporated into the Framework; these are chapters on uncertainty in the evaluation process, and sampling methods and approaches.

The chapters on the specific types of evaluation provide an overview of the evaluation methods and issues to be addressed for each type of evaluation, some indications of what defines a high quality evaluation for each type of research effort, and the presentation of key references for additional technical background and to provide examples of use. One of the key documents that heavily informed certain chapters of this Framework is the prior California Measurement & Evaluation Protocols.[5] Other key reference documents include the Energy Efficiency Policy Manual[6] and the International Performance Measurement and Verification Protocol.[7]

Other "protocol" type documents used for general reference in the early stages of this project included the following publications:

- USEPA (1995). *Conservation Verification Protocols: A Guidance Document for Electric Utilities Affected by the Acid Rain Program*.
- * FEMP (2000). *Federal Energy Management Program (FEMP) M&V Guidelines: Measurement and Verification for Federal Energy Projects*. Federal Energy Management Program. September. Version 2.2, DOE/GO-102000-0960.
- * ASHRAE (2002). Measurement of Energy and Demand Savings, Guideline 14. American Society of Heating, Refrigeration and Air Conditioning Engineers: Atlanta, GA. Available at: www.ASHRAE.org
- Nexant and Lawrence Berkeley National Laboratory (2002). *Detailed Guidelines for FEMP M&V Option A*. Federal Energy Management Program.Available at: http://ateam.lbl.gov/mv
- AIS, SRC International (2001). *European Ex-post Evaluation Guidebook for DSM and EE Services Programmes*. International Energy Agency. April. Available at: http://dsm.iea.org/ Report is available under the library.
- Xenergy, ADM Associates, VACom Technologies and Partnership for Resource Conservation (2001). *2001 DEER (Database for Energy Efficiency Resources) Update Study*. California Energy Commission. Study ID 3001. Within searchable database at: http://www.calmac.org
- * Violette, Daniel (1995). *Evaluation, Verification, and Performance Measurement of Energy Efficiency Programs*. International Energy Agency. Available at: http://dsm.iea.org/. Report is available under the library.

---

5 *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs.* (* CADMAC 1999). Referenced as the Measurement and Evaluation Protocols. Available at: www.calmac.org.

6 *Energy Efficiency Policy Manual, Version 2*. August. (* CPUC 2003). Available at: http://www.cpuc.ca.gov.

7 *IPMVP Volume 1: Concepts and Options for Determining Energy and Water Savings*. (International Performance Measurement and Verification Protocol 2001). Available at: http://www.ipmvp.org/download.html

A focus on how the literature would be used for the Framework as a whole and for individual types of evaluation guided the selection of documents for the literature review. The criteria for selecting literature for review from the thousands of available publications centered upon providing support for:

- Defining types of evaluations needed for various decisions being made;
- Criteria for prioritizing between evaluation needs;
- Graduate textbooks and seminal works in fields related to energy efficiency program evaluation, to include:
  - Research design
  - Sampling design
  - Survey design
  - Statistics and econometrics
  - Evaluation studies
  - Program theory/logic work and evaluation
  - Energy engineering;
- Work that provided testing or advancements of new methods, tests, and specific examples in the following types of efficiency evaluation:
  - Energy impact and program effects evaluation
  - Measurement & verification
  - Process evaluation
  - Market assessment/progress measurement
  - Evaluating training efforts;
- Advancements and tests within energy efficiency evaluation in the areas of:
  - Uncertainty
  - Sampling;
- Other aspects of an overall evaluation research as were supportive in the creation of the integrated evaluation planning, evaluation, and program planning components within the umbrella roadmap.

All key team members, several Advisory Group members, and a number of public commenters provided recommendations for the literature to be reviewed for this effort. From these recommendations a literature review list was developed for possible use in the project. This list was reviewed by team members and then by the Project Advisory Group and finalized for the project effort. This list is provided in Appendix B, the Bibliography. Those items in the bibliography (and any other areas of the Framework with citations) that were part of the initial literature review list are indicated by asterisks in the front of the citation (*).

Document accessibility was an important criterion for whether a possible reference became a primary information source for the project. If documents were readily available for free, were made available through the CALMAC website, could be made available through the project website (http://www.tecmarket.net/CAFrame.htm), or were available at reasonable costs (such as various conference proceedings), they could serve as a primary project reference. Expensive documents or documents of limited availability, (e.g., EPRI documents) may be referenced in the Framework, but are not emphasized as

primary reference documents. Similarly, older documents that might be difficult to access are generally not primary references for the project, but may be referenced.

Development of Principles and Roadmap

The next step in the Framework construction process was to develop guiding principles for each type of evaluation. These were the primary paradigm parameters set for the Team and Advisory Group evaluation discussions. Frequent teleconferences with the Project Advisory Group were held to discuss evaluation issues, concepts, and the design and construction of the draft roadmaps. This process provided significant input and discussion between the Project Advisory Group and the key project leads to inform the development of the basics of the Framework and its roadmap characteristics. During this process a discussion packet of material was developed that contained these principles and draft roadmaps for each type of evaluation, the umbrella framework, and the overarching evaluation studies and issues. This packet was also distributed for public review to self-selected interest groups (see the public input process discussion below). The material was then revised as appropriate based upon the public input and used to guide the development of the roadmaps for public presentation at a set of CALMAC workshops.

An assembly of this material was then prepared for public review and comment during an open CALMAC meeting and placed on the CALMAC website prior to the meeting. During the public meeting the material was presented and discussed, and public comments and recommendations on the material were received by the project team. The CALMAC meetings were held in southern California (on September 3, 2003) and northern California (on September 4, 2003). Following these meetings, minutes were prepared and placed on the CALMAC website for public consultation.

Development of Draft and Final Project Reports

All of the material developed in these early efforts, the comments received at the CALMAC meetings, and the literature review provided the raw material from which the project teams developed the draft roadmaps and chapters for the Framework document. During the development of the Framework chapters a significant effort was made to ensure that references were incorporated into the chapter text to provide appropriate citations to support the chapter text but also to site reference material that could be used to learn more about the methods, tools, and issues being presented. This included citing references to academic literature upon which the foundations of the methods for quality evaluations are based, and the energy efficiency evaluation literature to provide examples of the use of the methods, issues, and tools discussed. Citations are provided at the end of each of the chapters so that each chapter can act as a stand-alone document. A complete bibliography is included as Appendix B.

An initial draft document was distributed to the Project Advisory Group and to the public interest group list in November 2003. Comments were received and reviewed by the TecMarket Works Framework Team in December 2003. From this review, a list of areas of disagreement, conflicting recommendations, and important direction issues was

compiled. This list, with an outline of the areas needing decisions, was provided to the Project Advisory Group. Lengthy teleconferences with the Project Advisory Group were held on December 23rd and 29th of 2003 to resolve these issues and provide final document guidance.

A complete set of revised chapters was prepared for a second round of review by the Project Advisory Group. These were sent out for review at the end of January 2004.

Comments from the Project Advisory Group on the second draft of the Framework document were used to create the final draft. This final draft document will be posted on the CALMAC website in mid February for public review. A notice of this posting with a request for public feedback will be sent to the CALMAC service list at the time of the posting. Public feedback will be accepted by e-mail and used to guide the development of the final Framework document.

Public Input Process

This project has one of the highest levels of public input experienced within an evaluation project. First, there was an open project kick-off meeting that discussed the details concerning possible subjects and issues to be addressed in the projects. Over 150 comments were received during this meeting and incorporated into a publicly available set of meeting presentation materials and meeting notes posted on the CALMAC website.

Then midway through the project two open CALMAC meetings were held. One was held in northern California and one held in southern California in early September 2003. These meetings presented the evaluation guiding principles that were proposed for each evaluation type, and the issues to be presented in the Framework document. Comments and discussion were received at both meetings. Publicly available meeting presentation materials and notes were prepared and posted on the CALMAC website.

Additionally, this project developed a unique public input process based upon self-selected subject interest group lists. To construct these lists an e-mail was sent to the CALMAC service list asking that anyone interested in reviewing the Framework's progress and draft chapters as they developed could be sent this material if they responded by e-mail asking to be placed on one or more of the chapter review lists. The evaluation principles and the draft Framework were sent to the interest group lists. In addition, the draft materials for the September CALMAC meeting were sent to this group. Then again, as the Framework draft chapters were developed in November, they were each e-mailed to members on the interest group list for that section of the Framework.

After multiple revisions to incorporate public comments, a final draft Framework was provided for public input. This document was placed on the CALMAC web site and announced for public review and comments. This public feedback process provided the final input opportunity for the project just prior to finalizing the Framework document.

Chapter 3: Use of the Evaluation Framework by Different Stakeholders

One of the early discussions between the California Public Utility Commission (CPUC) staff, the Framework's Advisory Group and the TecMarket Works project management team involved identifying for whom the Framework document should be written and the level of detail to be included in the various chapters within the Framework document. As a result of these discussions, this document is written so that readers with a limited background in the field of evaluation can understand the scope and application of the evaluation efforts and concepts presented. At the same time, the Framework is meant to be a reference document for evaluators and a tool to help design high quality evaluations. These two conditions required a balance in the material presented that attempts to meet the information needs of both evaluation and non-evaluation professionals.

The document also seeks to provide enough information and references so that evaluators will know what methods and levels of rigor are expected. To this end, the document includes a significant list of references guiding readers to additional information about the evaluation methods, tools, or issues being discussed. It is also realized that different chapters need to be produced at different levels of specificity. For example, the impact evaluation chapter is written at a higher level of detail than other chapters because specific impact evaluation and data handling practices need to be identified and discussed to indicate their strengths and weaknesses for use within the impact evaluation process.

In identifying the focus and scope of the Framework effort the CPUC, the Project Advisory Group and the project team stipulates that professionals specifically trained and practiced in the type of evaluation for which they are responsible, should conduct those evaluations. These parties accept this stipulation because it is imperative that people who are experts at assessing energy programs conduct evaluation studies. The Framework also stipulates that program evaluations will be conducted by firms, organizations or groups that are independent of the implementation administrator or contractor and that the evaluation teams will maintain an arms-length relationship with implementation administrators and contractors in order to help assure objective and reliable evaluation efforts. This document assumes that skilled professionals, who do not require that the Framework document be a detailed instruction manual, will conduct evaluations, and that the Framework document be designed to focus and shepherd the evaluation process. This focusing and shepherding is intended to guide evaluations toward producing rigorous, reliable energy program evaluation results that support fair comparisons across programs to the extent possible within the available evaluation resources.

The Framework document is written with the following types of individuals in mind.

- Regulatory staff and policy makers responsible for policy-associated goal setting and resulting high-level program or program portfolio decisions.
- Portfolio managers responsible for guiding or developing program portfolios.

- Power planning personnel responsible for power supply acquisition.
- Evaluation oversight managers and reviewers responsible for guiding or overseeing the evaluation efforts or interpreting evaluation results.
- Program administrators and implementers responsible for evaluation support efforts for their programs.
- Cost-effectiveness and avoided cost personnel responsible for identifying programs or program components that provide cost-effective energy supplies.
- Evaluation designers and managers responsible for designing or recommending evaluation methods and approaches that provide reliable results.
- Impact evaluators responsible for conducting studies which estimate the energy impacts (kW, kWh, therms, etc.) associated with a program, a group of programs or a program portfolio.
- Metering and monitoring evaluators responsible for designing or conducting field efforts to support the impact evaluation efforts.
- Process evaluators who are responsible for designing and conducting process evaluations to improve operations and program cost-effectiveness.
- Information and education evaluators responsible for designing and conducting evaluations of information and educational programs to assess program effects.
- Market transformation program and market effects evaluators responsible for designing and conducting evaluations of program markets, including market conditions, market operations, market effects or market baseline studies.
- Non-energy effects evaluators who are responsible for designing and conducting studies of the non-energy effects of programs or groups of programs or program portfolios.
- Statisticians and research data managers and analysts who are responsible for data handling and analysis efforts.
- Other stakeholders and interested parties that have an interest in the evaluation of energy programs or in the results of these studies.

Developing a single document for such a wide diversity of interested stakeholders is a challenge in itself. The Framework team attempts to meet this challenge by providing a document that does not over-simplify the evaluation efforts, and at the same time does not delve into the details of conducting program evaluations.

With these conditions in mind the Framework team realizes that specific parts of the Framework document will be of greater or less interest to the types of professionals identified above. The authors have developed two mechanisms to aid the reading and use of this document by different stakeholders. First, many of the Framework's chapters contain a Preface describing the contents of the chapter and indicating the various stakeholders that might find value from that chapter. Second, the authors have developed

a summary table (below) indicating which of the types of individuals identified above might find value in each of the Framework's chapters.

The following table presents the specific chapters within the Framework document that the authors suggest will be of interest to each of the categories of professionals identified above. Across the top of the table are the types of professionals identified above. Within each square of the table matrix is a ball indicating the level of interest that the authors suggest is appropriate for each of the types of professionals identified.

There are three levels of "interest categories" in this table consisting of a black bullet (●), a small bullet in a circle (⦿), and a white bullet (❍). The interest category populated by a black bullet (●) indicates that the authors suggest the chapter be considered important information for the types of professionals identified. Professionals in the categories that have a black bullet in the cell corresponding to the chapter on the left should, in the opinion of the authors, read these chapters carefully and have a working knowledge of the information within those chapters. Professionals with a small bullet in a circle (⦿) in the chapter cell may want to consider the material in the chapter as suggested reading, and be familiar with the chapter contents and implications for the evaluation effort. Finally, the professional categories that have a white bullet in the chapter cell (❍) might want to have at least an overview familiarity with the chapter contents.

By reviewing the table and examining the bullets within the matrices readers can focus their document review efforts on the chapters of the Framework that are most appropriate to their areas of responsibility.

Table 3.1: Suggested Framework Reading by Various Stakeholders

● Important information ⦿ Suggested reading ❍ Some familiarity	Regulatory Staff & Policy Makers	Portfolio Managers	Power Planning Personnel	Evaluation Oversight Mgrs & Reviewers	Program Administrators / Program Implementers	Cost-effectiveness and Avoided Cost Personnel	Evaluation Designers & Managers	Impact Evaluators*	Metering & Monitoring Evaluators	Process Evaluators	Information & Education Evaluators	Market Transformation and Market Effects Evaluators	Non-Energy Effects Evaluators	Statisticians & Research Data Managers	Other Stakeholders & Interested Parties
1. Introduction and Purpose of Framework	⦿	⦿	⦿	●	●	⦿	●	●	●	●	●	●	●	❍	⦿
2. Study Methodology and Description of Public Input Process	⦿	❍	❍	❍	⦿	❍	⦿	⦿	⦿	⦿	⦿	⦿	⦿	❍	❍
3. Use of the Framework by Different Stakeholders	●	⦿	⦿	●	●	⦿	●	●	●	●	●	●	●	⦿	⦿
4. Evaluation Overview and Issues	●	●	❍	●	●	●	●	●	●	●	●	●	●	⦿	⦿
5. Umbrella Roadmap for Evaluation Planning and Prioritization Decisions	●	⦿	⦿	●	●	●	●	●	●	●	●	●	●	●	⦿
6. Impact Evaluation Roadmap	⦿	⦿	⦿	●	●	⦿	●	●	⦿	❍			❍	⦿	❍
7. Measurement and Verification Roadmap	⦿	⦿	⦿	●	●	❍	●	●	●	❍				⦿	❍
8. Process Evaluation Roadmap	⦿	❍		●	●	❍	●	⦿	⦿	●	●	●	⦿	⦿	❍
9. Information / Education Roadmap	⦿	⦿		●	●	❍	●	❍		❍	●	❍	❍	⦿	❍

Table 3.1: Continued

● Important information ⊙ Suggested reading ❍ Some familiarity	Regulatory Staff & Policy Makers	Portfolio Managers	Power Planning Personnel	Evaluation Oversight Mgrs & Reviewers	Program Administrators / Program Implementers	Cost-effectiveness and Avoided Cost Personnel	Evaluation Designers & Managers	Impact Evaluators*	Metering & Monitoring Evaluators	Process Evaluators	Information & Education Evaluators	Market and Market Effects Evaluators	Non-Energy Effects Evaluators	Statisticians & Research Data Managers	Other Stakeholders & Interested Parties
10. Market Transformation Program Evaluation Roadmap	⊙	⊙	⊙	●	●		●	⊙	❍	⊙	⊙	●	⊙	⊙	❍
11. Non-Energy Effects Evaluation Roadmap	⊙	❍	❍	●	●	⊙	●	⊙	❍	⊙	⊙	●	●	❍	❍
12. Uncertainty	●	●	●	●	⊙	⊙	●	●	●	●	●	●	●	●	⊙
13. Sampling	⊙	❍	❍	●	⊙		●	●	●	●	●	●	●	●	❍
14. Evaluation and Cost-Effectiveness	●	●	⊙	●	●	●	●	●	●	●	●	●	●	●	⊙
15. Overarching Evaluation Studies	●	●	⊙	●	●	⊙	⊙	⊙	⊙	⊙	⊙	⊙	⊙	❍	❍
Appendix A: Definition of terms	⊙	⊙	⊙	●	●	●	●	●	●	●	●	●	●	●	⊙
Appendix B: Bibliography	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍	❍

** Sections of the Impact Chapter may contain information that the authors consider critical knowledge for engineering and statistical billing evaluators, depending upon the analysis method selected.*

Chapter 4: Evaluation Overview and Issues

Preface

This chapter of the Framework presents and discusses an overview of energy program evaluation, the different types of evaluation, the history of energy efficiency program evaluation in California, the use of program theory within evaluation, evaluation ethics, and the role of evaluation in the formation of energy policy. This chapter serves as a foundation for understanding the Framework and provides information for the development of evaluation policy in California.

Due to the groundwork this chapter provides and the importance of the topics covered with regard to the development and implementation of a statewide evaluation Framework, a wide range of evaluation stakeholders, can benefit from this chapter, including:

- Regulatory Staff & Policy Makers;
- Evaluation Oversight Managers & Reviewers;
- Program Administrators / Program Implementers;
- Cost-effectiveness and Avoided Cost Personnel;
- Evaluation Designers & Managers;
- Evaluators for all types of evaluation - impact, metering and monitoring, process, information and education programs, market transformation and market effects, and non-energy effects;
- Statisticians & Research Data Managers; and
- Portfolio Manager(s).

Core Purpose of Evaluation

When distilled to its most basic level, the essential over-arching purpose of evaluation is to help ensure that good decisions are made regarding the investment of energy program resources by providing rigorous, independent evaluation studies and study results. While the evaluation purpose is pursued through many different types of activities and methodologies by different parties and in different timeframes across the life cycle of a program, it is useful to keep this over-arching purpose in mind as a unifying theme for why evaluations are funded and conducted.

One of the primary ways in which evaluations provide information for making good decisions is testing the implicit and explicit assumptions within the program's program theory. It may also be important for the evaluation to develop and test alternative theories or hypotheses that may compete with or supplement the program theory. For

example, important conditions to be tested by the impact evaluation are the energy and demand savings per measure, per customer, and for the program overall and the reason for those savings. The process evaluation may test whether the program is being implemented as planned, the various reasons driving customer participation and satisfaction, the usefulness of tracking data, etc. The market evaluation can test the assumptions held within the program's logic model and alternative hypotheses of how the market operates and how, and whether, the program interventions are obtaining the anticipated outputs, short-term outcomes, intermediate outcomes, long-term outcomes, and ultimate outcomes.

The Two Key Functions of Evaluation – Summative and Formative Evaluations

There are two key functions that are at the core of the evaluation enterprise:

1. To document and measure the effects of a program, and
2. To help understand why those effects occurred and identify ways to improve the program.

Most parties would agree that the single most important function of evaluation is to document and measure program effects. Indeed, determining the impacts of a program or group of programs is the threshold requirement for being able to assure accountability for the expenditure of resources on that program. These are summative evaluations.[8] Summative evaluations are done after the program has been operating to document program impacts and are used to inform decisions on whether to continue, expand, cut back or end the program. Generally, impact evaluations in the energy efficiency field are summative evaluations.

However, the second key function (understanding why the observed effects occurred and identifying ways to improve program effectiveness) is also important. This function is supported through formative evaluations. Formative evaluations are often provided either early in the program's operations or steady-state programs to obtain feedback and discover ways to improve a program. Process evaluations are typically used as formative evaluations.

The importance of this perspective can be seen through a simple example. Consider the situation where the observed impacts of a program are weak, but this effect is due to an easily correctible problem in the program implementation. Simply reacting to the impact evaluation results alone might well lead to a technically accurate but policy-poor decision to terminate the program, whereas understanding why those results occurred and what

[8] The terms and distinctions identified as summative and formative evaluations have become uniformly adopted. They were first coined by Scriven in 1967 with regard to educational evaluation (Scriven 1967). Further discussions and comparisons can be found in: *Program Evaluation: Alternative Approaches and Practical Guidelines*, 2nd edition (Worthen et al. 1997), pp. 14-18; and *Evaluation: Methods for Studying Programs and Policies* (* Weiss 1998), pp. 31 -33.

can be done to improve program results can lead to improvements in the program implementation and ultimately a successful program.

Similarly, experts realize that new programs typically have to experience a "start-up" phase, where effectiveness is not yet optimized and initial costs are higher. It would be inappropriate to simply look at initial impact results and make a quick judgment about a program's potential without understanding where the program was in terms of its life cycle.

The new California Evaluation Framework is built with the understanding that the two key concepts, elements of summative and formative evaluations, are important to the overall policy decisions that would be made based upon evaluation results. High quality evaluation efforts would incorporate both concepts into their energy program evaluation design: (1) the evaluation would document program impacts and (2) the evaluation would provide for a better understanding of the observed results, the reasons for the observed results, and identify applicable opportunities for improving program performance.

These principles translate well to the California context. In California, it is incontrovertible that the primary purpose of evaluation is to document the amount of ex-post or net energy saved through the public benefit charge programs and to provide information to help determine the cost-effectiveness of acquiring those resources. However, it is also important that evaluations provide information to help improve programs and assist in making the best possible choices in a public policy context. California is known for incorporating multiple and complex objectives in its public policy directives. This makes it ever more important to utilize as comprehensive and rigorous evaluation approach as cost-effectively as possible in order to provide policy makers and regulators the information they need to make sound decisions.

The Two Basic Types of Evaluation

Proceeding to the final level of this conceptual framework, there are two basic types of evaluation that are used in energy efficiency evaluation to support the two key functions and the fundamental over-arching purpose of evaluation. These two broad categories of evaluation are:

1. Effects Evaluation
2. Process Evaluation

As the names imply, these evaluation types correspond directly to the two key evaluation functions outlined earlier. Effects evaluation includes many different types of program "effects." These include the energy and demand savings that are the focus of impact evaluations. Impact evaluation comes in many forms, ranging from simple engineering estimates applied to counts of measures installed to sophisticated statistical models incorporating measured energy consumption data and numerous other variables and statistical corrections. Other types of programs can have different types of effects

evaluations, such as market effects and transformation studies for market transformation programs, and information and education program effects evaluation. Effects evaluation also increasingly has come to include important elements that go beyond the effects that are the primary program goals to secondary effects to "non-energy effects." The one unifying theme in all these activities is the fundamental purpose of effects evaluation: to serve the key function of documenting and measuring effects caused by the program.

Similarly, process evaluation can be accomplished using a variety of methodological techniques, but its purpose is to serve the key function of documenting and understanding how the program is operating and identifying opportunities for program improvement.

An expanded view of process evaluation can include studies to understand market operations and processes and how market transformation programs are operating within those market processes. In this way, the process and effects components of evaluation can work together to provide both the functions of effects and process evaluation for market transformation programs.

It is possible to further delineate several sub-categories of market transformation program evaluation, each with a somewhat different focus and purpose. In this regard, the Framework would distinguish:

1. Market effects evaluations, which examine the changes in a market caused by a program,
2. Market operations and baseline evaluation[9], which looks at how a market operates, the key information points and information hubs and how products flow. This type of evaluation is used by evaluators for assessing program performance or program designs and operations,
3. Marketing evaluations, which examine the effectiveness of a set of marketing, or market, outreach efforts.

In many cases a well-designed and comprehensive evaluation approach will incorporate both of the basic types of evaluation: effects and process to provide a complete assessment of the program so that program and policy decisions can be grounded in the rigorous and objective information needed to support a program or policy decision.

Program Theory/Logic Model Use in Evaluation

An important component of the evaluation effort is to draw upon the program theory and logic model, to include its review (or development if one is not available) and use as an evaluation-planning tool. This section presents an introduction to program theory and logic models and their use in evaluation. It also provides references that readers unfamiliar with the program theory literature can use to learn about its uses, methods, and importance.

[9] Sometimes referred to as "market characterization."

A program theory essentially tells the story of a program, detailing what a program is going to do, why it is going to do it, and what will be accomplished as a result.

A program theory is a presentation of the goals of a program, incorporated with a detailed presentation of the activities that the program will use to accomplish those goals and the identification of the causal relationships between the activities and the program's effects. The program theory describes, in detail, the expected causal relationships between program goals and program activities in a way that allows the reader to understand why the proposed program activities are expected to result in the accomplishment of the program goals. A well-developed program theory can, and should also describe the barriers that will be overcome in order to accomplish the goals and clearly describe how the program activities are expected to overcome those barriers. The program theory should also, where possible, reference the market research that was used to identify, and the social science, marketing, and communication theories that support, the causal relationships between the program activities and the program goals. A program theory may also indicate (from the developers perspective) what program progress and goal attainment metrics should be tracked in order to assess program effects.

Program theories (PT) are sometimes called the program logic model (LM). A stricter definition would be to differentiate the program theory as the textual description while the logic model is the graphical representation of the program theory showing the flow between activities, their outputs, and subsequent short-term, intermediate, and long-term outcomes. Often the logic model is displayed with these elements in boxes and the causal flow being shown by arrows from one to the others in the program logic. It can also be displayed as a table with the linear relationship presented by the rows in the table. The interactions between activities, outputs, and outcomes are critical to understanding the program logic and argue for the need to have, or construct, both a program theory and a program logic model. An example of a logic model diagram is provided in the Appendix to this chapter.

The full range of known external influences is also important to display in the logic model. Without them, these influences may be forgotten or underestimated, and an inaccurate evaluation can result from failing to understand the context in which the hypotheses and causal mechanisms are to be tested.

The development and refinement of program theories and logic models and their use within evaluation has occurred in the larger evaluation field outside of energy program evaluation. There is a rich history and experience in using these tools for better program design, program management, and evaluation in education, public health, and social programs. It is important for evaluators to refer to the program theory literature as they assess program-developed PT/LM or when they develop alternative PT/LM in order to plan their evaluations, unless they have extensive experience in this area. Textbooks providing a good background and foundation for this work can be found in works by

Weiss,[10] Rogers et al.,[11] and McLaughlin and Jordan.[12,13] Background information was also provided in the *2001 Framework Study* that is well worth examining.[14]

It is equally important to keep an eye on the developing state-of-art of using PT/LM in the energy evaluation field as significant developments continue to occur. Early articles using PT/LM within energy program evaluation can be found in Jordan et al.[15] and Ives-Petersen.[16] A program theory evaluation approach was used to expand the market barriers perspective to include communications theory and diffusion factors in a commercial program evaluation in California.[17] PT/LM have been very helpful in identifying indicators for market transformation (MT) programs,[18] and to align indicators with performance metrics for MT programs.[19] It is also a critical component in developing evidence of causality for MT efforts.[20] At the same time, work in efficiency evaluation has noted the importance that PT/LM work be appropriately grounded in market theory and recognize the business paradigms of the program interventions being implemented.[21,22]

This approach can be equally important in evaluating educational efforts.[23] In fact, the importance of the use of PT/LM within MT program evaluation and in evaluating information and educational programs is why assessing and using PT/LM are key elements within the chapters on these types of evaluations in this Framework.

By spelling out the theory, the program evaluator will be able to identify and examine the assumptions underlying the theory by comparing the theory to what happens as the program is implemented. If the assumptions are supported, then there is reason to believe that the program is working as planned and for the reasons indicated in the program theory. If the assumptions are not supported or other alternative assumptions are

---

10 *Evaluation: Methods for Studying Programs and Policies* (* Weiss 1998).

11 *Program Theory in Evaluation: Challenges and Opportunities.* (* Rogers et al. 2000).

12 "Logic Models: A Tool for Describing Program Theory and Performance." (McLaughlin et al. 2004).

13 "Logic Models: A Tool for Telling Your Performance Story." (* McLaughlin et al. 1999).

14 *A Framework for Planning and Assessing Publicly Funded Energy Efficiency.* (* Sebold, Fields et al. 2001).

15 "Measuring and Managing the Performance of Energy Programs: A Case Study." (Jordan et al. 1997).

16 "Using the Program Logic Model to Increase the Relevance and Use of Evaluation Findings of Market Transformation Projects." (* Ives-Petersen 1999).

17 "Using Diffusion and Communications Theory to Expand Market Barrier Examination in MT Measurement." (Megdal et al. 1999).

18 "A Systematic Application of Theory-Based Implementation and Evaluation of Market Transformation Programs." (* Hastie et al. 2000).

19 "Using Program Theories to Align Performance Metrics with Public Purpose Goals." (Erickson et al. 2002).

20 "He Did It! He Did It! – Providing Evidence for Causality." (* Megdal et al. 2001).

21 "Merging Program Theory and Market Theory in the Evaluation Planning Process." (* Hall et al. 2001).

22 "When Business Analysis Tools Need to Accompany Program Theory Evaluation within Energy Efficiency Market Transformation Efforts." (Megdal et al. 2001).

23 "Detecting Behavioral Change from a Visit to a Children's Museum Energy Conservation Exhibit." (* Peters et al. 2000).

identified, then the program may not be working, or not working as efficiently as it could. In these cases the program theory needs to be modified to better reflect the actual operation of the market so that the evaluation can be on the program activities that are expected to provide the greatest benefits to the program. However, if the program theory or the assumptions about the market are not accurate, the program may need to be redesigned to increase the effects from the program. Used properly, the program theory can be used to help identify the issues to address in the evaluation, and to identify activities of the program that need to be re-examined.

An evaluation design and program theory can work hand-in-hand to be able to have the evaluation differentiate between theory failure (incomplete or inaccurate theory), and program failure (poorly designed or implemented operational procedures). In this case of theory failure, the program theory is incorrect or not complete. The assumptions about the market or the causal mechanisms that create attitudes or behaviors are not valid, are only partially valid or missing key theoretical components. The program theory needs to change and the program needs to be refined accordingly.

On the other hand, if the evaluation identifies program failure, this means that the theory appears to be correct. However, the program implementation had problems that did not allow it to have the anticipated outputs and initial subsequent outcomes. Figure 4.1 shows the differences between theory failure and program failure.

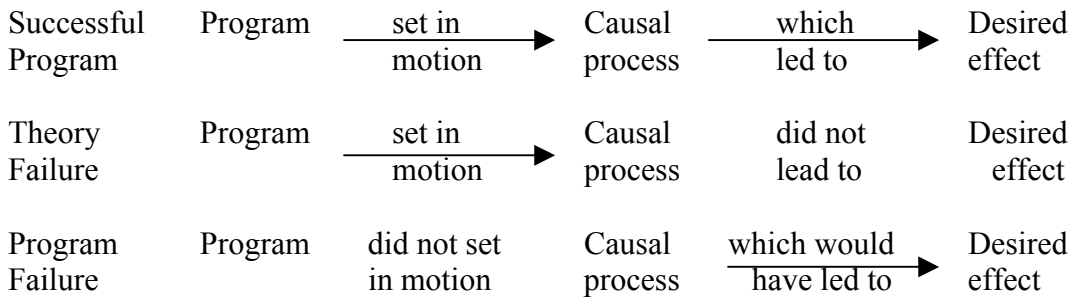

Figure 4.1: Theory Failure & Program Implementation Failure[24]

Developing a Program Theory

The easiest way to develop a program theory is to start by systematically describing a program in terms of resources, activities, outputs, short-term outcomes and long-term outcomes. (See the literature cited above in the first subsection of the PT/LM section.)

- *Inputs (resources)* are the elements required by an organization, program, or project to initiate and/or sustain activities. Examples are money, collaborations, skills, and time.
- *Activities* are the program activities that are used to produce the outputs that initiate the causal logic within the logic model.

[24] *Evaluation: Methods for Studying Programs and Policies.* (* Weiss 1998), page 129.

- *Market actors* are those market actors targeted by the interventions or that play a role in the causal logic of the program theory.
- *Outputs* are the immediate results of an activity. Examples are the number of contacts made, number of brochures printed, number of contractors recruited, and number of audits completed.
- *Outcomes* are the intermediate or once removed consequences resulting from program activities and program outputs. There may be a sequence of outcomes. Outcomes may be unintended or intended but they are not prompted by direct action on the part of the program. Examples are changes in awareness, attitudes, and behaviors, participants referring non-participants to the program, trade publications running articles about efficient equipment and practices, dealers changing their stocking practices, etc.
- *Long-term outcomes (impacts)* are the end-states to be realized. Impacts may take months and years to accomplish and may be influenced indirectly by the intervener's actions. Impacts are the long-term goals of the program. Examples are kWh saved, gallons of water saved, tons of CO_2 reduced, efficient technologies and practices are the industry standard, T-12 fluorescents are difficult to buy, etc.

Once a program has been systematically described in terms of resources, outputs, outcomes, and long-term impacts, the procedure for developing a program theory is a systematic one. One of the best ways to develop a program theory is to start with the long-term outcomes and work backwards to resources. Essentially, the process is one of repeatedly asking the same question, if "Z" is a long-term outcome (or short-term outcome, output, or activity), what is required to produce "Z." It is then a matter of writing the causal relation in the form of a statement; "Y" will cause "Z." One then backs up and asks what will cause Y and continues until one has described the required activities and resources.

One can then reverse the order and edit the statements until one has a sequence of causal statements that describe how the program works.

Several things are likely to happen as this is being worked upon. It is likely that gaps in the causal relationships between actions and expected effects will be found. Some steps will be identified that require substantial leaps that suggest that the theory needs further refinement. Some steps in the theory will seem quite improbable, suggesting that the theory, and probably the program design, needs improvement. Some steps will contradict what is known from the program, marketing and evaluation literature and other social science and business theories. Sometimes an assumption, key to the way the program is designed for one of its causal chains, can be in conflict with an assumption in a different causal chain. All of these instances identify places where the program theory and the program design may need improvement.

This process and the resulting program theory and logic model provides program information that helps isolate important measurement indicators. These are indicators for the outputs and outcomes of the program. More importantly, the program has been

described in a way that will enable the evaluator to identify key issues and to identify important questions on program assumptions or key hypotheses that the evaluation needs to be able to answer.

An example of a Program Theory is provided in the Appendix to this Chapter.

Reasons for an Evaluator to Use and Refine the Program Theory

Ideally, the program evaluator should not find it necessary to develop a program theory to support the evaluation planning process. All programs should have a detailed program theory that is used to justify the need for the program, justify the activities needed to accomplish the program goals, and determine the funding levels needed to successfully implement the program. In the past, program theories have often been embedded in the general logic associated with a program's efforts and not specifically developed into a formal program theory. However, in the last several years organizations that provide funding for energy efficiency programs have begun to require detailed program theories so that they can make program and program funding decisions to help assure that public benefits dollars are well spent and that there is a high probability that program goals will be obtained. When a program theory is already developed the evaluation planning efforts should be informed by that "official" program theory. However, evaluators should not rely only on the official program theory for their evaluation planning efforts. When evaluators examine the official program theory it is not unusual for the evaluator to identify alternative paths not reflected in the official program theory by which participants can reach the same desired outcomes as those reflected in the official program theory. However, when a program does not have a formal program theory the evaluation team should develop a program theory in cooperation with the program administrators that can be used to inform the evaluation planning process. Regardless of the presence or absence of an official program-developed program theory, it is up to the evaluator to identify research approaches that test for the different causes for the effects that are projected by the program theory.

There are a variety of reasons why a program theory should be used for developing an evaluation plan. These include:

- The program theory can be used to examine and clarify program assumptions and program goals so that the evaluation team has a complete understanding of the program to be evaluated.
- The program theory can be used to help identify potentially missing or extraneous activities that need to be addressed in the evaluation.
- The program theory provides a program description that can be used to:
  - Identify key hypotheses to be tested as part of the evaluation.
  - Identify key measures for which data need to be collected.
- Because the program theory provides a logic and sequence, it can be used to identify measures to be included in the evaluation that are appropriate to the stages of program development at the time the evaluation is being conducted.

- The program theory provides a model against which program activities, assumptions and goals can be compared through time and that can be used as a basis for identifying changes to the program.
- The program theory can be used as the foundation from which recommended program changes can be made and to estimate the potential effects of those recommendations on the operations or success of the program.
- The program logic model can summarize the key testable hypotheses; each of the outputs and outcomes are possible indicators, and each causal arrow can be tested for its occurrence, its mechanism, and whether there are important external influences on them that also need to be measured. These can form the list of potential measurement elements that are then prioritized and assessed for whether and when they should be measured.
- Permits multiple explanations for how a program is expected to work.
- Enables the incorporation of unintended consequences or new learning about the market and its participants.

Due to the importance of using the program theory and logic model as part of evaluation planning, program theories and logic models that are already developed by the program are critically assessed by the evaluator during the first steps of the evaluation planning process. In this way verification can be made on whether all the necessary elements are there and seem logical given relevant theories and experience in the evaluation field. This is necessary to ensure that the PT/LM can provide an accurate foundation for the evaluation plan.

Program Theories are Especially Important to the Evaluation of Complex Programs

Using a program theory in the evaluation planning process is important for all types of program evaluations. However, the use of a program theory is particularly important for evaluations of complex programs with long-term goals such as information and educational programs and programs that are trying to change how a market operates where the program activities and the desired long-term or ultimate outcomes of the program may be many steps removed from one another.

In the case of resource acquisition programs, the causal linkages between an intervention and the desired behavior are usually quite direct and the program theory quite compact. For example, a rebate is offered to reduce the cost of an efficient washing machine to increase the number of householders purchasing efficient washing machines in order to reduce consumer energy use (*2001 Framework Study*, Chapter 4).

For programs that attempt to change how a market operates (market transformation), the causal chain may be quite lengthy involving numerous interventions in different parts of a market with different market actors. For example, an education program targeting grade schools may depend on a series of events to occur over a considerable period of time before the effects of the program begin to appear. Likewise, a program that attempts

to change the way products are manufactured may need to rely more on efforts that change product demand levels beyond what can be achieved by a limited number of rebates offered to one or two segments of the market at a specific point in time. In these cases the program theory may be more elaborate than the theories associated with the resources acquisition program.

The Relationship Between Program Theory, Market Assessment and Market Theory

In order to develop a useful program theory, an understanding of the market is crucial. For example, the important decision makers for the facilities of major retail chains are often a small group of architects and engineers located in proximity to the retail chain's headquarters who do most of the design work for the chain. These "image designers" may employ local architects and engineers to help steer projects through local zoning and permitting processes but the local architects and engineers have little say in the overall design. If one is attempting to target new construction or major renovations in retail establishments, it is important to understand this relationship. The implication of this is that the local design is being driven by regional or national standards and not by the local architects or engineers who must focus on the permitting process. It may be difficult for locality-based organizations to influence these standards. Regional and national efforts might be more effective in influencing national retailers. In some cases the national firms are aware of the potential for local rebates and may have standard designs that can be used to capture the rebates thus making them free riders for the local program. Local efforts to improve energy efficiency might better be focused on local retailers and the local architects and engineers who serve them. The important point of this example is that without an understanding of the market that one is trying to influence substantial resources can be directed to parts of the market where little or no effects can be expected.

Thus, program theories need to be based on good market theories and good market theories need to be based in strong empirical evidence. In the course of developing a program theory, program designers need to have an understanding of how their target market operates. Where empirical data is lacking to support a market theory, program designers may need to conduct a market characterization study. (See Chapter 10 for a description and content of a market characterization study.)

A program theory is only as good as its empirical underpinnings. If the market theory is absent or wrong, then it is possible that the program theory will be incorrect as well. This may cause the evaluator to focus on issues that are not relevant and may lead to conclusions about the program that are not helpful.

Evaluation Ethics

Credibility of evaluations and evaluators are absolutely essential for evaluations to fill their role in providing credible findings on the results from the program and for providing recommendations for program refinement and investment decisions. This makes

evaluation ethics a critical foundation for evaluations and the field of evaluation. The Framework could not be complete without reinforcing the importance of evaluation ethics in the practice of evaluating energy efficiency programs.

The American Evaluation Association (AEA) has developed and approved a set of guiding principles for evaluators. These well-established guidelines are part of the new California Evaluation Framework and the authors recommend that these principles guide energy program evaluations and the evaluators who conduct these studies. These guidelines are provided in every addition of The American Journal of Evaluation, on the AEA's website, and were provided in the preface pages of the hard-copy *Proceedings of the 2003 International Energy Program Evaluation Conference*.

These guidelines are provided in their entirety in the Appendix to this chapter. A high level summary of the guiding principles is given below.

A. Systematic Inquiry: *Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.*
B. Competence: *Evaluators provide competent performance to stakeholders.*
C. Integrity/Honesty: *Evaluators ensure the honesty and integrity of the entire evaluation process.*
D. Respect for People: *Evaluators respect the security, dignity and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.*
E. Responsibilities for General and Public Welfare: *Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.*

History of Energy Efficiency Program Evaluation in California

The history of California's energy efficiency programs and related measurement and evaluation (M&E) requirements can be divided into four distinct periods: the pre-Protocol era (1970s-1994), the Protocol era (1994-1997), the Restructuring era (1998-2000), and the current transition period (post-2000). In particular, in the last five years the environment within which energy efficiency programs operate has changed enormously. The market structure for electricity in California changed, and simultaneously, the California Public Utilities Commission (CPUC) shifted the focus of energy efficiency programs from resource acquisition to market transformation, and a new generation of programs was added to the older ones. However, the energy crisis of 2001 forced regulators to revisit these policy decisions, with a renewed emphasis on resource acquisition. The optimal balance between resource acquisition and market transformation is still in question, and the relative emphasis is expected to change as conditions warrant. The optimal roles of utilities and other market entities also remain to

be decided. The CPUC's desire to update the Framework for assessing programs to meet the new realities is the genesis of this report.

1970s-1994: The Pre-Protocol Era

For over twenty years, the CPUC has approved the use of ratepayer funds to promote energy efficiency activities, and authorized the major investor-owned utilities (IOUs) under its jurisdiction to administer a wide variety of energy efficiency programs. By the early 1990s, a wide variety of these programs began to be planned and carried out as part of the biennial resource planning update with the CPUC and the California Energy Commission (CEC). By the early 1990s, energy efficiency programs and other demand-side management (DSM) activities were identified as viable, cost-effective alternatives to supply-side energy generation projects (the "resource acquisition" perspective).

From the late 1980s until 2002, the CPUC allowed the utilities to recover from ratepayers the costs of "shareholder incentive mechanisms." The terms and conditions under which the utilities were allowed to claim and recover these transfers varied greatly and the earnings claims were verified in separate proceedings, and each mechanism allowed for a very short earnings recovery period.

In June 1990, the CPUC permitted the Division of Ratepayer Advocates (DRA, later ORA) to receive funds to review utilities' energy efficiency programs. During this period, utilities were calculating the results of their programs through engineering estimates (often referred to as "*ex-ante* estimates"). Beginning in 1993, the CPUC required that utilities should rely more on measurement than engineering and required that the evaluations of energy efficiency programs be based on "*ex-post* measurement." The 1993 CPUC decision also required that measurement and evaluation protocols be established, and formally established CADMAC – the California DSM Measurement Advisory Committee – to develop these protocols. CADMAC prepared semi-annual reports on the progress of the protocols and also hosted informal workshops where participants could freely discuss the technical issues of the protocols and their implications.

1994-1997: The Protocol Period

For the 1994-1997 years, statewide consistency was established for the shareholder incentive mechanisms, more rigorous terms and conditions for the measurement and verification of costs and benefits were established based on the CADMAC-developed Protocols. These were adopted by the CPUC in D.93-05-063, providing the rules by which impact evaluations were done to determine the energy savings achievements of programs for which shareholder earnings were awarded. Earnings claims were addressed in the consolidated Annual Earnings Assessment Proceeding (AEAP), with a relatively lengthy earnings recovery period of up to 10 years. The protocols required first-year impact evaluations for most programs (emphasizing regression analysis and use of billing data) and persistence studies (fourth and ninth year measure retention studies, and research on relative technical degradation). Since the ultimate earning level depended on

the measured persistence of savings, studies under these protocols are still in progress and are scheduled to occur again in 2006.

1998-2000: Energy Efficiency, Electric Industry Restructuring, and The CBEE Period

Beginning in 1998, several significant program design and implementation changes occurred. First, funding for traditional rebates was reduced and replaced by Standard Performance Contract (SPC) programs, where savings and incentives were based on measured performance. Second, funding for "upstream market transformation" interventions was substantially increased. Third, utility performance awards were substantially delinked from cost-effectiveness considerations, and there were reductions in the earnings opportunities for the utilities. Fourth, there were significant increases in expenditures on M&E studies that attempted to quantify market effects and indirect benefits attributable to the expanded upstream market transformation programs.

Beginning in 1998, the CPUC moved to a market transformation goal for programs.[25] The CPUC codified a policy that emphasized removing barriers to energy efficiency in the market so that private sector entities would be able to provide energy efficiency services once public monies were no longer available to fund activities.

As a result of this move towards market transformation, an efficiency advisory board, the California Board for Energy Efficiency (CBEE), established a new set of expectations. The CBEE provided advice to the CPUC on the types of programs to fund as well as the type of measurement and evaluation requirements needed for evaluating market transformation programs. During this period, the California Measurement Advisory Council (CALMAC) was formed to address issues related to programs conducted in 1998 and onwards, particularly M&E efforts. M&E efforts focused on verification of the number of energy efficiency measures performed or installed as reported by the IOUs, and the measurement of actual energy savings achieved was de-emphasized.

The CBEE recommended a new cost-effectiveness test, the Public Purpose Test (PPT). The PPT includes elements that were not traditionally included in the TRC calculations, such as spillover savings, non-energy costs/benefits, positive/negative externalities, and reductions in the cost of measures or practices caused by the program. In addition, the PPT is applied at the portfolio level to encourage investment in interventions that may not produce measurable savings in the early years, but are more cost-effective over time as market effects compound in later years. In April 1999, CPUC Resolution E-3592 approved modifications to the Policy Rules and, in so doing, adopted the PPT as the standard for cost-effectiveness.

One of the last studies requested by CBEE involved a team of evaluation and cost-effectiveness experts who completed a major framework that discussed the justification for, the logic behind, and the techniques needed for evaluating and assessing the cost-

[25] See *A Scoping Study on Energy efficiency Market Transformation by California Utility DSM Programs* (* Eto et al. 1996).

effectiveness of market transformation interventions (the *2001 Framework Study*).[26] While the results are worthy of serious consideration, they were never incorporated in any set of M&E protocols or guidelines due to the continuing turmoil in DSM planning and regulation in California.

Post-2000: Transition Period and the Energy Crisis

The current transition period has been a period of great uncertainty, sparked by the Energy Crisis of 2001.

The Energy Crisis

By almost any measure, the events surrounding the electricity situation in California in the 2000/2001 time period were simply extraordinary.[27] Between the summer of 2000 and the early winter months of 2001, the California Independent System Operator declared over seventy days of system emergencies, and rolling blackouts were actually initiated on several occasions. In January and February 2001, the CEC projected electricity supply and demand for the summer of 2001 under various temperature scenarios, and analyses suggested that the State could face a potential shortfall of five thousand megawatts during the months of June through September.

In reaction to this, California added significant new funding for energy efficiency.[28, 29] In all, policy makers ordered more than a 250% increase in spending over the 2000 level. Most of this funding focused on resource acquisition activities to address the near-term crisis, rather than market transformation.

Changing Rules

In November 2001, the CPUC changed the rules for energy efficiency programs to allow non-utilities to compete with utilities for energy efficiency funding in Decision 01-011-066. The Commission eventually selected approximately 40 third parties to administer approximately $100 million in funds for local 2002 and 2003 energy efficiency programs.

For 2002, the CPUC adopted a new Energy Efficiency Policy Manual (contained in CPUC Decision 01-011-066)[30], and program evaluation and numerous other program activities are governed by this set of protocols. The manual contains the requirements for cost-effectiveness analysis, and requires measurement and verification using the U.S. Department of Energy's International Measurement and Verification Protocol.[31]

---

26 *A Framework for Planning and Assessing Publicly Funded Energy Efficiency* (* Sebold, Fields et al. 2001).

27 *Examining California's Energy Efficiency Policy Response to the 2000/2001 Electricity Crisis: Practical Lessons Learned Regarding Policies, Administration, and Implementation*. (Kushler et al. 2003).

28 *Evaluation: Methods for Studying Programs and Policies* (* Weiss 1998).

29 "Using Energy Efficiency to Help Address Electric System Reliability: An Initial Examination of 2001 Experience." (Kushler et al. 2003)

30 (CPUC 2001).

31 (International Performance Measurement and Verification Protocol 2001).

For Program Years 2003 and 2004, the penalty "stick" replaced the incentive "carrot." There are no shareholder earnings for either utility- or non-utility-administered energy efficiency programs. Instead, a holdback mechanism has been introduced for the non-utility-administered programs, and utilities are subject to a penalty process for subpar program performance.

Resource Procurement and Portfolio Management
In another major policy shift, the CPUC required (in October 2002, Decision 02-10-062) that California's three electric IOUs take responsibility for resource procurement to maintain the reliability of California's electric grid. To meet this CPUC mandate, the utilities filed portfolio management plans in April 2003, which included $700 million in energy efficiency spending over the next five years in addition to current commitments through the Public Goods Charge funds. With such a mix of implementers and policy objectives, new evaluation protocols are simply one of the many pieces that must come together to create a consistent and reliable program environment.

Policy Use of Evaluation Results

Policy Based on Evaluation Results

Good policy is based upon reliable, accurate information. It is this principle that makes quality evaluations imperative. Its importance can be seen in the current movement towards greater "evidence-based decision-making" in policy arenas[32] [33]and health care[34]. Evidence-based decision-making requires a focused question that is then addressed by assessing and applying the evaluation results to answering that question through the art of critical thinking. It is important for this process to include information concerning uncertainty and potential bias. (See Chapter 12: Uncertainty, for a discussion of the Framework's guidelines for these issues.) With a systematic approach using these methods, policy can be well-grounded, well-supported, and better allow consistent and incremental policy that is refined as research suggests, increasing the likelihood that the acceptance and adjustments to policy changes can occur more smoothly.

Interpreting and Adjusting Results for Summative versus Formative Purposes

It is important that users of evaluation results interpret study results correctly. As discussed earlier in this chapter, a summative evaluation answers the question of what occurred within a program over a particular time period. This information is appropriate

---

32 See (Public Policy Forum 2003) for information on UK Government movement on evidence-based decision-making.

33 "Evaluation, Policy Learning and Evidence-Based Policy Making." (Sanderson 2002).

34 A large effort of the last 5-10 years has developed here for clinical applications. Evidence-based decision-making gained prominence in Canadian health care with the Prime Minister's National Forum on Health in 1997, where it was defined as "the systematic application of the best available evidence to the evaluation of options and to decision-making in clinical, management and policy settings" from *Health Services Research and Evidence-Based Decision-Making* (Canadian Health Services Research Foundation 2000).

for use in understanding the program's accomplishments and its contribution to the overall portfolio. Further, information from a formative evaluation is needed to decide what that program's future results might look like, indicating whether it should be continued or not. Similarly, summative results only provide partial information on which to base a decision about offering a program in a different location. In this latter case, formative evaluation information is needed to understand why and under what conditions the results can be obtained in order to help determine if the program should be replicated at another locations and if the results from the new program would be similar or different.

An example of this concept that is used frequently within this Framework is a program that is in its early implementation stage that has significant start-up costs. If during the first year a benefit-cost analyses is used to decide whether to continue the program a significant risk occurs that the program might be highly cost-effective over the long-term, but cancelled because of the results of the first year assessment. This means the initial savings and benefit-cost analysis may be one of several important elements to report as part of the summative evaluation. However, there should also be a formative assessment concerning the probable long-term benefit-cost that may provide a different picture of the program to policy makers.

A less obvious example is when the effects of a significant change in the economy also affects the impact evaluation results. For example, in cases when a recession coincides with post-retrofit periods, a billing analysis may overstate savings. In this case correction techniques should be applied in the billing analysis to capture this effect and correct the energy savings estimates. Engineering methods that incorporate fieldwork may recognize this condition and give the evaluation team a notice of the current building condition. At the same time, both methods need to provide an estimate of what savings would be expected under more steady-state economic conditions. A similar but opposite challenge might occur during periods of fast economic growth. In these cases, crowding of employees might occur prior to physical expansion (by either new construction or renting new space). This could mean that savings are actually greater than what would show up in a billing analysis. Billing analysis, that doesn't correct for this condition, will understate savings (as post-retrofit usage is greater). Engineering estimates of savings can overstate long-term savings if the current operating conditions are assumed to be occurring over the long-term, when instead they may actually only be a temporary condition. Each of these scenarios points to the need for careful evaluation with accurate analysis and reporting that reflects changes going out beyond the energy technology changes.

Summative results also need to be assessed in conjunction with process evaluation results in order to understand the program's potential under differing program and operational conditions to support decisions associated with program continuation or funding.

Making the Framework a Living Document

This document provides a great many references to help guide policy makers, energy suppliers, evaluation planners, evaluators, and other stakeholders understand the

environment in which evaluations can be conducted and the scope of these efforts. However, as evaluation efforts are undertaken evaluation professionals will continue to find ways to improve upon the current state-of-the art. As additional research is completed and improvements are discovered for specific evaluation practices, a systematic process needs to be in place that will allow the findings from this research to be used to modify and refine the Framework, either as an actual document modification allowing for the Framework to be updated, or through a referral sending users to the potentially modifying documents.

Similarly, as elements of this Framework are used in the field, lessons will be learned that can lead to additional recommendations for refining this Framework. Likewise, a systematic process needs to be put in place that encourages and allows this evolution to occur.

Finally, the evaluation field is not static. The highest quality Framework needs to be able to change and grow in response to developments within the energy program evaluation field as well as the larger field of evaluation. A means for this to occur needs to be established.

Appendix to Chapter 4

An Example of a Program Theory

To illustrate a program theory, we examine a program with a goal to reduce energy consumption in multifamily buildings. The program targets owners of multifamily buildings who are making equipment replacement decisions typically for HVAC systems, including large air conditioning systems and central or distributed heating systems. The idea is to obtain immediate savings from owners who are currently making equipment replacement decisions by encouraging more efficient equipment installations, while encouraging these owners, their contractors, and their distributors to adopt more energy efficient decision practices and equipment in the longer term. In this example the program is operated via a program vendor who works with equipment installation contractors to identify facility owners who are or can be encouraged to be interested in an energy efficient assessment of their facility linked with rebates for efficient equipment replacements or upgrades that are recommended as a result of the assessment.

A portion of the program theory for this program might look something like the following:

Multi-family market research conducted by the program vendor indicates that facility owners are interested in increasing or maintaining the value of their property, in reducing the hassles associated with facility maintenance, in reducing the operational costs associated with the facility, in being seen in the market as a provider of high-quality rental units, and in making the facility more attractive to potential buyers. This market condition results in a market opportunity for a program to offer services that makes it economically attractive for owners to achieve their goals while saving energy at the same time. The program vendor hypothesizes that if they work through equipment installation contractors who have pre-existing relationships in the multi-family market they can use these relationships to convince owners to make the changes needed to achieve the program goals.

The market theory indicates that there are two multi-family market sub-segments that can be successfully reached by the program that will produce energy savings. These include:

1. Facility owners who are not currently considering an upgrade but who can be convinced to retrofit their facilities with more efficient equipment, and
2. Facility owners who are considering an upgrade to install low-cost/low-efficiency equipment, but who can be convinced to install more efficient equipment than they were originally considering.

The program vendor hypothesizes that through these two types of market actors the program can deliver a specific level of energy and demand savings to California's energy markets. The program vendor indicates that they can capture the energy savings through

the following program activities based upon the program assumptions concerning the market and the consequences (outputs and the causal chain of outcomes from these outputs).

Program Activities

- Initial program launch will have the vendor working with a small set of innovative customers already known by the vendor. Vendor will establish relationship with these customers as the innovative launch customers.
- The program vendor will make personal contact with installation contractors and describe in detail the benefits and operations of the facility assessment and reward program.
- The program vendor will design, develop and provide to the contractors program marketing and outreach materials.
- When a customer seeking an inefficient replacement system approaches an installation contractor or when the contractor judges that a customer or potential customer is a good candidate for the program, the installation contractor will contact the customer and recommend the facility assessment and reward program.
- The installation contractor will then instruct the program vendor to conduct the assessment, recommending equipment and/or installation practices that will save energy and accomplish the customer's participation objectives. The program vendor makes a presentation to the owner and the installation contractor offering a reward that represents the incremental cost between what would have been the standard installation and the efficiency upgrade.
- The program vendor assists the installation contractor as necessary, concerning where efficiency equipment can be obtained and the methods for its proper installation.
- The owner and the installation contractor confer and the owner decides to proceed with the upgrade.
- The upgrade is installed.

Program Assumptions

- The program vendor can organize and implement the program and make contacts with the target customers.
- Initial participants will be a small set of innovative customers known by the vendors, and these relationships with these customers will help launch the program.
- The program marketing and outreach materials (developed by the vendor) will be accepted by the contractors and help gain acceptance by their customers in such a way as to increase program enrollments and equipment installations.

- The vendor's activities with these contacts and the program marketing materials will be enough to motivate contractors to join the program and market to their customers and potential customers.
- Contractor enrollments and participation will occur at the rate assumed from both targeted market segments such that the program achieves its energy-related goals.
- The program is attractive to installation contractors because it gives them a new tool to use to gain additional business, added profits, and additional customers than they would have without the program.
- The program is attractive to customers because it provides a free assessment service and cash rewards to customers in two market segments that enables them to increase the value and marketability of their property while reducing operational costs.
- The market theory describes several market barriers for contractor enrollments and customer participation, but identifies several market and technology advantages associated with the program that the vendor thinks could overcome those barriers.
- The installations are made in a high quality manner and assumed energy and demand savings are thus achieved.

The goal of this program is defined in the program theory to reduce the energy consumption of 100 multi-family facilities having a total of more than 1,500 living units in the South Bay area of California by "X" kWh a year with a corresponding drop in coincident demand by "X" kW during the first year of program operation, by "X" kWh a year with a corresponding drop in coincident demand by "X" kW during the second year of program operation, and by "X" kWh a year with a corresponding drop in coincident demand by "X" kW during the third year of program operation. The program theory indicates that eighty percent of the projected savings will be produced over a fifteen-year period consistent with the expected life of the installed technology.

What these assumptions and activity statements represent is a series of hypotheses about how the program vendor thinks the program will work in the market place. A logic model can be created that represents how the program activities are hypothesized to lead to program outputs. Then there are assumptions of how the program outputs will lead to short-term outcomes. The first activities statement hypothesizes that personal relationships between the vendor and a small number of innovative contractors is important to the program launch. Another hypothesis is that the vendor's work with contractors and the program marketing materials will cause contractors to participate with the program and be able to "sell" their customers on participating in the program. These and all of the other assumptions and program logic included within the above become the evaluation's testable hypotheses.

An Example of a Logic Model

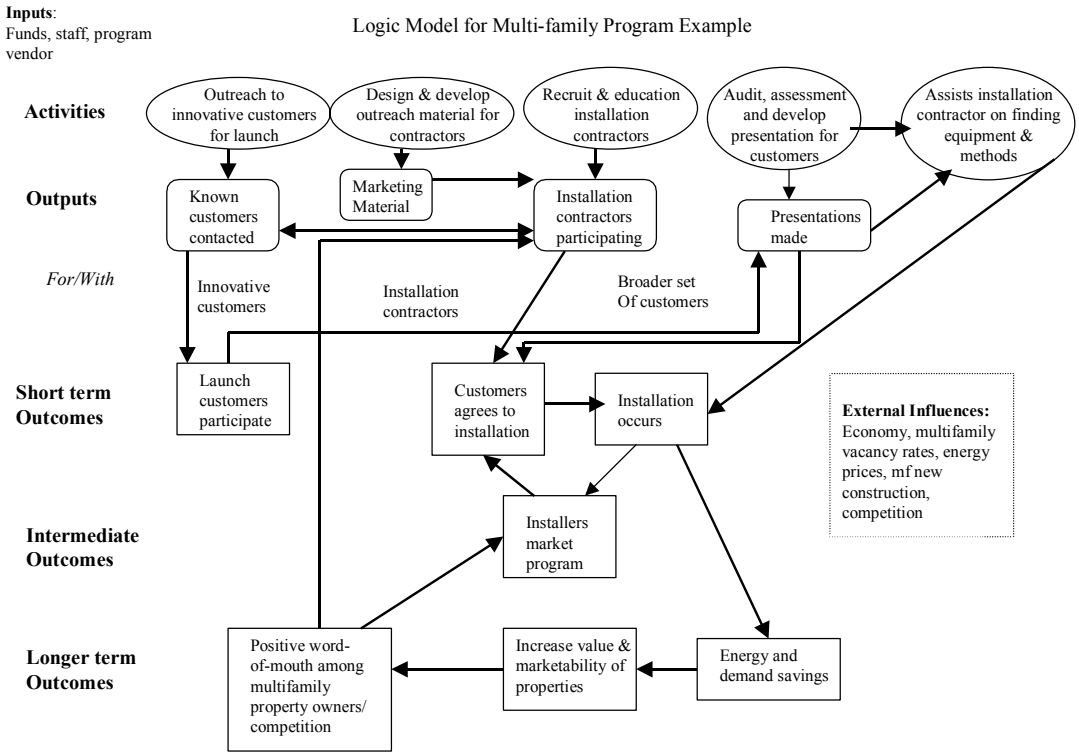

Figure 4.2: Logic Model for Multi-family Program Example

Complete Evaluation Ethics Document from the American Evaluation Association

These guidelines are copied here in their entirety as printed in The American Journal of Evaluation.

American Evaluation Association: Guiding Principles for Evaluators
The American Evaluation Association (AEA), the professional association for evaluators, works to ensure ethical work in the evaluations of programs, products, personnel, and policy. To proactively guide the work of professionals in everyday practice and to inform evaluation clients and the general public of expectations for ethical behavior, the Association developed the following guiding principles:

A. Systematic Inquiry: *Evaluators conduct systematic, data-based inquiries about whatever is being evaluated.*
   1. Evaluators should adhere to the highest appropriate technical standards in conducting their work, whether that work is quantitative or qualitative in nature, so as to increase the accuracy and credibility of the evaluative information they produce.
   2. Evaluators should explore with the client the shortcomings and strengths both of the various evaluation questions it might be productive to ask, and the various approaches that might be used for answering those questions.
   3. When presenting their work, evaluators should communicate their methods and approaches accurately and in sufficient detail to allow others to understand, interpret and critique their work. They should make clear the limitations of an evaluation and its results. Evaluators should discuss in a contextually appropriate way those values, assumptions, theories, methods, results, and analyses that significantly affect the interpretation of the evaluative findings. These statements apply to all aspects of the evaluation, from its initial conceptualization to the eventual use of findings.

B. Competence: *Evaluators provide competent performance to stakeholders.*
   1. Evaluators should possess (or, here and elsewhere as appropriate, ensure that the evaluation team possesses) the education, abilities, skills and experience appropriate to undertake the tasks proposed in the evaluation.
   2. Evaluators should practice within the limits of their professional training and competence, and should decline to conduct evaluations that fall substantially outside those limits. When declining the commission or request is not feasible or appropriate, evaluators should make clear any significant limitations on the evaluation that might result. Evaluators should make every effort to gain the competence directly or through the assistance of others who possess the required expertise.
   3. Evaluators should continually seek to maintain and improve their competencies, in order to provide the highest level of performance in their evaluations. This continuing professional development might include formal coursework and workshops, self-study, evaluations of one's own practice, and working with other evaluators to learn from their skills and expertise.

C. Integrity/Honesty: *Evaluators ensure the honesty and integrity of the entire evaluation process.*

1. Evaluators should negotiate honestly with clients and relevant stakeholders concerning the costs, tasks to be undertaken, limitations of methodology, scope of results likely to be obtained, and uses of data resulting from a specific evaluation. It is primarily the evaluator's responsibility to initiate discussion and clarification of these matters, not the client's.
2. Evaluators should record all changes made in the originally negotiated project plans, and the reasons why the changes were made. If those changes would significantly affect the scope and likely results of the evaluation, the evaluator should inform the client and other important stakeholders in a timely fashion (barring good reason to the contrary, before proceeding with further work) of the changes and their likely impact.
3. Evaluators should seek to determine, and where appropriate be explicit about, their own, their clients', and other stakeholders' interests concerning the conduct and outcomes of an evaluation (including financial, political and career interests).
4. Evaluators should disclose any roles or relationships they have concerning whatever is being evaluated that might pose a significant conflict of interest with their role as an evaluator. Any such conflict should be mentioned in reports of the evaluation results.
5. Evaluators should not misrepresent their procedures, data or findings. Within reasonable limits, they should attempt to prevent or correct any substantial misuses of their work by others.
6. If evaluators determine that certain procedures or activities seem likely to produce misleading evaluative information or conclusions, they have the responsibility to communicate their concerns, and the reasons for them, to the client (the one who funds or requests the evaluation). If discussions with the client do not resolve these concerns, so that a misleading evaluation is then implemented, the evaluator may legitimately decline to conduct the evaluation if that is feasible and appropriate. If not, the evaluator should consult colleagues or relevant stakeholders about other proper ways to proceed (options might include, but are not limited to, discussions at a higher level, a dissenting cover letter or appendix, or refusal to sign the final document).
7. Barring compelling reason to the contrary, evaluators should disclose all sources of financial support for an evaluation, and the source of the request for the evaluation.

D. Respect for People: *Evaluators respect the security, dignity and self-worth of the respondents, program participants, clients, and other stakeholders with whom they interact.*

1. Where applicable, evaluators must abide by current professional ethics and standards regarding risks, harms, and burdens that might be engendered to those participating in the evaluation; regarding informed consent for participation in evaluation; and regarding informing participants about the scope and limits of confidentiality. Examples of such standards include federal regulations about protection of human subjects, or the ethical principles of such associations as the American Anthropological Association, the American Educational Research

Association, or the American Psychological Association. Although this principle is not intended to extend the applicability of such ethics and standards beyond their current scope, evaluators should abide by them where it is feasible and desirable to do so.

2. Because justified negative or critical conclusions from an evaluation must be explicitly stated, evaluations sometimes produce results that harm client or stakeholder interests. Under this circumstance, evaluators should seek to maximize the benefits and reduce any unnecessary harm that might occur, provided this will not compromise the integrity of the evaluation findings. Evaluators should carefully judge when the benefits from doing the evaluation or in performing certain evaluation procedures should be foregone because of the risks or harms. Where possible, these issues should be anticipated during the negotiation of the evaluation.
3. Knowing that evaluations often will negatively affect the interests of some stakeholders, evaluators should conduct the evaluation and communicate its results in a way that clearly respects the stakeholders' dignity and self-worth.
4. Where feasible, evaluators should attempt to foster the social equity of the evaluation, so that those who give to the evaluation can receive some benefits in return. For example, evaluators should seek to ensure that those who bear the burdens of contributing data and incurring any risks are doing so willingly, and that they have full knowledge of, and maximum feasible opportunity to obtain any benefits that may be produced from the evaluation. When it would not endanger the integrity of the evaluation, respondents or program participants should be informed if and how they can receive services to which they are otherwise entitled without participating in the evaluation.
5. Evaluators have the responsibility to identify and respect differences among participants, such as differences in their culture, religion, gender, disability, age, sexual orientation and ethnicity, and to be mindful of potential implications of these differences when planning, conducting, analyzing, and reporting their evaluations.

E. Responsibilities for General and Public Welfare: *Evaluators articulate and take into account the diversity of interests and values that may be related to the general and public welfare.*

1. When planning and reporting evaluations, evaluators should consider including important perspectives and interests of the full range of stakeholders in the object being evaluated. Evaluators should carefully consider the justification when omitting important value perspectives or the views of important groups.
2. Evaluators should consider not only the immediate operations and outcomes of whatever is being evaluated, but also the broad assumptions, implications and potential side effects of it.
3. Freedom of information is essential in a democracy. Hence, barring compelling reason to the contrary, evaluators should allow all relevant stakeholders to have access to evaluative information, and should actively disseminate that information to stakeholders if resources allow. If different evaluation results are communicated in forms that are tailored to the interests of different stakeholders, those communications should ensure that each stakeholder group is aware of the

existence of the other communications. Communications that are tailored to a given stakeholder should always include all important results that may bear on interests of that stakeholder. In all cases, evaluators should strive to present results as clearly and simply as accuracy allows so that clients and other stakeholders can easily understand the evaluation process and results.

4. Evaluators should maintain a balance between client needs and other needs. Evaluators necessarily have a special relationship with the client who funds or requests the evaluation. By virtue of that relationship, evaluators must strive to meet legitimate client needs whenever it is feasible and appropriate to do so. However, that relationship can also place evaluators in difficult dilemmas when client interests conflict with other interests, or when client interests conflict with the obligation of evaluators for systematic inquiry, competence, integrity, and respect for people. In these cases, evaluators should explicitly identify and discuss the conflicts with the client and relevant stakeholders, resolve them when possible, determine whether continued work on the evaluation is advisable if the conflicts cannot be resolved, and make clear any significant limitations on the evaluation that might result if the conflict is not resolved.
5. Evaluators have obligations that encompass the public interest and good. These obligations are especially important when evaluators are supported by publicly generated funds; but clear threats to the public good should never be ignored in any evaluation. Because the public interest and good are rarely the same as the interests of any particular group (including those of the client or funding agency), evaluators will usually have to go beyond an analysis of particular stakeholder interests when considering the welfare of society as a whole.

Chapter 5: Umbrella Framework for Evaluation Planning, Funding and Research Prioritization Decisions

Introduction

This chapter presents and discusses the evaluation planning cycle, the program evaluation goals placed within that cycle, and issues relating to the planning and review of the evaluation plan. The chapter then presents an example of an evaluation approach integrated within a program planning and implementation cycle. These sections allow the reader to understand the structure of the program cycle in which the evaluation process is placed.

The chapter then focuses on the evaluation planning and development process, including a review of the key issues associated with the evaluation planning and review process, the completion of the evaluation plan and issues relating to the plan review and approval process. The chapter then displays a high-level diagram of the Framework, showing the relationships of the different parts of the Framework. The chapter then discusses evaluation priorities, budgets and opportunities for evaluation consolidation.

Finally, the authors have included an appendix to the chapter providing a more detailed example of an integrated "systems approach" program implementation and evaluation cycle.

The primary goal of this document is to set the stage for providing guidance of what types of evaluations are appropriate in California and to provide guidance to those responsible for preparing and approving evaluation plans, for conducting the evaluations and for reporting the results. The Framework is an evaluation guidance document and is not a prescriptive protocol for conducting evaluations in California. However, the California Public Utilities Commission (CPUC) or another designated organization charged with the responsibilities of developing new evaluation protocols may wish to use this Framework as one of the guiding documents for establishing new evaluation protocols for California's energy efficiency programs.

As with other chapters in this document, the umbrella chapter is written for a wide range of stakeholders within the California energy efficiency program arena. However, because of the topics covered in this chapter and the impact of this chapter on the development and implementation of a statewide evaluation framework, a wide range of affected stakeholders should have a working knowledge of the structure and issues described in this chapter, including:

- Regulatory staff & policy makers responsible for evaluation;
- Evaluation oversight managers & reviewers;
- Program administrators and program implementers;
- Cost-effectiveness and avoided cost analysts;

- Evaluation designers and managers;
- Evaluators for all types of evaluation—impact, metering and monitoring, process, information and education programs, market transformation and market effects, and non-energy effects;
- Statisticians & research data managers involved in evaluations; and
- Portfolio managers.

The Evaluation Planning Cycle

The evaluation processes and roadmaps associated with this Framework are placed within a cyclic evaluation-planning process consistent with the program funding and contracting cycles. This Framework takes no position on whether those cycles should be one, two, three or more year cycles, but does suggest that the evaluation planning cycle be consistent with the program approval and funding cycles. This allows the program evaluation planning process to structure the evaluation plans so that the evaluations are planned, implemented and completed within a program cycle. Later in this chapter we discuss one approach for coordinating the program implementation and evaluation cycles in order to illustrate the need to coordinate the two processes over the same period of time.

For two-year cycles the program evaluation plans would cover all the evaluation efforts that are planned over the two years of the program contract period. Likewise, a three-year program cycle would mean the evaluation planning process cover a three-year period. The evaluation plan would then specify the types of evaluation efforts that are scheduled throughout the program period. The evaluation planning process would take into consideration the need for evaluation results to document the operations and effects of the program, provide feedback for ongoing program improvement, provide information to support energy efficiency portfolio assessments and to help support the planning efforts for future program cycles. For impact evaluations that examine technology-specific impacts (rather than program impacts), the information can also be used to update the DEER. The timelines and processes for these efforts can be part of the evaluation protocol formation process that may be developed under the direction of, and adopted by, the CPUC.

The Framework process does suggests that program administrators plan their evaluation efforts over the duration of the program period. This may take the form of an annual evaluation plan for each year of the program cycle, or a multi-year plan with detailed evaluation efforts presented for each year of the program.

The overall plan for the evaluations of a program should describe the types of studies that are proposed for the program, the justification for why the different types of evaluations are being proposed and why others are not, the timeframe for these studies, a presentation of the CPUC's evaluation goals that each study addresses, and a description of how each goal will be addressed.

These current CPUC Policy Manual[35] evaluation goals are to:

1. Measure energy and peak savings;
2. Measure cost-effectiveness;
3. Provide upfront market assessment and baseline analysis;
4. Provide ongoing feedback and guidance to the program administrator;
5. Measure indicators of effectiveness and testing program theory and approach;
6. Assess the overall levels of performance and success;
7. Inform decisions regarding compensation and final payments; and
8. Help assess the continuing need for the program.

These specific goals may be modified from time to time to support changes in public policy or to fine-tune the goals so that they provide more defined instructions to support the evaluation planning process. However, detailed evaluation plans for individual programs should include the following information, as appropriate for the type of evaluation being conducted:

1. The type of evaluations being conducted, what CPUC evaluation goals it will meet, and the logic and support for the proposed set of evaluation studies and why these have higher priority over other types of evaluation (whenever the evaluation budget limits conducting all of the evaluations that would be desired).
2. A short description of the program(s) being evaluated;
3. A presentation of the program theory;
4. A short description of the expected program effects;
5. A list of the technologies offered by the program with an indication of which technologies will be addressed in the evaluation(s);
6. A presentation of the researchable issues to be addressed in the evaluation(s) and reported in the evaluation reports;
7. A presentation of the overall scope of the evaluation reported by type of evaluation (impact, process, market effects, etc.);
8. A detailed presentation of the evaluation activities to be undertaken, by program evaluation type (impact, process, market effects, etc.) including the monitoring, metering and verification efforts to be employed;
9. A presentation of the sampling approach and sample selection methods for each evaluation activity that includes sampling efforts. Sample sizes should be reported at the technology level and summarized to the program level;
10. A description of how control group, comparison group or non-participant information will be used in the evaluation(s) and in the analysis;

[35] *Energy Efficiency Policy Manual, Version 2*. (* CPUC 2003), page 26.

11. A description of how spillover and freerider effects will be addressed in the evaluation activities and in the data analysis;
12. A brief presentation and discussion of the threats to validity, potential biases, methods used to minimize bias, and the level of uncertainty associated with the sample selection methods and the evaluation approaches;
13. A description of the data handling and data analysis approach to be used to address the researchable issues;
14. A statement of whether the evaluation can be used to update and/or validate specific values, including measure costs, measure life, and measure impacts, within the DEER. This statement should identify the specific DEER technologies and technology characterizations that will be addressed. If the evaluation cannot be used to update the DEER information, the reasons should be cited.
15. An activities timeline with project deliverable dates; and
16. A presentation of the evaluation costs for each type of evaluation (process, impact, metering and verification, market effects, etc.) and the overall costs for all evaluation efforts proposed.

The Framework does not include a description of the process associated with the submission, review and approval of program-specific evaluation plans. However, the Framework is established with an understanding that there will be a formal process for reviewing and approving evaluation plans to make sure that they provide the level of rigor required to support a portfolio assessment and planning process, and to reliably understand the energy resources available through different types of energy efficiency programs, technologies and behaviors. The formal review and approval process for evaluation plans will be developed and described by the CPUC.

Components of the Evaluation Plan

The program evaluation plan is a formal document that clearly presents the evaluation efforts and details the activities to be undertaken during the evaluation. The evaluation plan is a stand-alone decision document, meaning it needs to contain the information needed to allow the evaluation plan reviewer to make a decision regarding if the evaluation plan should be approved and implemented or modified. There are several components of the evaluation plan that are needed for the review process, these include the following components:

- Cover page containing the program name, program administrator, evaluation contractor, date of evaluation plan, and the program tracking number (if any);
- Table of Contents;
- Summary overview of the program and the evaluation effort;
- Presentation of the evaluation goals and researchable issues addressed in the evaluation;
- Brief description of the program(s) being evaluated;

- Description of how the evaluation addresses each of the CPUC evaluation objectives (currently as in the CPUC Energy Efficiency Policy Manual[36] Chapter 6 EM&V Objectives),
  - Where an evaluation plan prioritizes evaluation needs within a constrained evaluation budget such that not all evaluation objectives are met with completed evaluation studies, then the evaluation plan should provide evidence to support its prioritization to include, as appropriate:
    - comparative program evaluation findings,
    - the history of evaluation findings for the type of program, and
    - a discussion of program maturation and stability/changes;
- Brief presentation of the program theory and underlying assumptions;
- Task description of the evaluation efforts;
- Description of the analysis activities and approaches;
- Description of the M&V efforts (impact evaluations only) including:
  - Reference to IPMVP option, if used
  - Description of deviation from IPMVP, if any;
- Description of the sampling rational, methods and sample sizes;
- Discussion of the uncertainty and bias issues associated with the evaluation approach and methods incorporated to minimize bias and uncertainty;
- Evaluation activities timeline;
- Program budget and the evaluation budget; and
- Program administrator and evaluation manager contact information.

The evaluation plan should be written in a style and with enough detail that it can be clearly understood by program administrators, policy makers, and evaluation professionals.

Reviewing and Approving The Evaluation Plan

It is important for evaluations to focus on the key research goals for the types of programs being implemented and to focus research on the types of information needed at various times along the program path. It is the evaluation plan that details the researchable issues that are addressed in the evaluation effort and also presents the methods and approaches that are expected to provide the needed information. Policy makers and regulators need to be assured that the evaluations conducted will deliver the type and quality of information needed. Evaluations that are under-designed can waste valuable resources by not being able to reliably provide the information needed or by providing information that is not needed. Likewise, evaluations can also be over-designed to address issues that are not priority issues or employ research methods that can be substituted with less costly approaches. In addition, there is a need to coordinate the overall evaluation approach so that limited evaluation resources can be focused on the issues of importance across the collection of evaluation activities. As a result, there is a need to have a process for reviewing and approving evaluation plans. While the

[36] (* CPUC 2003).

Framework does not address the organization and structure of how that review and approval process should be provided, it is important to have a process that guides the evaluation research being conducted to meet the overall policy and portfolio assessment needs as well as the program-specific evaluation needs. The evaluation plan review process could be developed and directed to address the following issues in the review process.

- Does the evaluation address the key policy, regulatory and oversight needs for evaluation information? Are there key issues missing? How will the missing issues be addressed? Can the approach be strengthened to better address these needs?
- Does the evaluation plan represent a reasonable approach to addressing the information needs and the researchable issues?
- Is the evaluation capable of providing reliable conclusions on energy and other impacts? Are there key threats to the validity of the conclusions? Are they being minimized given budget constraints and study tradeoffs? Will they be documented and analyzed?
- Are the researchable issues clearly presented and described?
- Are there missing opportunities associated with the evaluation approach that should be added or considered?
- Does the cost of the study match the methods and approaches planned?
- Does the evaluation plan make sense as an independent evaluation study, given the other evaluation efforts planned for the same time period for other programs and for the program portfolio (i.e., or does a different complementary or consolidated study provide a better use of evaluation funds)?

Evaluations to Facilitate an Integrated Planning Process

The timing of evaluation studies can be viewed in the context of the overall strategic cycle associated with program planning, implementation and evaluation. In a strategic framework approach, evaluation results are used to make informed decisions on program improvements and future program designs and offerings. The program implementation cycle is one in which programs are designed, then implemented, and then evaluated. Following the results of the evaluation, programs are re-examined for design changes and then modified so that those design changes result in improved program implementation efforts. This cycle provides for a continuing process of program improvement, so that the programs match available market opportunities and continually improve their cost-effectiveness over time.

The timing of these activities is influenced by several, often-competing considerations. These considerations include:

- Input to policy considerations and decisions;
- Early feedback to program implementers;
- Program lifecycle stage;

- Evaluation data time lags;
- Portfolio planning requirements;
- Evaluation planning requirements and regulatory oversight;
- Program design/solicitation, selection, review, and implementation preparation;
- The value of dispersed timing for evaluations;
- Contract requirements for "pay for performance" programs;
- Market inertia;
- Timing needs for retention, measure life, and technical degradation analyses; and
- Regulatory oversight and review.

Each of these aspects and their relationship to the evaluation process is briefly discussed in the following paragraphs.

Input to Policy Considerations and Decisions

One of the most important goals of the evaluation effort is to provide policy makers and portfolio decision makers with the information they need to identify and select which programs and program portfolios to offer as an alternative energy resource to the people of California. Some policy makers argue that this is the most important reason why evaluations are funded. Policy makers and portfolio managers need reliable evaluation results to support the policy decisions and their associated timelines. Likewise, portfolio managers need information when they are designing the mix of program services to provide and selecting the target markets for these services. Program administrators and evaluators should consider these needs when designing and conducting evaluation efforts.

Early Feedback to Program Implementers

Program managers benefit from evaluation results as early in the implementation process as practical in order to make timely program design and delivery adjustments to improve the cost-effectiveness of their programs. Process evaluation and Measurement and Verification (M&V) activities conducted early in the program cycle can identify potential problems with program operations and operational procedures as well as hardware selection and installation issues, allowing program implementers to correct these problems early on.

At the same time, early scheduling of process evaluation and the M&V efforts may complicate the sample design process. Participant samples for on-site or phone surveys may need to be drawn in stages, before the full participant population is established. If problems are identified and corrected, then follow-up surveys may be required to verify that program changes are effective.

Program Lifecycle Stage

Programs need sufficient time to start up and establish efficient operation. Implementation contractors, subcontractors, and trade allies need time to "get up on the learning curve" to the point where the program is operating at maximum efficiency. The

market may also require stability in program presence and operation before steady-state operation can be achieved. This may be especially important with new programs, but it is also important to allow for a period of operational presence in the market with continuing programs that have implemented significant program design or delivery changes. However, as indicated in the Process Evaluation Chapter, it is also important to have the evaluation team start their work during the early phases of the program design and delivery period to help the program be structured in a way that collects needed evaluation information. Early evaluator involvement can also provide, where appropriate, early reviews of program designs and intended operational systems and procedures and allow for early implementation feedback so that needed program changes can be made as early as possible.

Established and continuing programs may not require start-up periods at the beginning of each program cycle. For these programs, unless there are major changes in the program design and operational procedures, the timing of the process and M&V studies may be less critical. Indeed, for continuing programs without major changes in program design and operational procedures, evaluation during each program cycle may not be an efficient use of resources.

Evaluation Data Time Lags

For some programs, especially small programs or new programs, there is a natural time lag between the completion of a program and the completion of the evaluation study. When a program is small enough or new enough that the evaluation needs to obtain information from participants entering the program during the end of the program cycle, evaluation results may be provided at a later time compared with other evaluations in which the program has enough participants to complete the evaluation efforts earlier. In these cases, the evaluation results may not be available until during or after the planning process for a future version of the program.

Portfolio Planning Requirements

Planning a portfolio of programs requires an understanding of the overall goals of the portfolio, the technical and market potential of various measures and delivery mechanisms to deliver energy and demand savings, the cost-effectiveness of various delivery mechanisms, and the risk and uncertainty associated with various technologies and program approaches. Evaluations can provide much information to support this process, provided the information is delivered in a timely manner and in a format compatible with the portfolio analysis process. Market saturation studies, which inform technical and market potentials, should be conducted on a schedule that provides timely information to the planning process.

Portfolio planning can be considered an ongoing process. Portfolio objectives can be updated with the best available information during each planning cycle. Portfolio objectives will likely change as policies change, markets change and the saturation of

energy efficient technologies increases. The portfolio plan can be viewed as a "living document," subject to continuous refinement and revision.

Evaluation Planning Requirements and Regulatory Oversight

Evaluation studies need to be planned from a portfolio perspective, considering the portfolio objectives, risks, and evaluation data needs. Overall evaluation goals and objectives, evaluation priorities, and program design issues should be addressed prior to planning individual program evaluation studies. Issues such as evaluation resource allocation, evaluation study consolidation, the need to evaluate a particular program during a given program cycle, and evaluation scheduling should be considered prior to planning individual evaluation studies. After this, time needs to be scheduled for the processes of regulatory review and approval, selection of evaluation contractors, and development of detailed evaluation plans. The schedule of these activities will likely be defined by the prevailing administrative structure.

Program Design/Solicitation, Selection/Review and Implementation Preparation

Timely introduction of programs into the market and smooth, uninterrupted operation of continuing programs requires sufficient time for program design/solicitation, selection/review, and preparation for implementation. (Implementation preparation includes such time-consuming activities as contracting, working out informal arrangements with trade allies, and developing program materials and processes). This means that timing of these activities and the evaluation activities used to support policy decisions needs to be carefully planned if the process is to operate with optimum information, consistency, and efficiency.

The Value of Dispersed Timing for Evaluations

A large portfolio of programs launched with concurrent start and end dates can create large "spikes" in the workloads of various entities, including portfolio administration staff, contracting entities, program implementers, evaluation contractors and other stakeholders. This can lead to inefficiency and, sometimes, efforts that have a higher probability for error due to strain on the resources available. If the program duration is fairly long, the lull between program evaluation contracts can create difficulties for evaluation firms who must manage their firms to match the ebb and flow of the evaluation market. These could discourage the most competent firms from remaining in the field, lowering the availability of high quality evaluators when they are needed.

Contract Requirements for "Pay for Performance" Programs

The contract requirements for "pay for performance" programs such as the Standard Performance Contract (SPC) program or other programs utilizing an Energy Services Company (ESCO) performance contract may require evaluation results on a time line different from the program or portfolio planning process. The same issue may also apply to incentives paid to utilities under a resource acquisition or procurement scenario. The

program planning, implementation and evaluation cycle may not match the contracting requirements for payment levels based on ex-post or net evaluation results.

Market Inertia

Many information and market transformation programs take time to work due to significant market inertia. The timing of the effects evaluation needs to consider the likely response time of the market to the program intervention. Market effects evaluations may need to be conducted after multiple program cycles to provide enough time for the programs to overcome inertia in markets in which they operate

Timing Needs for Retention, Measure Life, and Technical Degradation Analyses

Measure retention studies may need to be conducted at several points in time years after the program has ended, due to the dynamics of building remodeling, occupant turnover, operations and maintenance practices, business investment decisions, new technology entries into the market, and measure efficiency degradation. These studies need to be conducted periodically after there is a reasonable expectation of removal, failure or degradation of a significant fraction of the measures.

Oversight and Review

Not only evaluation study plans, but also progress and reports may require some level of oversight and review. For example, evaluation study results may need to be reviewed for accuracy or their ability to support the portfolio planning process. Time to conduct these reviews should be included in the overall process.

Example of an Integrated Evaluation Timing System

The evaluation timing system described in this section explores an example of a method to integrate program planning, implementation and evaluation into a continuous cycle of activities that provide timely feedback of information into the overall strategic planning process. The overall strategic planning process in this example has been divided into five major cyclic program activities, including:

1. Goal Setting - Updating and Potential Analysis;
2. Portfolio Analysis - Sector and Program Priorities;
3. Portfolio/Program Design, Selection, Review and Approval, and Preparation for Implementation;
4. Program Launch Preparation, Overall Evaluation Planning and Regulatory Oversight; and
5. Program Implementation, Evaluation, M&V and Market Assessment & Ongoing Regulatory Oversight.

These activities and their possible placement in this example are discussed below.

Activity I: Goal Setting - Updating and Potential Analysis
Activity I consists of establishing the high-level goals for the efficiency portfolio and estimating the achievable potential for the efficiency resource. Overall energy efficiency policy goals (such as targeted percentage reduction in per capita energy use) are restated in terms of specific energy and demand reduction targets. Overall policy goals are reconciled with load forecasts and other information on utility energy and capacity requirements.

Once the specific energy and demand goals are defined, the potential analysis helps refine the goals based on the overall availability of the conservation resource within technologies, market sectors and geographic area. Technology and market performance information from program evaluations, market saturation studies and other overarching studies are incorporated into the efficiency resource potential estimates. Information is also assessed regarding the energy and demand goals, potential analysis, and other policy objectives, such as geographic and sector equity issues.

Activity II: Portfolio Analysis – Sector and Program Priorities
A preferred portfolio of energy efficiency programs for the upcoming program cycle is defined based on the goals and potentials established in Activity I and further review of evaluation data on program impacts, process and market effects, and past overarching studies. Lessons learned from current and past program efforts along with efficiency potentials by technology and market sector from Activity I are used to define a portfolio of programs that meet the overall efficiency resource goals. The results of the portfolio analysis are presented for public comment and regulatory review. Preliminary evaluation planning and budgeting may also take place during this activity.

Activity III: Portfolio/Program Design, Selection, Review and Approval
Designs for specific programs meeting the portfolio goals defined above are developed and/or solicited and selected. The proposed portfolio of programs is finalized and presented for public comment and/or approval by the responsible entity.

Activity IV: Program Launch Preparation, Overall Evaluation Planning and Regulatory Review
Once the program portfolio is designed, selected and approved, the programs are prepared for implementation. Contracts (if needed) are negotiated, trade allies and key stakeholders are notified, and materials and internal processes are developed to prepare for program introduction. Concurrently, an overall evaluation planning process takes place. This process establishes overall goals and objectives for the program evaluations, sets evaluation priorities based on perceived risks to achieving the portfolio savings objectives and other objectives, and addresses program design issues related to evaluation. Issues such as evaluation resource allocation, evaluation study consolidation, the need to evaluate a particular program during a given program cycle, and evaluation scheduling are addressed during the overall planning process. The overall evaluation plans will need to be reviewed by an appropriate body or bodies to assure that they will

meet the information needs of policy makers, portfolio managers, program administrators, and program implementers. This examination considers not just the information needed about the planned program effects, but also the information needs associated with portfolio planning and technology targeting, including such tasks as updating the DEER and maintaining accurate persistence estimates.

Activity V: Program Implementation, Evaluation, M&V, Market Assessment and Ongoing Regulatory Oversight

After the overall evaluation plans have been approved, more detailed evaluation plans must be developed and evaluation professionals selected to conduct the studies. The overall evaluation plan developed in Activity IV will likely schedule new or redesigned programs to initiate early process evaluation coordination and support activities as well as M&V planning during the initial program start-up period. These early actions can help assure that program designs support the evaluation function, in addition to helping establish the platform from which early evaluation feedback can be initiated. These early actions can also lead to improved database designs that help support the evaluation efforts. Following these early efforts, the evaluations will then likely swing into standard evaluation activities after the program is up and running, but early enough in the program cycle to provide feedback and corrective recommendations to program implementers in time for the program to benefit from those recommendations. Early impact evaluation activities to support program progress tracking consist primarily of measure installation tracking and verification combined with ex-ante savings estimates by measure. Adjustments to ex-ante savings estimates may be made based on early issues identified during M&V activities.

The full net impact evaluation analysis proceeds according to the schedule laid out in the program evaluation plan. Ex-post savings by measure and/or program are estimated, and the final program impacts are estimated based on the final program accomplishments. Assumptions underlying the efficiency potential analysis can then be updated based on the full net impact analysis. These data then feed back into the goal setting and potentials analysis activities, and the cycle repeats to allow for an integrated planning process.

The above example of a coordinated and integrated system is one example of how this process can be structured. This document makes no recommendations on the organizations responsible for these efforts, but presents this example of how a system might function to allow for a coordinated process in which the steps that need to be integrated are established along a single timeline.

The steps and feedback paths for this type of an integrated planning process are shown in Figure 5.1.

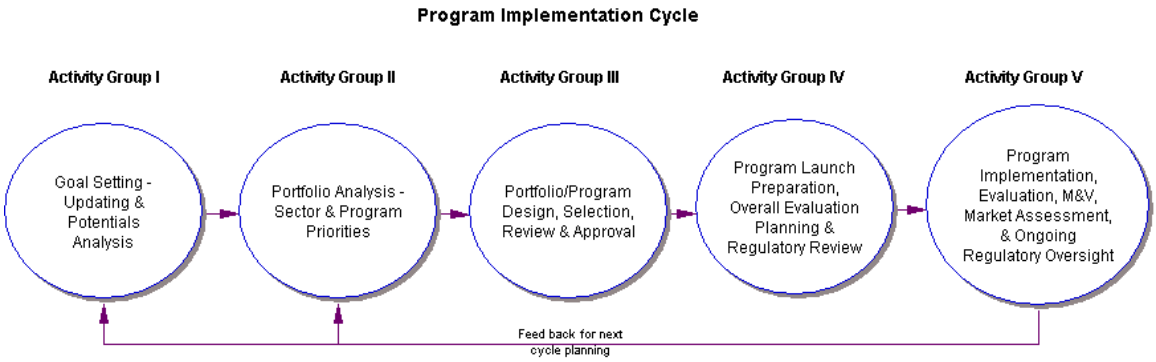

Figure 5.1: Program Implementation Cycle

Note: See the appendix to this chapter for a more detailed example of this process with key steps and dates incorporated into a sample timeline.

On top of this cycle of repeating activities, longer timescale analyses of market effects and measure persistence can be conducted to provide important information to the integrated planning process. A discussion of the issues and the timing of these efforts are described below.

Program Duration

Programs should be conducted over a period of time needed for the programs to establish themselves in the market and reach stable operations prior to conducting the program net impact evaluation. In addition, this period should allow for program improvements to be made to the programs, as a result of the evaluation recommendations and finding, while the program is still operational.

Timing of Evaluation Activities

Ex-post measure savings analysis and net-to-gross analysis should be completed prior to program completion to allow the evaluation information to feed back into the planning process for the next program cycle.

Multiple Offset Program Cycles

Separate program cycles with offset start dates can be considered to provide opportunities for new program proposals on a more frequent basis and to level the load for all parties involved in planning, selecting, developing, implementing and evaluating programs. "Off-cycle" programs may be targeted to fill particular gaps in the portfolio and/or focus on innovative programs, pilot programs, demonstration programs or emerging technologies

The interactions between two separate program tracks, with a two-year offset between each track are shown in Figure 5.2. Note that the second track program planning cycle

can take advantage of impact evaluation information from the first track in the overall planning process.

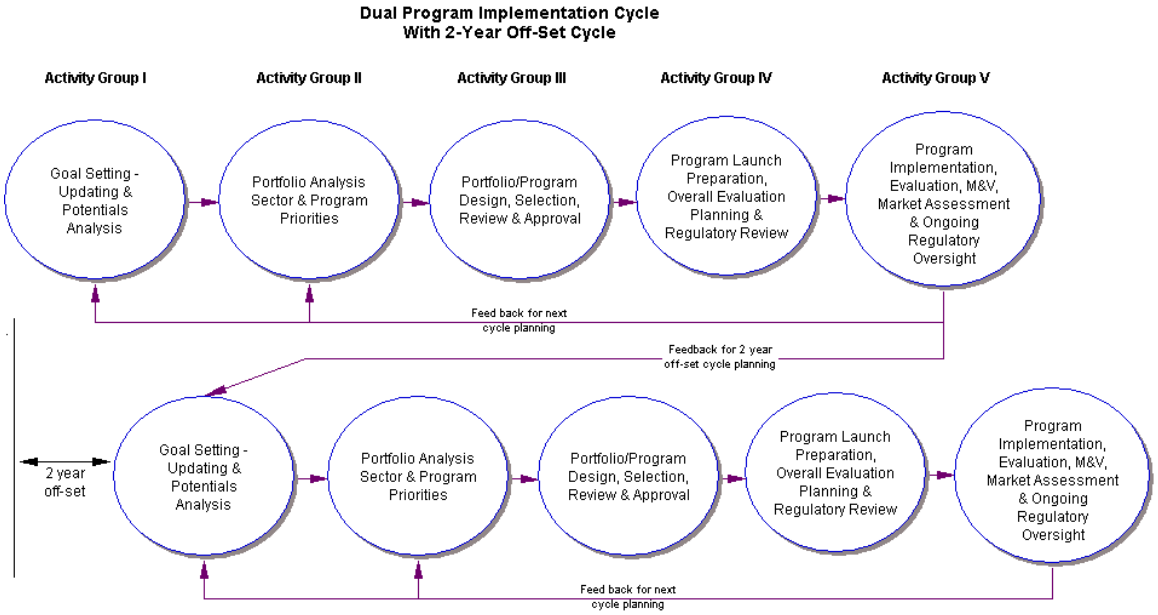

Figure 5.2: Two-Track Program Implementation Cycle

Impact Analysis for Final Contract Purposes

Final program results for the purpose of settling contract terms and determining incentive payments should be derived from the ex-post analysis and final program accomplishments shortly after program completion. These efforts can be held separate from the impact analysis conducted to inform the program planning process.

Program Planning Period

A coordinated effort between the evaluation, efficiency resource potential and portfolio analyses will be required to complete the necessary analysis in time for it to impact the next integrated planning effort to guide the next cycle of programs. For programs that are contracted, solicitation and contracting procedures may need to be streamlined to ensure that an integrated planning process can be accomplished.

Detailed examples of the timing of an integrated planning process are presented in the appendix to this chapter.

Roadmaps and the Evaluation Planning Process

This document contains several chapters specifically designed to support the evaluation planning process. Also, each chapter discussing a specific type of evaluation (impact, process, market effects, etc.) contains a planning decision guidance "roadmap" that

allows the reader to see when and under what conditions a specific type of evaluation may be desired. These chapters provide valuable guidance concerning the CPUC's expectations for the rigor, quality and focus of the evaluation efforts and the associated activities needed to support the evaluation function.

In summary, the chapters presenting the evaluation roadmaps and decision systems cover the following evaluation topics:

- Impact evaluations;
- Measurement and verification efforts (M&V);
- Process evaluations;
- Market transformation evaluations;
- Evaluations of information and education programs;
- Non-energy effects evaluations;
- Sampling requirements and methods;
- Uncertainty and bias issues associated with the evaluation efforts; and
- Designing evaluations with cost-effectiveness analysis needs in mind.

The information presented in these chapters is provided to help guide the evaluation planning and implementation process for California's energy efficiency programs. These chapters will help the evaluation planner make decisions about, and develop the methods and procedures for each evaluation. These chapters will help guide the evaluation planning process and the activities associated with conducting the field efforts, analyzing the data collected, and in developing the evaluation reports.

The program-specific evaluation planning process should move through each type of the program evaluations presented in this document according to the type of program being implemented. These efforts should be planned to take place within the contracted program implementation period so that the evaluation results can be used to help improve the program to the extent possible within the implementation period, to document program accomplishments and net benefits, and to inform the portfolio assessment of which programs to offer in the future. The evaluation planning and implementation process should match the evaluation needs associated with the goals of the program. This means different types of programs should plan for different types of evaluations efforts. These efforts and the key policy and regulatory issues associated with each of the types of programs and program evaluations are discussed below:

All Programs

There are key parts of the Framework designed to feed the evaluation planning process for all types of programs and all types of evaluations. These chapters focus on sampling methods and approaches to support the evaluation effort, on having the evaluation planning process focus on, and address, uncertainty and bias issues associated with each evaluation effort, and on designing evaluations that inform the program-level and the portfolio-level cost-effectiveness analysis needs.

The key policy-related, regulatory and evaluation plan review and approval issues regarding these chapters include:

- Is the sampling plan representative of the population served?
- Is the sampling plan able to support the evaluation policy objectives?
- Is the sampling plan objective and unbiased focusing on key evaluation needs?
- If the sampling plan is biased, how can it be improved?
- If the sampling plan is biased, how do the biases harm the evaluation's objectivity or influence our understanding of the program's effects?
- Are there threats to the validity of the evaluation results that are incorporated into the evaluation design?
- Can the evaluation design be modified to limit threats to the validity of the evaluation results?
- Have or will the instruments proposed be examined for potential threats to validity? Is there an estimate of the magnitude of these effects or to the degree that these are occurring? Are methods proposed to address these effects? Are cost justified mitigation methods proposed?
- Are the threats to validity or biases described in the evaluation plan so that policy makers understand these threats or biases and the limitation these provide to the use and usefulness of the evaluation data and to the findings developed from the evaluation?
- Does the evaluation provide the data needed to inform a cost-effectiveness assessment of the program and to compare this program with others?

Programs with Impact Goals

Programs with energy impact goals should focus their evaluation plans to include an energy impact evaluation and a process evaluation over the program's implementation cycle. Evaluation planners for these programs will want to read the Framework chapters on impact evaluations, on conducting measurement and verification (M&V) efforts, and on conducting process evaluations.

The key policy-related, regulatory and evaluation plan review and approval issues regarding programs with impact goals include:

Impact Evaluations

- If no impact evaluation is proposed, is this well-justified based upon documentation of the following:
  - Prior net impact evaluation history for this program, date of last impact evaluation, and stability/precision of results,
  - Comparability of program prior net energy impacts with other similar programs,
  - Size of program impacts and program budget and current evaluation budget limitations,
  - Maturation of the program, and

  - Similarity of program with past programs for which net impact evaluation results are available (technology mix, delivery mechanism, customer mix).
- Is the evaluation using appropriate analysis methods considering the type, size and timeline of the program being evaluated and for the data that is available or can be collected? (See Framework Appendix C for an example assessment.)
- Will the evaluation reliably report gross and net impacts associated with this program?
- Will the evaluation reliably report the gross and net impacts associated with the program and program technologies and can these be used to update the DEER for impact, cost and measure life?
- Are the verification efforts adequate to document service and to support impact assessments?
- Are the M&V efforts designed to support the net impact assessment needs so that impacts are accurate for this program, as implemented?
- Are the impact and M&V methods and approaches appropriate to support the policy formation and regulatory needs?
- Will the evaluation test the program theory's causal relationships and assumptions regarding program events and activities, and program results so that the theory can be supported or modified?

Process Evaluations

- If no process evaluation is proposed, is this well-justified based upon documentation of the following:
  - Prior process evaluation history for this program, satisfactory or good process evaluation findings within last three program cycles,
  - Program has not changed since last process evaluation was conducted,
  - Size of program impacts compared with the program budget and the current evaluation budget,
  - Maturation of the program,
  - Similarity of prior process evaluation results for past programs for which process evaluation results are available, and
  - Program impact and cost-effectiveness results have shown to be both close to expectations and similar to other programs offering similar services.
- Is the process evaluation designed to be able to identify weakness in the program's designs, methods of operation or implementation procedures that limit the ability of the program to obtain cost-effective impacts?
- Will the process evaluation identify program methods and procedures that can be modified to increase energy impacts, improve customer satisfaction, increase enrollment rates, or improve market appeal?
- Will the process evaluation document the program's methods of operation so that they are well-understood by policy makers?

Information and Education Programs

The information and education program evaluation plan should contain a program effects evaluation and a process evaluation over the program cycle (unless they have energy

impact or market transformation goals, see chapter 5 for details). Evaluation planners for these programs will want to read the Framework chapters on evaluating information and educational programs and on conducting process evaluations.

The key policy-related, regulatory and evaluation plan review and approval issues regarding programs with information or educational goals include:

Program Effects Evaluations

- Will the evaluation document and, to the extent possible, quantify the direct effects of the program on participants or the recipients of the program's services?
- Will the evaluation test the program theory's causal relationships and assumptions regarding program events and activities, and program results so that the theory can be supported or modified?

Process Evaluations

- Is the process evaluation designed to be able to identify improvements in the program's methods of operation or implementation procedures that can improve the ability of the program to accomplish its informational or educational goals or increase the intended effects from these efforts?
- Is the process evaluation designed in a way to obtain the information to allow it to identify new methods and procedures, or identify changes to current methods and procedures that could increase knowledge gained, the extent of the information used, or other improvements that would help the program more cost-effectively achieve its effects goals?
- Is the process evaluation designed such that it can obtain the information and identify and recommend changes to the program that might increase the level of desired behaviors?
- Will the process evaluation document the program's methods of operation so that they are well-understood by policy makers?

Market Effects or Market Transformation Programs

Market transformation programs should plan for a baseline study or a baseline analysis approach that can serve as an evaluation baseline, a process evaluation and, at some point in time after which market effects can be expected, a market effects evaluation. Evaluation planners for these programs will want to read the Framework chapters on market transformation evaluations and on conducting process evaluations as well as the 2001 publication on evaluating market effects and market transformation programs[37].

The key policy-related, regulatory and evaluation plan review and approval issues regarding programs with market transformation goals include:

[37] See *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*. (* Sebold, Fields et al. 2001).

Market Effects Evaluations

- Does the baseline study used to support the market effects evaluation provide an adequate presentation of the way the market operates relative to the markets that are targeted?
- Is the baseline information based on research that is conducted with enough reliability that the error bands around the baseline market data can be compared with the results from the market effects evaluation and used to determine market movement?
- Is the baseline study capable of establishing a measurable baseline from which market effects metrics influenced by the program can be tracked and measured for program effects?
- Will the evaluation reliably (to the extent possible) document the net changes that have occurred in the market as a result of the program?
- Does the evaluation have a reasonable approach for disaggregating market effects caused by the program with the market effects caused by non-program events so that net market effects can be estimated?
- Does the market effects evaluation have a reasonable approach for forecasting how the current market effects will continue in the future?
- Will the evaluation test the program theory's causal relationships and assumptions regarding program events and activities, and program results so that the theory can be supported or modified?

Process Evaluations

- Is the process evaluation designed to be able to identify weakness in the program's designs, methods of operation or implementation procedures that limit the ability of the program to change the way markets operate to be more energy efficient?
- Will the process evaluation identify program methods and procedures that can be modified to increase market effects?
- Will the process evaluation document the program's methods of operation so that they are well-understood by policy makers?

Programs That Rely On Non-Energy Effects To Achieve Their Goals

Some programs may rely on non-energy effects to gain participation and achieve their energy savings goals. For these programs there may be a desire to conduct a non-energy effects evaluation. These evaluations can be approved for funding if they pass specific conditions regarding the need and use for the evaluation findings. Evaluation planners for these programs will want to read the Framework chapters on conducting evaluations on non-energy effects.

The key policy-related, regulatory and evaluation plan review and approval issues regarding programs with non-energy effects impacts include:

- Is the non-energy effects evaluation needed to more accurately understand or to increase the energy effects of the program?

- Does the CPUC have a policy supporting the spending of Public Goods Charge funds to document the specific non-energy effects targeted?
- Can the evaluation design reliably identify and quantify the targeted net non-energy effects?

Overview of the Framework

This document and the associated roadmaps provide a structured approach to planning and conducting evaluations of California's energy programs. Figure 5.3 provides an overview of the process and allows the reader to see the components of the evaluation framework structure and linkages between the components. These steps and processes are presented and discussed in detailed in the associated chapters of this document.

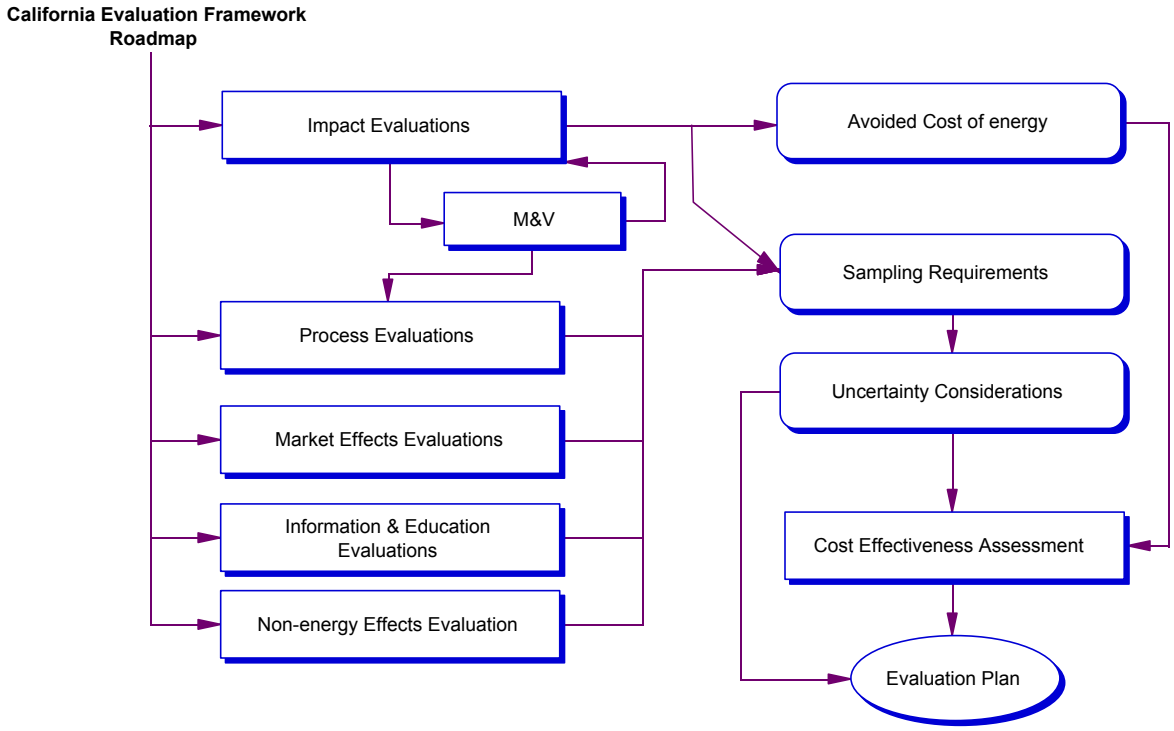

Figure 5.3: Overview of the Evaluation Framework

As discussed earlier, the Framework is an evaluation guidance document and is not a prescriptive evaluation protocol. While the Framework methods and approaches are not mandatory, they are developed to help guide the evaluation planning process for California's energy efficiency programs. This document is intended to help guide the evaluation process to provide reliable and comparable evaluation results across the different types and sizes of programs, and to focus on and address specific issues related to the evaluation planning and implementation process.

Determining Evaluation Priorities

Evaluation funds are generally limited to a small fraction of the total program budget and this small fraction may not be adequate to provide all the evaluation results needed to make well-informed policy or program decisions. Therefore, there is almost always a need to carefully allocate evaluation funds among the various potential evaluation activities that must be pursued. This frequently raises a question of what relative emphasis to place on each of the two key evaluation functions (i.e., documenting and measuring program effects versus understanding/improving program processes) and on each of the types of evaluation, as well as the question of which particular programs should receive the most emphasis.

There is a discussion of a systematic approach for allocation of evaluation resources in Chapter 12, Uncertainty. This approach is recommended, at least for the consideration of resource allocations between various program impact evaluations.

A program-level allocation of evaluation resources should be examined when program-level budgets are set. However, there are also many qualitative decisions that need to be made, or decisions that are difficult to quantify.

The decision on how to allocate evaluation resources within a program-specific evaluation budget should be made after a careful consideration of the specific program and the policy decisions being confronted. Some examples of key factors that should be considered include:

- What are the most important information needs of the key policymakers/regulators associated with the program?
- Where are the programs in terms of their implementation cycle (e.g., "new" programs generally need a relatively higher level of process evaluation emphasis than existing, well-established programs)?
- What key decision points are approaching and on what time schedule?
- What level of evaluation information already exists and where are there important information gaps?

In short, there are no hard and fast rules about relative allocations of evaluation resources to various functions and types of evaluation. Those decisions must be made on a very situation-specific basis. However, there are several key information needs that represent the main purposes for conducting evaluations in California. Evaluation resources should be focused so that these needs are satisfied. These include:

- The need for accurate net program-level energy and demand impact information that reliably indicates the resources gained as a result of spending public dollars for the program funded,
- The need to understand the net energy and demand impacts associated with the individual technologies and behaviors adopted, as implemented and used,

- The need to accurately compare the cost-effectiveness of the program with other programs making up the portfolio of programs,
- The need to understand and have measurements of the program's key effectiveness indicators,
- The need to understand the program theory supporting the structure, organization, operations and activities of the program and to test the program theory and the theory's underlying assumptions,
- The need to understand the market and the key market factors effecting the program's ability to achieve energy and demand goals and achieve the program's operational and market-related goals,
- The need for an understanding of the program and its operational characteristics and methods to identify those that can be improved, and to understand those that should be used by other programs,
- The need for accurate information on the non-energy program goals achieved by the program and the effects of those achievements,
- The need to understand if the program has reached its best cost-effective performance or if it can be expected to improve its cost-effectiveness through program design, operational or implementation changes,
- The need for information that helps policy makers determine the ability of the program to produce energy and demand savings in the future,
- The need to document the activities and results received in exchange for Public Goods Charge or procurement funds, and
- The need for information to help program managers understand the results of their program's efforts and to provide timely feedback to managers to help them improve their programs.

These above needs represent the primary reasons for conducting evaluations in California and provide guidance to the evaluation planning process. These needs are demanding on the evaluation effort and on the resources that can be allocated to the evaluation function. In some cases these needs may not be met within a single program implementation and evaluation cycle. As a result, the evaluation planning efforts may need to address the timeline over which the needs can be addressed, and focus the current program-cycle evaluation efforts on those needs that are identified as the highest priority for the current cycle.

Recommendations for Setting Evaluation Budgets

The budgeting of evaluations is a subject of interest to all stakeholders in the program planning, implementation and evaluation cycle. Estimates of where evaluation budgets should be set have in the past generally ranged from about ten percent to a low of about two percent of the program budget. Budgets for the 2002 - 2003 programs averaged just over four percent. A great many of these evaluation plans indicated that the lack of evaluation resources significantly restricted the ability to reliably estimate program effects, especially at the technology level. Many of the 2003 – 2004 program evaluations relied on deemed savings estimates or adjustments to deemed savings. In addition, many

program administrators indicated that they had insufficient resources to conduct process evaluations or to obtain needed baseline information for their programs.

The most important consideration for establishing evaluation budgets is the need for reliable estimates of energy effects and program potentials to be used for making energy resource policy decisions and for identifying cost-effective energy supply options. The second most important reason for evaluation is to improve the operations and cost-effectiveness of energy programs. Evaluation budgets that are set too low to accomplish these goals are of little value to any stakeholder. Evaluation budgets that are set higher than needed for these goals may be a waste of valuable resources.

There is no single specific percentage of a program's budget that should be allocated for the independent evaluation process. Evaluation budgets for small pilot programs, testing new designs or delivery concepts, may need to be set at a level higher than the program costs in order to collect and analyze the information needed to determine if the program should be continued or expanded. In these cases it would not be unusual for a pilot test program to have a program budget of $100,000 and an evaluation budget of $100,000. At the other end of the scale are large statewide programs offering technologies and services that have been evaluated in the past, offered in well-understood environments in which the energy impacts and operational procedures are well-documented, and in which the procedures are working smoothly and efficiently. In this case a $15,000,000 program may need only two percent of the program budget to verify what is already understood. Most energy programs can be pegged somewhere between these two extremes.

One of the key questions that needs to be addressed in establishing a program-level evaluation budget is found in the answer to the question: What are the energy supply decisions that must be made in relation to the performance of this program and what is the supply risk of making a wrong decision relative to this program and these technologies? This is a decision for policy makers and for portfolio managers charged with providing an overall portfolio of low-cost and reliable energy supply and demand-side resources.

In establishing an evaluation budget policy, one must also look at the cost of the program and the number of years over which the program may be implemented. Some may suggest that a $100,000 program is too small to worry about conducting a rigorous or reliable evaluation. Yet if that program is funded for a period of ten years, the program becomes a $1,000,000 program that is never evaluated. Policy makers are then placed in a position to defend the spending of a million dollars without an understanding of the program's effects or potential effects.

At the end of the day, and after all arguments have been aired, policy makers still need to allocate a budget to the evaluation effort. Adequate budgets are generally lower than ten percent but higher than four percent. Budgets can stay toward the lower end when prior evaluations are drawn upon and circumstances are explored to determine the necessity for repeat evaluations.

In order to establish an evaluation framework that incorporates a wide range of evaluation needs into a program cycle evaluation planning process, the evaluation budget needs to be in the range of six to eight percent of the total portfolio budget. However, this budget should not be equivalently distributed across the programs. Rather, the budget should be allocated at the program-level based on the need for reliable information after careful examination of past program evaluations, an assessment of the rigor (reliability of findings and levels of uncertainty) of those evaluations, and a understanding of the current mix of programs and technologies, and their associated mix of targeting and delivery practices. This approach will help assure that evaluation resources are wisely used and that the information gained from the evaluation improves the ability of policy makers and energy supply planners to make informed decisions.

Evaluation Spending Priorities

While it is important to understand the budgets that are typically considered appropriate for energy efficiency programs, there are evaluation priorities that can serve to inform decisions about how those funds are distributed among the various evaluation goals.

1. As a general rule of thumb, evaluation efforts for resource acquisition programs first need to focus on establishing reliable net energy impact estimates for both the program as a whole and for the technologies addressed in the program. However, it is also important to understand the key factors that drive net savings at the technology level for different types of participants. For information and education programs, it is important to understand the effects that are being achieved by these programs, to be able to identify what information or education efforts provide these effects and the causal relationships between the two. For programs that are designed to change how markets operate, the evaluation should first focus on being able to identify, monitor and track the net changes in the market being caused by these programs and the causal linkages between program actions and market effects. In addition to knowing the impacts/effects from program efforts, it is important to know that the investments are wise investments. This means that providing the information to assess the program's cost-effectiveness and cost-efficiency of the program also needs to be a top priority of the evaluation focus.

2. Maintaining high participant satisfaction scores and reducing barriers to participation represents the second tier of efforts on which evaluation resources can focus. California is dedicated to providing customer-focused services that are designed to achieve high levels of customer satisfaction. To maintain these levels, it is necessary to understand the conditions that drive participant satisfaction and identify methods for increasing customer satisfaction scores. California energy efficiency programs should continually strive for the highest levels of participant satisfaction while obtaining the highest level of net benefits (cost-effectively). To accomplish this goal it is important to understand how energy efficiency programs, products and services need to be structured and

delivered to achieve cost-effective energy resources or achieving other program goals while maintaining high participant satisfaction. Additionally, programs must be aware of the barriers that keep customers from becoming participants and design programs that are able to reach and enroll eligible participants within the market segments that are targeted.

3. The third tier of evaluation efforts focus on identifying ways to improve the cost-effectiveness and operational efficiencies of the programs in order to evolve the program toward its most cost-effective or resource efficient mode of operation. These efforts can focus on a wide range of activities but typically involve the examination of the program designs and methods of operation and included investigative activities that focus on improving a wide range of program characteristics, including; program design, management, administration, staffing, targeting, marketing, processing, service delivery, reporting, and other efforts.

4. Finally, program evaluation resources need to document what the program is accomplishing in exchange for the public dollars spent. In the end, all energy efficiency programs funded by the Public Goods Charge and procurement dollars are accountable to the people of the state of California. The documentation function associated with providing energy efficiency programs is provided by several functions, including administrative tracking and program progress reporting. However, in addition, it is the evaluation studies and reports that can serve as the independent source of information on what is provided and what is achieved as a result.

Consolidating Programs for Evaluation

The process of grouping programs together or consolidating the evaluations of multiple programs may provide opportunities to increase the efficiency of the evaluation effort or the ability of the evaluation to produce more reliable results (e.g., decrease uncertainty levels). Coordination or consolidation options that may improve evaluation efficiencies include:

- Overlapping instrument and methodology development to serve multiple programs;
- Overlapping or complementary sample selection (which may increase ability to obtain sample points for specific technologies or types of customers/facilities, etc.);
- Overlapping field research or data collection efforts; and
- Data sharing or analysis grouping to increase reliability of findings.

In addition, consolidating of the evaluation function can, in some cases, increase the level of interest in the project from firms that may not typically bid on small evaluation projects, thereby increasing the level of competition for the study.

Consolidating the evaluations of similar types of programs or consolidating evaluations by their evaluation efforts has significant potential for increasing evaluation efficiencies as well as for providing additional depth, rigor of measurement, comparability, and value from the study. For example, it is possible to consolidate impact studies across different types of programs. Likewise, process evaluations can be consolidated, especially when they are testing similar types of program designs, operations or systems. Field metering and verification efforts can also be consolidated across a wide range of program types.

Consolidation of the evaluation effort can also be aggregated by market sector. For example, a group of residential programs can be consolidated. Likewise, a group evaluation of programs focusing on industrial or agricultural markets may be practical.

Evaluation consolidation may also be structured by program size, especially for small programs. Small programs are often resistant to spending evaluation dollars that may be needed to keep the program running. However, these programs may be consolidated to obtain an evaluation budget that can address a portion of the evaluation needs. In this case, it may also be possible to structure the evaluation over multiple program implementation cycles and focus each cycle on a set of evaluation objectives.

Finally, another approach to consolidation would be for multiple programs that are offered by the same implementer, so that the program evaluation efforts, across multiple programs, would deal with the same organizational structure. The value of this type of consolidation is less certain, because the evaluation contractor or contractor teams may need to demonstrate expertise across a wider range of program types, technologies and evaluation approaches with relatively little overlap.

Consolidating the evaluation functions across multiple programs can also add value to the evaluation effort. Examples of increasing the value of the evaluation via consolidation include:

- Information and educational programs targeting schools and teachers implemented by different organizations, perhaps using different strategies and approaches that can be compared;
- HVAC tune-up programs offered across the state so that the same or similar services can be evaluated across different markets, different administrators, different size units or different targeting and delivery approaches;
- Multi-family programs that must overcome similar barriers but may do so in different ways so that different approaches, incentive levels, or levels of services can be compared;
- Programs that target mobile homes but have a mix of covered technologies or service designs that can be assessed relative to other programs serving the same market; and

- Small commercial lighting programs that have a similar mix of technologies but may focus on different sectors within the commercial market or may have different approach and delivery strategies that can be compared.

While there are often benefits associated with consolidating evaluation efforts across multiple programs, it is not always advisable. Despite its promise, consolidation often faces obstacles. These may include:

- Programs that are at different points in their implementation cycle, reducing some of the potential benefits of consolidation;
- Grouping evaluations into too large a group so that only a limited number of evaluation firms can handle the load at a given point in time, thereby restricting competition and giving advantage to large firms;
- Grouping evaluation into too large a group so that the study requires small evaluation firms to group and subcontract pieces of the work, coordinating across complex teams of evaluation professionals;
- Timing issues in which the evaluation funds are scheduled to be available at different points in time reducing the potential for consolidated efforts;
- Ability of the evaluation plan to be developed and approved across different programs within the timeline allowed for these efforts;
- Implementers' concerns over loss of focus of the evaluation on their program;
- Implementer's concerns over the potential for unfavorable or inappropriate comparisons;
- Programs with characteristics that need to be evaluated by firms with special expertise, indicating that an evaluation may need a different evaluation team than the other programs; and
- Programs that have been evaluated in the past for which special research approaches or methods have been developed that may not apply to other programs.

Appendix to Chapter 5

An Example of an Integrated Planning Cycle

This appendix provides two examples of how the program design, implementation and evaluation cycles can be structured to provide an integrated planning approach.

Example 1 – Three Year Program Cycle with Integrated Planning, Solicitation and Evaluation Process.

An example of a three-year program cycle with an integrated planning, solicitation and evaluation process is shown in this example. An example describing the individual steps within the five activities described in Chapter 5 is as follows:

Activity I. Goal Setting – Updating and Potential Analysis
Activity I consists of two steps. These steps make up the process by which the high-level energy efficiency portfolio goals are established and updated, based on overall policy objectives and analyses of the maximum achievable energy efficiency resource. Goals and resource potential baselines, once established, will undergo periodic evaluation, review, and updating to provide timely policy information related to the accomplishments of the portfolio of energy efficiency programs.

1. **Goal Setting and Updating.** In this step, the high-level goals for the energy efficiency portfolio are translated into specific energy savings and demand reduction goals. For example, an overall goal of a one percent annual reduction in per capita energy use within each IOU service territory are translated into specific MWh/MW numbers and targets. These targets, in turn, provide the necessary guidance to the Portfolio Analyses activities taking place in Activity II.

2. **Potentials Analyses, Baselines and Updates.** In this step, the available data on the potential for the energy efficiency resource are reviewed and a benchmark of achievable potential is established. The benchmark for achievable energy efficiency provides the information needed to refine the portfolio savings goals based on the overall availability of the resource, and its availability within specific sectors. This step also includes analyses of load forecasts and other information related to utility capacity and energy needs. This information is analyzed and used to determine the best combination of efficiency resources, sectors, and technologies for the Portfolio/Program Design activity, undertaken in Activity III.

Activity II. Portfolio Analysis – Sector and Program Priorities
This activity has one step, as described below:

3. **Program, technology and sector priorities.** Based on overall efficiency resource goals and the specific energy and demand targets identified in Activity I,

program impacts analyses and evaluations, process and market effects evaluations, as well as other completed studies will be reviewed to design a preferred portfolio of programs for the upcoming program implementation cycle. This activity involves a thorough analysis of the status and success of the current and past energy efficiency portfolio efforts. The programs, market sectors and technologies identified in step three are those that will be included in the program portfolio. This portfolio, in turn, is expected to provide the energy and demand savings reductions necessary for meeting established savings targets.

Activity III. Program/Portfolio Design, Selection, Review and Approval

Activity three consists of three steps. These steps make up the process by which the portfolio is designed and approved, and programs are selected. This activity includes: selecting and/or soliciting programs and program portfolios identified in step three that are expected to acquire the resources identified in step one, obtaining the necessary regulatory approval of the portfolio design, and reviewing and selecting proposals received for competitively-bid programs.

4. **Program and portfolio design.** A portfolio of programs will be designed based on the results of the portfolio analysis conducted in Activity II. The portfolio design will address issues such as program types, technologies, markets, and delivery mechanisms, as well as issues such as equity and risk necessary to create a balanced portfolio.

5. **Program and portfolio selection.** Programs will be selected and/or solicited to meet the criteria developed in the portfolio design. Program selections consider the policy goals and energy savings targets specified in the portfolio design, as well as the abilities of the program or portfolio providers to accomplish the goals.

6. **Portfolio review and approval**. The preferred portfolio of programs is presented for review and approval by the responsible entity.

Activity IV. Program Launch Preparation, Overall Evaluation Planning and Regulatory Review

Activity IV consists of several steps conducted in parallel involving program preparations prior to program launch, overall evaluation planning and review of the overall evaluation plan.

7. **Program launch preparation**. Preparations to place the programs into the market are made, including contract negotiations with program implementers, trade ally and key stakeholder communications, and marketing material development.

Overall evaluation planning. Goals and objectives for conducting program level evaluations are established. Evaluation resources priorities are set and evaluation resources are allocated according to these priorities. Issues such as evaluation study consolidation, the need to evaluate a particular program during a given program cycle, and evaluation scheduling are addressed during the overall planning process. An overall

plan for evaluating the programs within the portfolio is developed. Evaluation contractor selection criteria are included in the overall evaluation plan.

8. **Regulatory review.** The overall evaluation plan is reviewed by an appropriate body or bodies to assure that it meets the information needs of policy makers, portfolio managers, program administrators, and program implementers.

Activity V. Program Implementation, Evaluation, M&V, Market Assessment and Ongoing Regulatory Oversight

Activity five includes the combined efforts for the program evaluation function, and covers a wide range of steps, from the selection of the evaluation contractor to the completion of the various types of evaluation studies.

9. **Evaluation contractor selection**. Evaluation firms are selected to conduct each program evaluation designated in the overall evaluation plan according to criteria contained within the overall evaluation plan.

10. **Develop detailed program-level evaluation plans**. Evaluation firms selected to conduct the program evaluations develop detailed evaluation plans for each program based on the goals and objectives specified in the overall evaluation plan.

11. **Process evaluations**. Process evaluations consistent with this Framework document are conducted. These efforts will generally take place after the program is allowed to develop and implement its operational and implementation procedures, typically six months after program initiation.

12. **Measurement and verification**. Data collection activities preceding the full net impact evaluation are conducted, which generally includes measure installation verification, field data collection, on-site surveys, metering and monitoring, billing data collection and preparation, and other such activities. M&V activities are coordinated with the process evaluation to provide early feedback to program implementers. These efforts are conducted early enough to inform the impact evaluation efforts, but late enough that the program is successful at gaining participants and is installing or getting ready to install program measures.

13. **Impact evaluations for first program year**. First year impacts are estimated based on measure installation verification and ex-ante savings estimates. Measurement and Verification efforts are in progress and may be used to perform a preliminary adjustment of the ex-ante savings estimates based on early observations of measure installation quality and field performance. This process will make use of the established program progress tracking system.

14. **Impact evaluation for second program year**. The second program year impact evaluation consists of the use of the measurement and verification results along with the more rigorous net effects impact evaluations. Net program impacts for this program cycle are estimated according to this Framework. Data collection

and analysis covering the first two years of program operation are used to inform this process. The impact evaluations provide the load impact information on the schedule required to conduct a portfolio analysis for the next cycle of programs. Ex-post energy savings data are developed to inform the third year end of cycle true up (below at step 16), provide revised savings estimates for program continuation proposals, provide information for consideration in DEER[38] revisions, and inform the market potential analysis for the next cycle program portfolio.

15. **Impact evaluation for third program year**. The third year impact evaluation consists; in general, of confirming the net impact analysis conducted in the second program year and includes a review of the technologies installed and their use conditions in order to "true up" the second year assessment into more reliable third year net impacts. Ex-post energy savings and net-to-gross estimates developed from the second year full net impact evaluation above are combined with third year program accomplishments to estimate third year program impacts. Programs receiving funding for an additional three-year cycle may also include third year activities with first and second year activities in the subsequent program cycle to perform the full net impact evaluation covering multiple program cycles.

16. **Annual update of program impacts vs. program goals**. In some cases, there may be a need to update program impact estimates and compare those estimates with program goals to determine if the programs are providing energy and demand impacts consistent with contractual requirements. This effort is done annually, using the best available impact evaluation results at the time of the update.

17. **Market effects baseline studies (market transformation and information programs).** Baseline studies are conducted to understand the relevant baseline characteristics of the markets prior to the introduction of market transformation and/or information programs.

18. **Market effects studies (market transformation and information programs).** Program evaluation efforts for market transformation programs or energy information and educational programs are conducted. In this step the evaluation focuses on the changes to the market that are caused by the programs. These evaluations are not energy impact studies, but studies that document the changes made to the market as a result of the program. These studies also examine the effect of information programs that are designed to change the way customers use energy, or act as promotional efforts to channel participants into other programs that directly deliver energy impacts.

19. **Individual program future potentials estimates**. This step is an assessment of the future energy efficiency or demand reduction potentials associated with the

[38] DEER: Database for Energy Efficient Resouces. (Xenergy et al. 2001). See http://www.energy.ca.gov/deer/ for information.

program based on the impact, process and in some cases the market effects evaluations. In this assessment the evaluation teams examine the impact and available process and market effects information as well as the latest energy efficiency potentials study results and makes an assessment of what they consider to be the potential for the program under review over a longer period of time. This information is used to compare with the statewide market potential assessments and used to guide the setting of energy and demand targets for the next cycle of programs undertaken in Activity I.

20. **Optional longer term impact & persistence studies**. Because the three year process is cyclic and expected to reoccur such that there is a steady stream of energy programs and program portfolios, an opportunity exists to schedule longer-term impact evaluations during the following program cycles that return to the participants of the previous cycles to assesses the degree to which the impacts are still being provided, and the persistence of the measures that are providing the impacts. These longer-term follow-up cycle impact studies are used to true up the impacts of the program or program portfolios over a period of time longer than a single program cycle. This provides a method to look at six, nine and twelve year effects and to document the longer-term reliability of the energy resources from California's energy and demand programs.

A detailed schedule for this example, assuming a January 2006 program start date is shown in Table 5.1 below.

Table 5.1: Example 1 Schedule

Activity	Step	Description	Start Date	End Date	Duration (mo)
I		**Goals Setting and Potentials Analyses**			
	1	Goals setting and updating	11/1/2004	2/1/2005	3
	2	Potentials analyses, baselines and updates***	8/1/2007	2/1/2008	6
II		**Portfolio Analyses and Assessment**			
	3	Program, technology and sector priorities	2/1/2005	4/1/2005	2
III		**Portfolio/Program Design, Review and Selection**			
	4	Program & portfolio design	4/1/2005	6/1/2005	2
	5	Program & portfolio selection	6/1/2008	8/1/2005	2
	6	Portfolio review and approval	8/1/2005	10/1/2005	2
IV		**Program Launch Preparation, Overall Evaluation Planning and Regulatory Review**			
	7	Program launch preparation	10/1/2005	1/1/2006	2
	8	Overall evaluation planning	10/1/2005	11/1/2005	2
	9	Regulatory review	11/1/2005	1/1/2006	1

Table 5.1: Continued

Activity	Step	Description	Start Date	End Date	Duration (mo)
V		**Program Implementation, Evaluation, M&V, Market Assessment and Ongoing Regulatory Oversight**			
		Program start-up	1/1/06	1/7/06	6
	10	Evaluation contractor selection	1/1/2006	4/1/2006	3
	11	Develop detailed program-level evaluation plans	1/1/2006	4/1/2006	3
	12	Process evaluations	7/1/2006	7/1/2007	12
	13	Measurement and verification	9/1/2006	7/1/2007	10
	14	Impact evaluation for first program year	1/1/2007	2/1/2007	10
	15	Impact evaluation for second program year ***	1/1/2007	1/1/2008	12
	16	Impact evaluation for third program year	11/1/2008	3/1/2009	4
	17	Annual update of energy impacts vs. goals	1/1/2007	3/1/2007	2
			1/1/2008	3/1/2008	2
			1/1/2009	3/1/2009	2
	18	Market effects baseline studies (MT & Info programs)	7/1/2006	11/1/2006	4
	19	Market effects studies (MT & Info programs)	7/1/2007	1/1/2008	6
	20	Individual program future potentials	9/1/2007	12/1/2007	3

*** Second year impact results feed step 2 in subsequent cycle.

A Gantt chart showing the schedule for each of the steps in the three-year program cycle described in this example is shown in Figure 5.4.

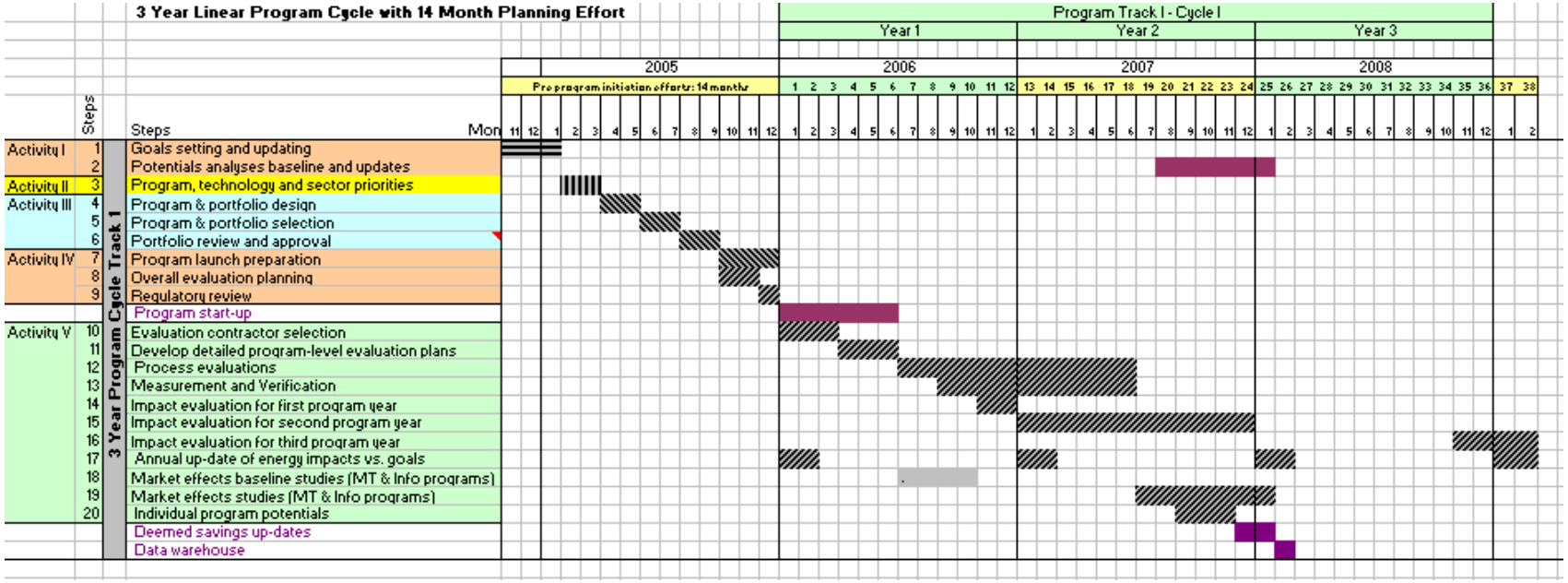

Figure 5.4: Example of Three-year Program Design, Contracting, Implementation and Evaluation Timeline

Example 2 – Offset Program Cycles

This example shows two overlapping tracks of three-year program cycles that allows for more frequent program solicitations and portfolio updates. In this example, the schedule for the second program track will start two years after the first program track.

The example presents the schedule of activities resulting operating two different program tracks, and explores the resource requirements needed to respond to the assessment, contracting, implementation and evaluation efforts associated with these activities. The schedule presented in the previous example is merged with the schedule for the second program track in Table 5.2.

Table 5.2: Schedule of Activities for Two-Track Example

Assumes a January 2006 start date for Track 1 and a January 2008 start date for Track 2
Track 2 activities are shown in italics

Track	Activity	Step	Description	Start Date	End Date
1	I	1	Goals setting and updating	11/1/2004	2/1/2005
1	II	3	Program, technology and sector priorities	2/1/2005	4/1/2005
1	III	4	Program & portfolio design	4/1/2005	6/1/2005
1	III	6	Portfolio review and approval	8/1/2005	10/1/2005
1	IV	7	Program launch preparation	10/1/2005	1/1/2006
1	IV	8	Overall evaluation planning	10/1/2005	11/1/2005
1	IV	9	Regulatory review	11/1/2005	1/1/2006
1	V		Program start-up	1/1/2006	7/1/2006
1	V	10	Evaluation contractor selection	1/1/2006	4/1/2006
1	V	11	Develop detailed program-level evaluation plans	1/1/2006	4/1/2006
1	V	12	Process evaluations	7/1/2006	7/1/2007
1	V	18	Market effects baseline studies (MT & Info programs)	7/1/2006	11/1/2006
1	V	13	Measurement and verification	9/1/2006	7/1/2007
2	*I*	*1*	*Goals setting and updating*	*11/1/2006*	*2/1/2007*
1	V	14	Impact evaluation for first program year	1/1/2007	2/1/2007
1	V	15	Impact evaluation for second program year ***	1/1/2007	1/1/2008
1	V	17	Annual update of energy impacts vs. goals	1/1/2007	3/1/2007
2	*II*	*3*	*Program, technology and sector priorities*	*2/1/2007*	*4/1/2007*
2	*III*	*4*	*Program & portfolio design*	*4/1/2007*	*6/1/2007*
1	V	19	Market effects studies (MT & Info programs)	7/1/2007	1/1/2008
1	I	2	Potentials analyses, baselines and updates ***	8/1/2007	2/1/2008
2	*III*	*6*	*Portfolio review and approval*	*8/1/2007*	*10/1/2007*
1	V	20	Individual program future potentials	9/1/2007	12/1/2007
2	*IV*	*7*	*Program launch preparation*	*10/1/2007*	*1/1/2008*
2	*IV*	*8*	*Overall evaluation planning*	*10/1/2007*	*11/1/2007*
2	*IV*	*9*	*Regulatory review*	*11/1/2007*	*1/1/2008*
1	V	17	Annual update of energy impacts vs. goals	1/1/2008	3/1/2008
2	*V*		*Program start-up*	*1/1/2008*	*7/1/2008*
2	*V*	*10*	*Evaluation contractor selection*	*1/1/2008*	*4/1/2008*
2	*V*	*11*	*Develop detailed program-level evaluation plans*	*1/1/2008*	*4/1/2008*

Table 5.2: Continued

Track	Activity	Step	Description	Start Date	End Date
1	III	5	Program & portfolio selection	6/1/2008	8/1/2005
2	*V*	*12*	*Process evaluations*	*7/1/2008*	*7/1/2009*
2	*V*	*18*	*Market effects baseline studies (MT & Info programs)*	*7/1/2008*	*11/1/2008*
2	*V*	*13*	*Measurement and verification*	*9/1/2008*	*7/1/2009*
1	V	16	Impact evaluation for third program year	11/1/2008	3/1/2009
1	V	17	Annual update of energy impacts vs. goals	1/1/2009	3/1/2009
2	*V*	*14*	*Impact evaluation for first program year*	*1/1/2009*	*2/1/2009*
2	*V*	*15*	*Impact evaluation for second program year ****	*1/1/2009*	*1/1/2010*
2	*V*	*17*	*Annual update of energy impacts vs. goals*	*1/1/2009*	*3/1/2009*
2	*V*	*19*	*Market effects studies (MT & Info programs)*	*7/1/2009*	*1/1/2010*
2	*I*	*2*	*Potentials analyses, baselines and updates****	*8/1/2009*	*2/1/2010*
2	*V*	*20*	*Individual program future potentials*	*9/1/2009*	*12/1/2009*
2	*V*	*17*	*Annual update of energy impacts vs. goals*	*1/1/2010*	*3/1/2010*
2	*III*	*5*	*Program & portfolio selection*	*6/1/2010*	*8/1/2007*
2	*V*	*16*	*Impact evaluation for third program year*	*11/1/2010*	*3/1/2011*
2	*V*	*17*	*Annual update of energy impacts vs. goals*	*1/1/2011*	*3/1/2011*

*** Second year impact results feed step 2 in subsequent cycle.

The two-track process has the potential to inform consecutive or off-cycle program tracks by allowing for the use of evaluation results to refine the ex-ante or deemed savings estimates after the second and third year program evaluations are completed. It is possible to schedule and plan ex-ante savings updates to be coordinated with the second and third year program evaluation results that are needed to adjust these energy savings projections.

Likewise, the program-specific potentials analysis conducted within the program implementation and evaluation cycle can be used to inform the wider market potential assessments that are conducted periodically to assess the ability of the market to provide efficiency and demand resources. Using a coordinated approach to the implementation and evaluation of programs, in a sequenced series of coordinated efforts, means that the information needed to inform market potential assessments will be available at specific periods of time within each program cycle. It therefore becomes possible to plan and schedule potentials assessments to complement the information that is available to feed these assessments.

A Gantt chart showing the schedule for the two-track process is shown in Figure 5.5. Arrows showing interactions within and between the program tracks are shown in the Figure.

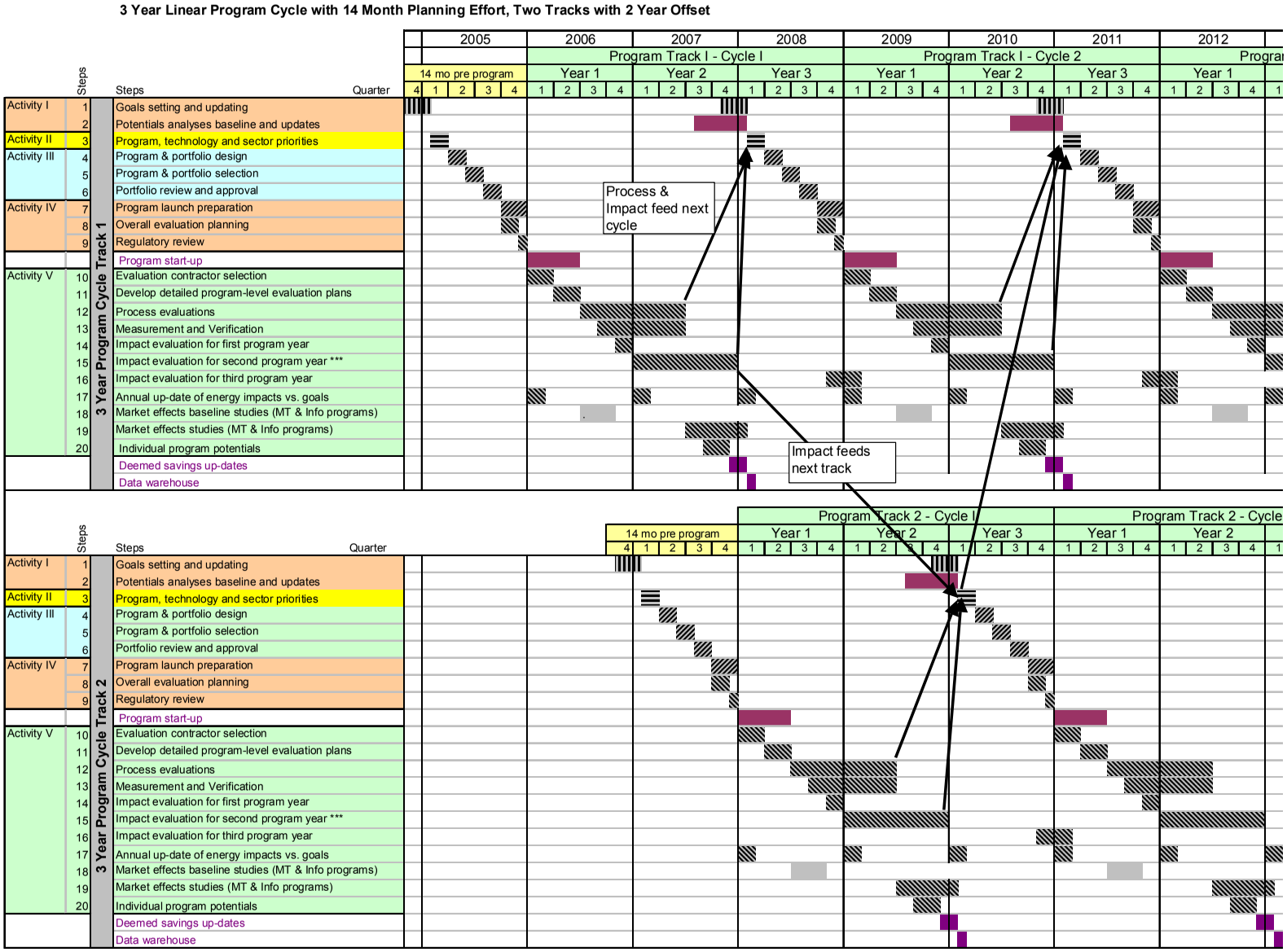

Figure 5.5: Example of Two Program Track Information Interactions

Chapter 6: Impact Evaluation

Preface

The Umbrella Roadmap (Chapter 5) provided some guidance on when an impact evaluation should be conducted. This chapter provides more details on how to conduct impact evaluations and identifies the studies and general steps needed to perform an impact evaluation. As such, this chapter provides a useful review of the impact evaluation methods and issues for policy makers and implementers.

California has a long history of conducting impact evaluations, dating back to the 1980s. The California Demand-Side Management Measurement Advisory Committee's (CADMAC) protocols, first written in 1993 and amended several times through 1998, provided the framework for conducting evaluation studies of programs funded by the state IOUs and filed with the CPUC. The purpose of these evaluation studies was to verify the energy and demand savings provided by these programs as part of a performance earnings process. The evaluation studies and methods prescribed by the CADMAC protocols were impact evaluations. Other types of evaluation, such as process and market evaluations, were not addressed by the Protocols.

Impact evaluation was the focus of California's Measurement and Evaluation Protocols and will continue to be a cornerstone of evaluation in California. Impact evaluations focus on estimating the "net" energy and demand savings of a program and its cost-effectiveness, thereby supporting California's program design, planning, and evaluation objectives. However, while the basic impact evaluation objectives have not changed, the operating environment has changed significantly in recent years and will affect the practice of impact evaluation:

- ***Energy price volatility***: the temporal variation in energy prices has made impact evaluation requirements potentially more stringent. Simple reports of annual total energy (kWh) and peak demand (kW) savings may not be sufficient for program managers and policy makers who are interested in impacts at the hourly or daily level.
- ***Energy price uncertainty***: the uncertainty in the balance of future supplies and demand and the effects of known and unknown market pressures make projections of future program benefits more difficult. Evaluations that use new pricing and price forecasting methods, with their own time/price assumptions, may be necessary.
- ***Introduction of non-utility implementers***: the audience for impact evaluation has shifted from highly skilled evaluation professionals at the CPUC and IOUs to include a diverse group of non-utility program implementers with varying backgrounds in evaluation. For many of these new entries, evaluation is a new and confusing field, so that impact evaluations will need to be carefully described.

- ***Appropriate evaluation scope for program***: Many non-utility programs deliver traditional energy savings technologies in innovative ways. These programs often have fairly small budgets, and do not have the program budgets for extensive impact evaluation requirements. If they are delivering new technologies or new configurations or new uses of existing technologies, impact evaluations for these programs may be important.
- ***Relevant research questions***: New program design and delivery mechanisms will require a new set of research questions across each type of impact evaluation.

As the environment has changed, the following issues will be very important for impact evaluation:

- ***Market noise***: The impacts of a particular program in a market filled with many implementers offering similar programs under different names with different incentive structures and marketing methods make estimating the influence of any particular program problematic. Identification of non-participants may be difficult, since customers may not be able to discern between the various programs operating in the marketplace, and may not accurately recall what influences various programs may have had on their decision processes or even remember the program in which they participated.
- ***Spillover effects***: Market transformation efforts and the market impacts of the statewide IOU programs highlight the need for developing new methods for estimating spillover effects into the non-participant population ("free drivers"). The level of free drivers may be extensive. Enhanced marketing of energy efficiency programs by non-utility implementers further increases the technical challenges of measuring spillover.
- ***Free ridership***: Free riders are project participants who would have installed the same energy efficiency measures if there had been no program. How free ridership is handled is a critical component of making the evaluations cost-effective and accurate. Uncertainty surrounding free ridership is a significant component of net energy and demand savings uncertainty; decreasing this uncertainty through additional studies could be expensive.
- ***Accuracy requirements***: The accuracy requirements for impact evaluations may need to be relaxed under certain conditions, depending on the ability to use statewide impact evaluation measure results, size of the effort, the relevant research questions, budget, and the decision processes tied to those questions.
- ***Data source and analysis methods***: The suite of tools available to conduct impact evaluations will need to evolve as the research questions change. Traditional engineering analysis, building energy simulations, metering and monitoring protocols, and statistical analysis of billing and survey data may need to be combined in new ways to answer emerging research questions.[39] The application of these tools will need to be described in a direct and straightforward manner to meet the needs of a diverse group of program planners and implementers. In some cases, new research questions may call for new analysis tools. Data

[39] See (Vine et al. 2000).

standardization efforts should be used whenever possible to gain the maximum usefulness of any data collection effort.

- ***Uncertainty***. Explicit approaches are needed to quantify uncertainty, rather than relying on single point estimates. See Chapter 12 on Uncertainty for more information.
- ***Integrated Planning Process***. Chapter 5, The Umbrella Framework, describes an integrated planning process that uses impact evaluation results to inform the planning process for the next round of programs. Conducting impact evaluations in this environment will place specific requirements on study timing and data reporting requirements.

Skills Required for Impact Evaluation

This chapter provides an overview of the methods used in billing and engineering analysis. The primary distinctions between the billing analysis path and the engineering analysis path are that the former uses both pre- and post-installation billing data and relies more on statistical/econometric methods. The billing analysis section provides background and references for policy makers, evaluation managers, and evaluators new to this field but with some knowledge of regression analysis. It is not meant to provide a "how-to" manual. The discussion here is meant instead to provide an overview and a perspective on how regression analysis methods are applied in the task of evaluating energy efficiency programs.

The authors refer readers to formal regression analysis training and textbooks in statistics and econometrics to ensure proper use of the methods discussed. This includes examining alternative correction methods for violations of regression assumptions, and a more thorough understanding for the general application of these methods.[40] The reliance on statistical/econometric methods requires personnel trained in these methods, as does the critical review of evaluations using regression methods.

The engineering analysis section provides background and references for policy makers, evaluation managers, and evaluators new to this field, but with some knowledge of basic energy engineering principles. References to the literature are provided for more information on engineering methods. The engineering analysis section covers both simple engineering methods and building energy simulation models. Simple engineering equations can be understood and used by most people with a general science

[40] Statistics textbooks that include significant discussions of regression analysis and econometric textbooks can be found in almost all college bookstores and larger libraries. Most of these probably provide adequate treatment for the use of regression in billing analysis. Two examples include: *Econometric Models and Economic Forecasts*. (Pindyck et al. 1981); and *Elements of Econometrics*. (Kmenta 1971).

background.[41] Building energy simulation models generally require personnel with a graduate degree in mechanical or architectural engineering.

Introduction and Key Issues

Impact evaluation estimates the amount of electrical energy, electrical demand, and fuel energy saved due to a program, thereby verifying whether the expected program energy and demand savings are actually occurring in the field. Impact evaluation, therefore, involves estimating a ***change*** in energy use. Since it is possible to only directly measure consumption, to estimate savings one must observe the energy use characteristics of a program participant over time and from this generally infer what the energy consumption of the participant might have been in the absence of the program.[42]

The primary purpose of impact evaluation is to obtain the most accurate and unbiased estimate of energy and demand savings due to the program. Methods used in impact evaluation can be useful for many other purposes in planning, load research, market research, and as information to work with process or market evaluation results. However, it is important to ensure that other desires not lead the evaluation away from obtaining the most accurate and unbiased estimates achievable within a reasonable cost for the evaluation needs and uses.

There are two types of savings estimates that are normally desired from impact evaluation: gross savings and net savings. ***Gross savings*** are calculated for program participants relative to their prior participation usage. ***Net savings*** controls for savings that would have occurred for these participants over the same time period whether the program was offered or not. Estimating net savings generally requires the use of a comparison group as a proxy for what the participants would have done absent the program, or self-reported information on what would have happened in the absence of the program when comparison groups cannot be reasonably identified.

---

41 The Association of Energy Engineers (AEE) offers a certificate for a Certified Energy Manager (CEM). The material covered in the CEM program is good background for understanding engineering methods in evaluation.

42 Many of the basic references on impact evaluation were written for EPRI in the 1990s. These references are not easily available to the general public. However, there are several references from the International Energy Agency < http://dsm.iea.org > that cover general evaluation topics that are in the public domain. See "A European Ex-Post Evaluation Guidebook for DSM and EE Service Programmes," (AIS 2001); and "Evaluation, Verification, and Performance Measurement of Energy Efficiency Programmes" (* Violette 1995) for more information. The international climate change mitigation effort has also produced some excellent general references. See "Guidelines for the Monitoring, Evaluation, Reporting Verification and Certification of Energy Efficiency Projects for Climate Change Mitigation." (Vine et al. 1999). *Handbook of Evaluation of Utility DSM Programs* is an early public-domain reference that provides good background material. (* Hirst et al. 1991).

Overall Impact Evaluation Objective

For the purposes of this Framework, the overall objective of impact evaluation is to estimate the net change in electricity and gas consumption and electricity demand on a program level. Electricity consumption impacts are generally defined in terms of kWh saved over some specified period. Fuel consumption impacts are generally defined in terms of therms of natural gas, gallons of propane, gallons of fuel oil, and so on. Electrical demand impacts are generally defined in terms of kW savings averaged over a short interval (e.g. 15 – 60 minutes). Demand impacts may be specified for a particular day and hour, or may be reported as the maximum savings occurring over a specified time period (e.g. summer on-peak hours).

Impact evaluation results are generally reported by costing period, which break the year into several bins depending on the season of the year and the utility system loading during these seasons. An example of the costing period definition used by Pacific Gas and Electric Company for commercial customers is shown in Table 6.1.[43]

Table 6.1: PG&E Commercial Customer Costing Period Definition

Costing period	Summer (May 1 - Oct 31)	Winter (All Other Months)
Peak	12 PM to 6 PM Weekdays	none
Partial-Peak	8:30 AM – 12 PM Weekdays 6 PM – 9:30 PM Weekdays	8:30 AM – 9:30 PM Weekdays
Off-Peak	9:30 PM –8:30AM Weekdays All Weekends and Holidays	9:30 PM to 8:30 AM Weekdays All Weekends and Holidays

The costing periods generally vary by utility and customer class, which makes it difficult to define a single costing period for evaluation results reporting. The costing period definitions used by utility planners may also change over time depending on price and load volatility. Calculating load impacts on an 8760 hourly basis provides the flexibility to apply the results to a variety of costing period definitions.

Coincident peak demand impacts are generally evaluated at the specific day and hour of maximum system demand. For non-weather-dependent measures, the daily peak load generally corresponds to a summer workday at 4 pm. For weather-dependent measures, it generally refers to 4 pm on the hottest summer workday.[44] The precise calendar day associated with the peak outdoor temperatures varies by location, and not all locations are expected to have the maximum temperature occur on the same day. Building energy simulations are generally driven with typical or long-term average weather data rather than actual weather data for a particular time period.[45] The California Energy

---

43 This definition is used by PG&E in their commercial time-of-use rates. It varies somewhat from the definition of on-peak savings from the CPUC Energy Efficiency Policy Manual (noon – 7 pm, M-F, June-September) (* CPUC 2003).

44 This definition is consistent with the *California Statewide Commercial Sector Energy Efficiency Potential Study* (Coito et al. 2002).

45 An important exception is when simulation models are calibrated to actual billing or end use energy consumption data. See Chapter 7 – Measurement and Verification for more information on simulation model calibration.

Commission (CEC) Climate Thermal Zone (CTZ) weather data sets are commonly used long-term average weather data set. An analysis of 8760 hourly weather data records in the CTZ files indicated the following days corresponding to peak temperature events in each of the 16 climate zones.[46]

Table 6.2: Day of Year for Coincident Peak Analysis (Based on 1995 Calendar year)

CTZ	Peak Day
1	7/21/95
2	7/24/95
3	7/18/95
4	7/18/95
5	9/5/95
6	9/8/95
7	7/31/95
8	7/20/95
9	8/8/95
10	8/14/95
11	8/3/95
12	7/24/95
13	8/15/95
14	8/7/95
15	7/21/95
16	8/7/95

Demand impacts can be estimated from an analysis of billing data when the data include actual building peak demand for each costing period, although the coincident peak is generally not reported. Metering studies (e.g. pre/post time-series measurements of electrical demand)[47] can be used to estimate demand in sites with permanent or temporary recording interval demand meters. Demand impacts can also be estimated from energy impacts by applying a series standard load shapes to allocate energy consumption into costing period bins. Sources for the load shape data include energy savings load shapes by measures that will be included in the California Database for Energy Efficiency Resources (DEER),[48] or from standard measure energy consumption load shapes applied to energy consumption savings estimates.[49] Note: load shapes based on building energy consumption may not provide a good estimate of the load shape of measure savings when energy savings are not linearly proportional to consumption.[50]

46 Analysis conducted by AEC for PG&E CASE initiative study.

47 The role of interval demand metering in measuring energy and demand impacts is covered in more detail in Chapter 7 – Measurement and Verification.

48 (Xenergy, ADM Associates et al. 2001).

49 Demand impacts in the latter case have long been estimated at PG&E by using a multiplier on the energy consumption savings that PG&E calls the "H" factor. H-factors vary by end use, building type, and climate zone.

50 A simple example is the economizer, or "free cooling" cycle of an HVAC system. Energy savings from these systems generally occur during periods of reduced HVAC loads.

Units of Measure

Energy metrics for impact evaluation are defined as annual end use energy and demand savings by costing period. The CADMAC Protocols required that impact data be normalized according to designated units of measure. The designated unit of measure specified by the Protocols varies by the program type and end use. Typical units include impact per dwelling unit (for residential programs), impact per lamp (for lighting programs), and impact per unit floor area for commercial programs. These designated units of measure were selected to help load forecasters at the California Energy Commission predict the impacts of demand-side management programs on the long-term statewide electrical load forecast.

Conducting impact evaluations within the context of an integrated planning process as described in Chapter 5 will require results reporting in a format compatible with the activities included in the process. Efficiency potential studies and program planning activities generally require impacts on an individual measure or measure bundle level. Portfolio analysis generally requires impacts on a program level. Load forecasting occurs at a statewide and utility level, and the data needs of the load forecasting models used by utilities and CEC planners should be considered. Measure and building specific information reported during measurement and verification (M&V) can provide the detailed engineering data useful to inform the efficiency potential and program planning processes. See Chapter 7 for more information on M&V reporting.

If the impact evaluation and its related M&V efforts are designed to provide technology-specific energy savings estimates, then the evaluation should include an effort to coordinate the results with updates to the DEER. The evaluation plan should address the technology-specific DEER data that can be updated by the evaluation and provide information on how the reliability of the data in the DEER is improved as a result of the evaluation. Results used to update the DEER can come from M&V studies, impact studies or both. Unit savings estimates consistent with the DEER schema or other engineering data supporting the refinement of the DEER unit energy savings are desired. If the evaluation results should not be used to update the DEER, the evaluation plan and the evaluation report should indicate why the data should not be used to update the DEER.

A preliminary list of the non-residential energy efficiency measures included in the 2004 DEER update study[51] is shown in Table 6.3. Note that many of the unit savings estimates can be scaled to represent a variety of technology efficiency options. Measure impacts are provided by building type, climate region and building vintage.

[51] (Itron Incorporated 2004).

Table 6.3: Energy Impact Units for Selected Nonresidential Measures Considered in the DEER Update Study

Measures	Measure group	Impact Units	Scalable	Comments
Lighting	Lighting	kW connected	Yes	Impacts expressed as % reduction
Lighting controls	Lighting	kW controlled	Yes	Impacts expressed as % reduction
Plug loads	Plug loads	kW connected	Yes	Impacts expressed as % reduction
Roof insulation, cool roof	Shell	SF roof	No	
Air curtain	Shell	Each	No	
Glass solar heat gain	Shell	SF glass	% change in SHGC	SHGC is the "solar heat gain coefficient," a measure of the solar energy transmitted by a particular glazing system
HP glass plus daylighting controls	Shell	SF glass	No	
Chillers	HVAC	ton	Difference in kW/ton	Impacts scalable by the difference in the chiller efficiency, expressed in kW/ton
Chilled and hot water distribution loop temperature control	HVAC	SF building	No	
Chilled and hot water distribution loop flow control	HVAC	Hp pump	No	
HVAC system conversion	HVAC	SF building	No	
VAV fan control	HVAC	Hp fan	No	
Evap cooling of makeup air	HVAC	Ton	No	
Ventilation rate	HVAC	SF building	No	
Heat recovery	HVAC	Each	No	
Economizers	HVAC	Ton	No	
HVAC maintenance	HVAC	Ton	No	
Heat rejection system	HVAC	Ton	No	
High efficiency furnace	HVAC	KBtu/hr	Difference in AFUE	Impacts scalable by the change in furnace efficiency, express as the Annual Fuel Utilization Efficiency (AFUE)
Large boilers	HVAC	SF building	Difference in efficiency	Impacts scalable by the change in boiler thermal efficiency, express as a percent.
HVAC system control	HVAC	SF building	No	
High efficiency packaged HVAC	HVAC	Ton	Difference in EER	Impacts scalable by the change in cooling efficiency, express as the Energy Efficiency Ratio (EER)
High efficiency motors	Motors	Hp motor	Difference in efficiency	Impacts scalable by the change in motor efficiency, express as a percent
Water heater tank insulation	Water heating	SF building	No	
Water heater efficiency	Water heating	SF building	Difference in efficiency	Impacts scalable by the change in water heater efficiency or energy factor (EF), express as a percent.

Table 6.3 represents an example of the measure impact normalization scheme designed into the DEER study at the time of publication of the Framework. Be sure to consult the latest version of the DEER for the current normalization scheme before planning an evaluation project.

Interactions Between Impact Evaluation and Planning Processes

Impact evaluation is viewed as one part of a continuous process of program planning, implementation and evaluation. Thus, the results of impact evaluation studies do not stand alone, but are used as inputs into planning and improving future programs. See Chapter 5, Umbrella Framework, for a description of the integrated program planning and evaluation process.

Selecting Methods

Chapter 5 provides information for how to decide when an impact evaluation is appropriate. This decision process is an important part of evaluation planning. Chapter 5 provides a broader examination of evaluation planning needs and the overall evaluation planning process.

The next question is what impact evaluation methodology or methodologies to use. This decision will be based on several factors. Some of these may include:

- Type of program
- Data collection costs and available budget
- Size of program
- Program evaluation history (verify/triangulate with another method, questions posed from past evaluations, etc.)
- Demonstrated effectiveness of methods when applied to other similar programs
- Evaluation of a single program or a consolidated evaluation of several programs

Billing analysis will tend to be preferred when:

- Both pre and post-retrofit billing data are available
- Expected program impacts can be expected to be observed in a billing analysis (e.g., at least 10% of total consumption, depending upon method used, cleanliness of billing data, and accuracy of measured variables in analysis)
- More often used than engineering analysis for programs with larger numbers of participants that are more homogenous

Engineering analysis will tend to be preferred when:

- No pre-measure billing data is available, e.g., new construction
- Expected impacts are too small to likely be observed in a billing analysis (e.g., less than 10% of total consumption)

- Programs with small numbers of participants or unique measures, e.g., with industrial process improvements
- Programs with significant investments in engineering methods within the program that can provide cost savings for a similar evaluation, e.g., programs that include substantial engineering M&V or building energy simulation modeling

Larger programs may want to use a combination of engineering and statistical methods, given the size of the resources being generated. It might also be useful to have programs, where possible, alternate evaluation methods during each evaluation cycle. Alternating methods can help provide greater perceived confidence in the results when these results are similar.[52]

Appendix C provides a sample set of evaluation planning guidelines that examine various program typing categories and how these can be used to set guidelines for selecting the impact methodology. It then provides a table where these criteria show the types of impact, M&V, and net-to-gross method that would then be selected.

A very general presentation of this initial decision process is graphically displayed in Figure 6.1.

[52] The calculation of statistical confidence is not affected. However, as noted in Chapter 12 on Uncertainty, bias is of significant concern and is immeasurable. Obtaining similar estimates from alternative methods can help provide evidence of a lack of bias if the potential sources of bias are not correlated between the two methods used.

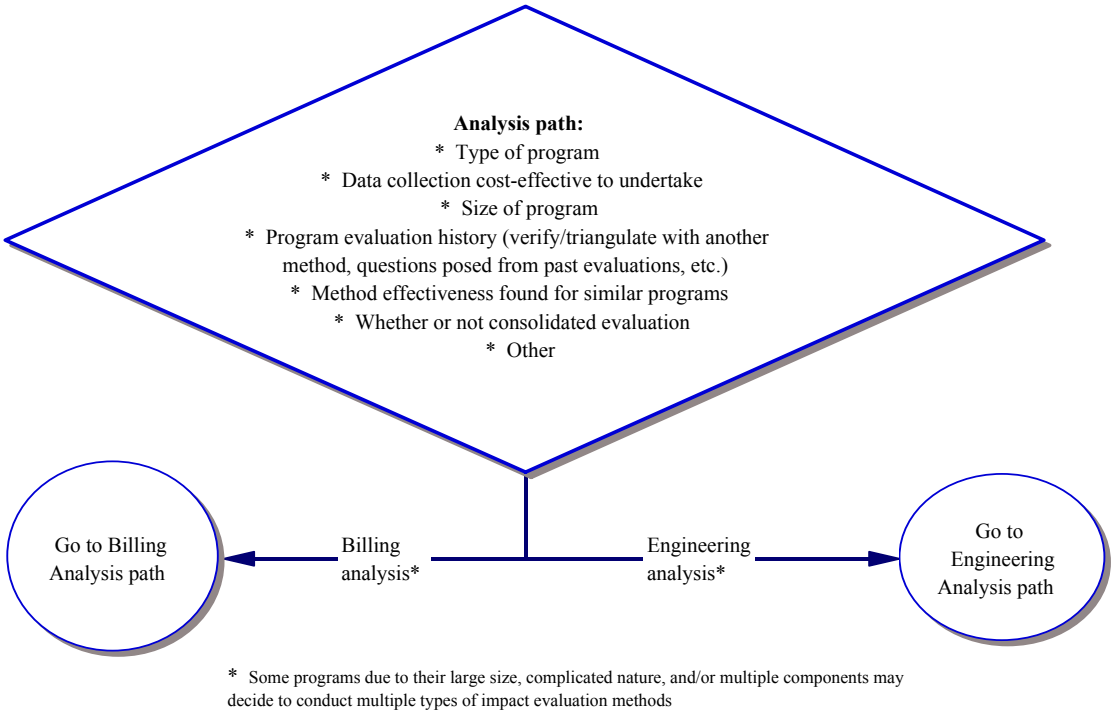

Figure 6.1: Example of General Decision Process for Selecting Impact Evaluation Approach

The Billing Analysis Path

This section of the Impact chapter provides a discussion of the critical issues that need to be assessed and addressed, wherever possible, in order for billing analysis results to be used for impact evaluation. The section provides numerous examples of energy efficiency program evaluations applying these different methods, regression diagnostic tests, and corrections. There are many more possible examples available. Those presented here aim to provide a starting point for evaluators and reviewers, to use alongside regression textbooks and research, when more information is needed for a specific analysis issue.

The authors are not recommending one particular type of analysis method or regression model for billing analysis. As was done in the prior Measurement & Evaluation Protocols,[53] the Framework provides an overview of the reasonable methods that have been used, and, more importantly, points to the need for evaluation professionals to use critical thinking to assess measurement issues and model assumptions, in order to provide

[53] *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs* (* CADMAC 1999) referenced as the Measurement and Evaluation Protocols.

the highest quality, most reliable impact estimates at a reasonable cost, given the nature of the program, data availability, and the evaluation questions to be addressed.

The Billing Analysis Roadmap is presented in Figure 6.2. It provides an overview of the likely methods and decision steps to be made. Nonetheless, this is a somewhat simplistic view of an analysis process that includes looking forward to anticipate data needs for data collection, and method and estimation adjustments as needed based upon early examinations of the data at hand.

The various billing analysis methods, literature references and examples are presented in this section. The net-to-gross analysis methods, issues, and references are provided in a later section near the end of this chapter.

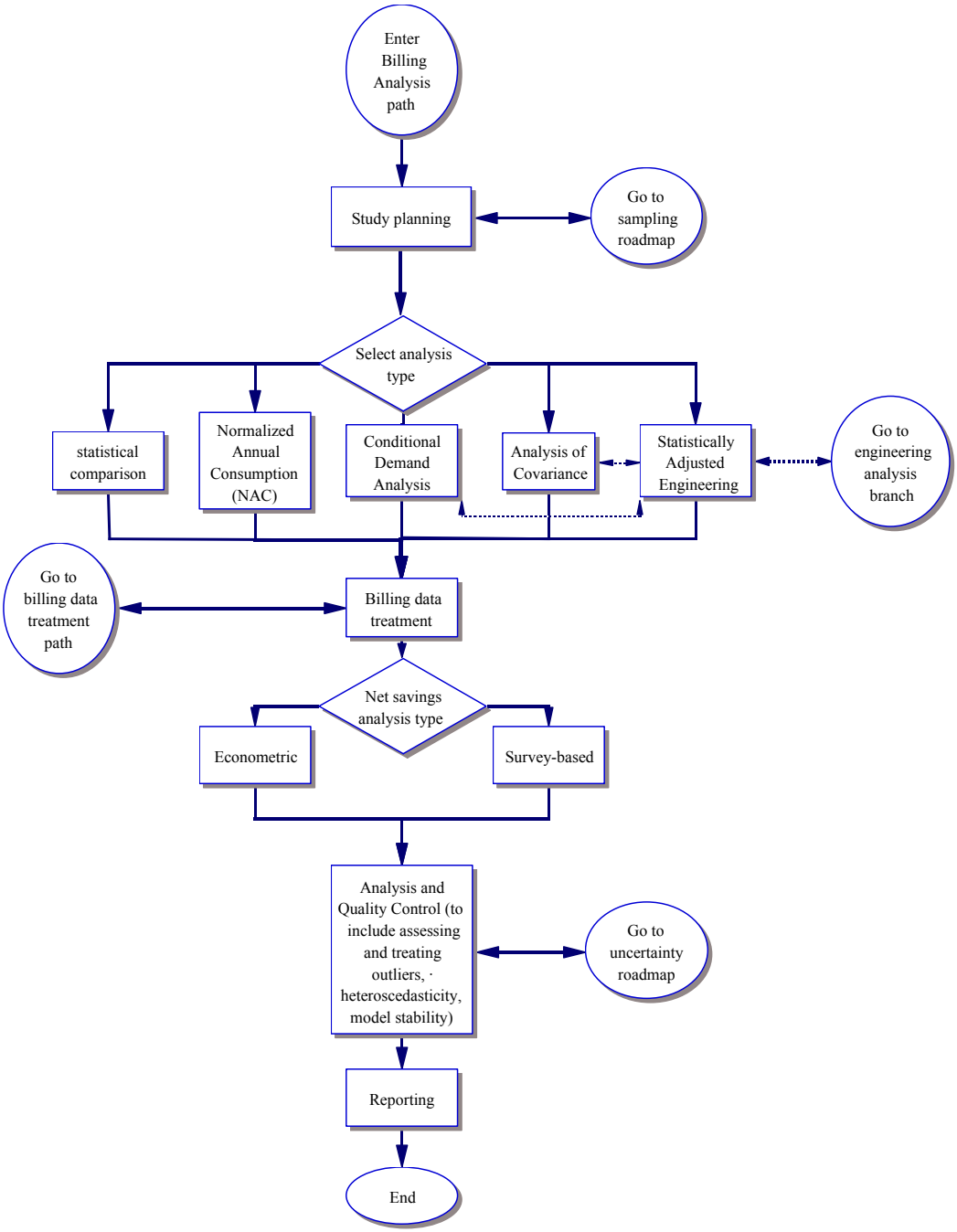

Figure 6.2: Billing Analysis Roadmap

Simple Aggregate Pre-Post Comparisons

The simplest use of billing data is to estimate program impacts by comparing average usage from the pre-installation billing data to post-installation bills. This is the One-Group Pretest-Posttest Design as described in the classic research design text by

Campbell and Stanley.[54] As discussed in this text, this simple research design contains many threats to the validity of its results.

One of the larger issues in using billing data (after the billing data have been cleaned) is that one of the largest drivers of energy consumption is the weather, and weather can vary significantly from one time period to another. This makes a pre-post comparison without adjusting for effects of weather generally meaningless. Accordingly, billing analysis normally adjusts for the weather in the pre- and post-installation periods.

A shortcoming of aggregating a customer's monthly usage to annual usage for pre-post analyses is that a lot of information is being "thrown away" that could be used in the analysis.[55] Having many monthly bills in the pre- and post-installation periods can provide more information for conducting the evaluation analysis and adding to its depth and, potentially, its reliability.

Another improvement on the pre-post design is the use of a matched comparison group in lieu of a simple comparison group.[56] A well-matched comparison group can be used to estimate what would have occurred in the absence of the program intervention (the counter-factual). Such a comparison group can come closest to serving as a control group, although the validity threat remains as some unmeasured factors relevant to measurement of effect might have been missed. A well-matched comparison group over the exact same time period might even mean that changes in the effects of weather, overall economy, and other external factors are already taken into account.

Comparing Pre-Post Billing Data for Programs with Experimental Design

Most of the regression discussion for billing analysis revolves around trying to isolate the program impacts from other elements that drive changes in energy use. With a true randomized experimental design there would be a control group rather than worrying and testing how well the comparison group matches the participant group. It would also provide net savings, avoiding the self-selection bias issues that dominate the discussion in the net-to-gross methods section near the end of this chapter.

Some have argued that publicly offered programs can not restrict who participates, that it may become a political problem and legal liability. It may be possible to create random experimental designs (or quasi-experimental design) within program design without creating political problems. The participation application process and program design offer opportunities for using experimental design that minimize some potential pitfalls in impact evaluation, such as the following:

---

54 *Experimental and Quasi-Experimental Designs for Research*. (Campbell et al. 1963).

55 *An Evaluation of Statistical and Engineering Models for Estimating Gross Energy Impacts*. (* Ridge et al. 1994).

56 In a true experiment, a control group is established through random assignment between the treatment group and the control group. Seldom is this possible in social science interventions. A comparison group may be used as a "control group." Using the term comparison group helps to remind users that assessing how close the comparison group mimics the treatment group on all relevant variables to the treatment or selectivity of group selection is important.

- When people apply to participate, they are required to agree to participate in the study associated with the program whether they are selected to participate in the program or not. Then a random sample of these qualified applicants (those agreeing to the study) are selected to participate. Thus, the comparison group would be the same type of people as the participant group, and the program effects would be relatively simple to calculate. This is an especially defensible method if the number of potential participants is greater than resources can support.
- A program could accept all prospective participants, but assign them to a random participation time period (which could be multiple periods to minimize effects from month or season). The "not-yet-participants" form the comparison group, while the "already-participants" form the treatment group. Billing analysis across these groups would not have the self-selection issues discussed in the net-to-gross section below.
- Matched community lists could be used in a neighborhood program. Then a random sample of communities could be developed to receive the program in year one, another in year two, and a third in year 3 as a phased roll-out. Effects due to the different time periods would still need to be examined. However, there are still significant advantages in using this method, for evaluation and potentially for program implementation, over a shotgun approach where the effort is a flat promotion across all communities rather than targeted efforts.
- An informational billing insert could be sent to a random sample of customers. Then surveys for recipients and non-recipients could be conducted for actions taken, or billing comparisons could be made.
- In-store rebates or in-store informational displays could be shown on random days, and sales of targeted materials affected by those days, relative to other days, could be calculated.
- An advertising rollout could be randomized across market areas. Then program/action response by market area and how that response changes with the advertising level of effort and/or message could be assessed.

Most likely, the program designs described above would initially cost more. At the same time, evaluation costs might be significantly reduced while providing potentially more defensible results. There is very little experience with experimentally designed energy efficiency programs. The advantages suggest, nevertheless, that a combined approach between program design and evaluation would be worth pursuing.

Types of Billing Analysis Models

As noted earlier, energy use analysis needs to adjust for the impact of weather on energy use. Weather adjustments can occur within a normal billing analysis framework by including weather variables, such as heating and cooling degree days. Weather normalized energy consumption can be estimated through regression for this type of comparison and is referred to as normalized annual energy consumption (NAC).

There are a variety of ways to adjust for weather including billing analysis models that perform individual regression analysis for each participant to adjust for weather and then provide the average program impacts from this grouping of regressions. One of the first and most popular of these methods in the late 1980's and early 1990s was the Princeton Scorekeeping Method (referred to as PRISM).[57] Given the popularity of this particular model, other methods that perform in the same manner are often called PRISM-like models.

Conditional demand analysis (CDA) describes a type of billing analysis in which observed energy consumption is estimated as a function of major end uses. A simplified model is presented here with binary (dummy) variables representing the presence or absence of end uses.[58] The resulting coefficients represent the marginal contribution to overall energy use associated with each end use. The dependent variable in these models is energy consumption (often monthly consumption).

The regression equation for a CDA model would appear as:

$$E_i = B_0 + B_1Ht_i + B_2AC_i + B_3WH_i + + B_3CD_i + \sum_{K=1}^{K} B_{Kn}K_n + e_i$$

where:

E_{it}	=	Energy consumption for customer "i"
B_0	=	Constant picking up energy consumed through unspecified equipment.
Ht_i	=	Dummy variable = 1 if customer "i" has electric space heating and 0 if not.
AC_i	=	Dummy variable = 1 if customer "i" has air conditioning and 0 if not.
WH_i	=	Dummy variable = 1 if customer "i" has electric water heating and 0 if not.
CD_i	=	Dummy variable = 1 if customer "i" has electric clothes dryer and 0 if not.
K_i	=	Dummy variable = 1 if customer "i" for each electric appliance in the model (and 0 if not), for appliances 1 to n.
e_i	=	Statistical error term, for unexplained variance in observed energy consumption.

This type of regression model was developed to predict energy use, and explaining energy use to customers. CDA-type models have often been used to estimate unit energy consumption (UEC). The model above, however, has frequently been found to be too simplistic. UEC's are often not linear and may depend upon a host of variables in complicated ways. A second problem lies in the nature of incorporating many variables,

57 There are several papers in *Energy and Buildings* dedicated to this approach. Some of these include: "Exploratory Scorekeeping for Oil-Heated Houses." (Fels et al. 1986); "A Midwest Low-income Weatherization Program Seen through PRISM." (Goldberg 1986); and "The Applicability of PRISM to Electric Heating and Cooling." (Stram et al. 1986).

58 "The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector." (Parti et al. 1980).

one for each electric using device. The possibility of over-specification and multicollinearity then becomes a real issue. In fact, a simplistic representation of "dumping" many dummy variables in for every appliance has at times been found to produce impossible estimates such as negative UECs. These problems are especially troubling where the primary evaluation question is not the predictive ability of the model for energy use but accurate and reliable estimates of energy savings or UECs.[59] Yet, careful work with CDAs that test and correct for these issues can make it a useful tool in the evaluation toolkit.

A study supported by San Diego Gas & Electric demonstrated through Monte Carlo simulation studies that a more direct estimation of savings than CDA could be as accurate in estimating savings as a CDA model with various CDA-type variables. Yet, it also demonstrated that if there is either measurement error for the variables added to the CDA model (errors-in-variables bias) or misspecification of the model then the more direct estimation of savings method is significantly more accurate.[60]

The term CDA has occasionally been expanded to cover a wide range of hybrid models that incorporate program data, consumption data, and engineering data. It is important to limit application of the term CDA to its original, more precise meaning. In this way, impact evaluation reporting can more easily be explicit about the type of model used in billing analysis.

In the traditional interpretation of terminology, a "Change Model" is different from a CDA Model. The CDA Model is designed to explain energy uses, while a Change Model is designed to explain changes in energy usage. That means there is a difference (or change over time) being explained by the model. This can take the form of having the change in energy consumption (pre versus post) as the dependent variable (e.g., December pre-retrofit usage – December post-retrofit usage), or having consumption as the dependent variable and pre-retrofit consumption as one of the independent variables.

This type of model could be specified as:

$$\sum_{K=1}^{K}$$

[59] An OLS regression procedure uses only the variation unique to a regressor as it relates to the dependent variable. This means that if two or more regressors are highly correlated for the purpose of calculating the coefficients the common variation is ignored. The overall variation is included in the R^2 and the overall estimate of usage may still be predictive. However, the coefficient of the regressors has high variance and may produce coefficients that are not reflective of the true effect. Given the coefficient of the regressors are the actual evaluation goal in either UEC estimation or energy savings estimation (from either a participation dummy variable or SAE variable), this is completely counter-productive. See the discussion below on misspecification and multicollinearity. Also see standard regression/econometric texts for discussions in this area, e.g., *A Guide to Econometrics*, 3rd edition. (Kennedy 1996), pages 176 – 187.

[60] The direct savings estimate is opposed to a CDA-approach which seeks to estimate baseline consumption and post-program consumption and using the difference as the estimate of savings. See *Appendix Z Simulation Study: Comparison of Alternative Methods for Measuring the Gross and Net Energy Impacts of Demand-Side Management Program (with Addendum)*. (Schiffman et al. 1993), available at www.calmac.org.

$$E_{i,\,Post} = B_0 + B_1 INSTALL_i + B_2 E_{i,\,Pre} + B_3 W_{it} + \quad B_k X_k + e_i$$

where:

$E_{i,\,post}$ = Energy consumption for customer "i" in post-periods
B_0 = Constant picking up energy consumed through unspecified equipment.
$INSTALL_i$ = Dummy variable = 1 for participant (installer) and 0 for non-participant.
$E_{i,\,pre}$ = Energy consumption for customer "i" in pre-periods.
W_{it} = Average weather for customer "i," in month "t," as defined by that customer's billing cycle.
X_k = Vector of other explanatory variables

The B_1 coefficient on the INSTALL dummy variable provides the measure of energy savings.

To control for differences between customers, other explanatory variables (the vector of X_k variables in the above) can be used in these models, such as variables that affect energy usage (e.g., square feet, operating hours, industry) or that affect energy usage over time (e.g., the price of energy, economic conditions).

A special type of explanatory variable, incorporating the engineering estimate of savings, provided a category of models called Statistically-Adjusted Engineering (SAE) Models.[61] The regression coefficient in these models is the percentage of the engineering estimate of savings observed in changes in energy usage. For example, if the coefficient on the SAE term is 0.8 this means that the customers are on average seeing 80% of the savings realized from their engineering estimates. The SAE model can estimate realization rates for an overall customer savings estimate or for individual measure savings estimates if these estimates are placed within the model.

A basic SAE model would look as follows:

$$E_i = B_1 S_i + B_2 \sum_{K=1}^{K} W_{it} + \quad B_K X_k + e_{it}$$

where:

E_i = Average energy consumption for customer "i" from the billing data.
S_i = The measure savings estimates in an SAE model for customer "i" in month "t" are used instead of the INSTALL variable seen in the model above. For a measure-level SAE model, as proposed here, a variable

[61] "Statistically Adjusted Engineering (SAE) Models of End-Use Load Curves." (Train et al. 1985); "The Economic Value of Energy-Saving Investments by Commercial and Industrial Firms." (Train et al. 1987).

and coefficient B for each measure-level savings estimate would be used.

W_{it} = Average weather for customer "i," in month "t," as defined by that customer's billing cycle.

X_k = Vector of other explanatory variables

$B_1 ... B_n$ = Estimate coefficients.

e_i = Statistical error term, for unexplained variance in observed average energy consumption.

This technique was originally intended for use in billing analysis to calculate gross energy savings. It provides a method to incorporate all available information (most importantly engineering estimates) into the regression, while providing feedback on the accuracy of the engineering estimate. Confusion has occurred, as some have tried to re-interpret the SAE approach for analyzing net savings (e.g., with comparison group data in the regression), and others have said that the SAE was a simplified form of an expanded CDA Model. The authors recommend that evaluators use a "back to the basics" approach when using terminology to describe the regression methods used. This should be less confusing to readers. (For example, a regression model was used, a CDA model, a CDA-SAE model, or a SAE Change model, etc.)

Most regression models are estimated as ordinary least squares (OLS), generalized least squares (GLS), or other forms of maximum likelihood estimation.[62] These methods generally produce similar results under similar circumstances. Generalized least squares, as its name implies, is a more generalized statistical equation. If the error term is normally distributed, both OLS and GLS may be identical to the maximum likelihood estimate (MLE).[63] There are differences in these estimation methods, however, that lead to the decision of which model specification is more appropriate for different circumstances. The more generalized the method, the more it can often be used to correct for different issues. At the same time, it can become more computationally difficult.

The billing analysis models discussed thus far can use OLS, GLS or maximum likelihood estimation. Discussion, comparisons, and derivations for these regression models can be found in most graduate regression statistics or econometric textbooks.

OLS is the most commonly used of these techniques and the easiest to use in most statistical software packages. It is important that the estimator of the model's coefficients is unbiased (centered on the correct answer) and consistent (that it approaches the exact population coefficient as the sample size gets larger). According to the Gauss-Markov

---

62 In general, there are three methods of estimation: moments, least squares, and maximum likelihood. (Moments is not discussed in the *Framework*. This method uses the sample mean as an estimate of the population mean.) See *Elements of Econometrics*. (Kmenta 1971).

63 The maximum likelihood estimate (MLE) is based upon the concept that different populations would generate a different sample. It is estimating what the population would look like that would most likely provide the sample values obtained. The technique provides estimators that maximize the probability of getting the sample, i.e., what estimators provide a population distribution that would maximize the probability of seeing the sample being analyzed. See pages 174 – 182 of *Elements of Econometrics* (Kmenta 1971), or the MLE discussion in any graduate econometrics textbook.

theorem, the best (minimum variance producing) linear unbiased estimator can be achieved with ordinary least squares, as long as four assumptions are correct. These are:

1. The relationship between the dependent and independent variables is linear;
2. The independent variables are fixed, non-stochastic variables (i.e., they are non-random variables known by the researcher);
3. The error term has zero expected value, and constant variance for all observations (i.e., the residuals fall uniformly around the correct answer, without a significant pattern being formed in the residuals according to unmeasured characteristics of the observations);
4. The error term is random and the errors (residuals) are uncorrelated to one another.

If these four assumptions are correct, OLS will provide a better estimator than GLS. But, if any of these assumptions are incorrect, GLS should be used. If OLS is inappropriately used in cases where one of these assumptions is violated, it will produce an estimator with greater variance than GLS would have produced. There are also common techniques that can be used along with GLS to correct for cases that violate these assumptions.

One billing analysis method that is designed to use GLS rather than OLS is Analysis of Covariance (ANCOVA) models. An ANCOVA model is also referred to as a "fixed-effects" model. This model allows each individual to act as its own control. The unique effect of the stable, but unmeasured characteristics of each customer is their "fixed effects;" from which this method takes its name. These fixed effects are held constant. The fixed-effects nature of the model means the ANCOVA model does not need to include unchanging customer characteristics such as square footage, number of floors, equipment in the home, etc.[64, 65]

Controlling for fixed effects controls the amount of variance (noise) the model is faced with, since each customer has a different baseload, a different response to weather, and a different pattern of consumption that changes over time. This approach also provides for a much closer fit to the data than most models as individual responsiveness is incorporated. At the same time, using individual responsiveness is more meaningful than including lagged usage variables. This method is similar to the PRISM-type models that

[64] *Loan and Whole House Rebates: Program Evaluations* (Megdal et al. 1993); and "Estimating Takeback (Comfort Increase) For a Low-Income Program, Loan Program, and a Single Family Rebate Program." (Megdal et al. 1993).

[65] "A Comparison of Model Specifications In a Billing Data Analysis of Impacts From a Commercial and Industrial Rebate Program," (Sumi et al. 1993); "The Treatment of Outliers and Influential Observations in Regression-Based Impact Evaluation," (Schutte et al. 1994); "A Monte Carlo Based Comparison of Techniques for Measuring the Energy Impacts of Demand-Side Management Programs," (Schiffman 1994).

allow separate baseline consumption for each customer and separate regressions for each customer.[66]

The basic model framework for an ANCOVA SAE model is as follows:

$$E_{it} = B_1 S_{itj} + B_2 W_{it} + \sum_{K=1}^{K} B_k i + e_{it}$$

where:

E_{it} = Average daily energy consumption for customer "i," in month "t," from the billing data, with the consumption for the billing cycle, divided by the number of days in the billing cycle.

S_{itj} = Dummy variable = 1 if customer "i," in month "t," had installed measure "j"; = 0, if the conservation measure had not yet been installed. For a SAE model, the measure savings estimates would be included in place of the "1" for the months after installation. For a measure-level SAE model, there is a variable for every engineering estimate "j."

W_{it} = Average weather for customer "i," in month "t," as defined by that customer's billing cycle.

B_{ki} = For ANCOVA, customer "i," included as own control for fixed-effects. The coefficient adjusts for the customer's base usage as differentiated from the usage for the sector based upon the other variables in the model. This means there is a coefficient for every customer in the model, so a model with 500 customers would have 500 coefficients. Interacted with weather, the coefficient adjusts for the customer's weather sensitive usage, as differentiated from the usage for the group as a whole, based upon the other variables in the model. In our example, this would add another 500 coefficients to the results.

$B_1 ... B_n$ = Estimate coefficients.

e_{it} = Statistical error term, for unexplained variance in observed average daily energy consumption, for customer "i," in month "t."

The coefficient "B_1" will provide either the average daily consumption savings from the measures installation (standard billing analysis), or the percentage of the engineering estimate obtained for an SAE model - depending on whether a dummy variable is used, or whether all sample participants have program engineering estimates available for all measures installed. From this coefficient the average savings can be estimated.

Regression Summary

Most billing analysis conducted to evaluate energy efficiency programs is done with one of the above models. Variants include incorporating comparison groups in the analysis

[66] One of the differences between PRISM-like models and ANCOVA is the greater flexibility of the ANCOVA model, which allows separate customer-specific interactions with weather and other variables to be easily assessed. The GLS form generally used to perform ANCOVA in standard statistical software packages also allows corrections for various regression assumption violations to more easily be examined and controlled.

and using methods that attempt to control some of the selectivity bias that occurs in participation (i.e., not having the comparison group a perfect match for the participant group). There are also tests and methods used to correct for violations in the assumptions of the regression analysis.

In fact, most of the work that many econometricians do is intended to deal with the reality that the key assumptions of OLS are usually not met in field data. Most graduate textbooks on regression analysis spend the first quarter on the basics of regression and then the rest of the book is in providing ways to address violations in the OLS assumptions, model form, data or other "problems" when using OLS.

One of the more critical assumptions in regression analysis is that the model specified is the correct model. Using regression analysis for causality assumes that the independent variables cause the actions being measured in the dependent variable, not just a correlation. It also assumes that this is the one and only true model for the cause of the movement seen in the dependent variable. This assumption is for the variables included, the mathematical form of the interaction, and the treatment of any non-random error term effects. Model specification is a critical component of regression analysis. Problems concerning model misspecification will, therefore, be examined again later in this section.

Accordingly, there is a need for experienced professionals to conduct billing analysis by critically analyzing their data and situation to be able to find when alternative models, tests, and corrections need to be employed. They must do so in a professional manner, reporting on their research design, logic, and potential issues (such as potential bias issues, as discussed in the Chapter 12 on Uncertainty).

A few of the more common issues, tests, and corrections will be discussed below. The authors are not recommending any specific billing analysis model, specific tests that must occur, or methods for corrections when statistical issues are found. However, it is essential that the evaluator's work be a credible professional job in line with the evaluation needs for that project and research results in the fields of regression analysis, econometrics, and energy efficiency evaluation.

Bayesian Analysis

The above discussion on billing analysis models has come from the more traditional statistical perspective. Another approach can be undertaken from a Bayesian perspective. (For a basic description of Bayesian statistics, see the sidebar provided in the Chapter on Uncertainty.) The foundation of the difference between the two paradigms is that classical statistics is based upon the notion of repeatedly drawn samples and the sampling distribution of the estimate for the true parameter, while Bayesian statistics is founded on subjective probability and the use of prior information.[67]

Bayesian output can be quite meaningful in an evaluation framework, though the interpretation is somewhat different.

[67] While the SAE model uses prior information by incorporating the engineering estimates of savings, it is not conducted as a Bayesian analysis and has not been interpreted in a Bayesian manner.

> *"Instead of producing a point estimate of ß, a Bayesian analysis produces as its prime piece of output a density function for ß called the "posterior" density function. This density function relates to ß, not B, so it most definitely is not a sampling distribution; it is interpreted as reflecting the odds the researcher would give when taking bets on the true value of ß."*[68]

Very little energy efficiency evaluation has occurred within a strict Bayesian approach.[69] This type of analysis may be applicable for impact evaluation, but the evaluation plan would need to show how the method would work for the evaluation being conducted, the advantages of doing so, and how it is to be conducted. Then it could be recommended to be part of the waiver process.

Critical Billing Analysis Issues and Quality Control Efforts

The perspective throughout this document has been to provide a discussion of the foundation that supports the roadmap and references for more information. It then directs evaluations to be conducted by skilled professional evaluators for the tasks to be performed and to be done in a manner guided by the science in that area (e.g., engineering for engineering analysis, statistics for sampling, and regression analysis for billing analysis). An extremely prescriptive manual of methods would not allow evaluations to be designed that would best fit the program and evaluation circumstances. This is particularly true in the area of billing analysis.

In billing analysis, as in all other approaches to evaluation approaches, it is important to describe in the evaluation plans and reports the potential areas of bias, questions about the specification of models, sampling issues, etc. In a high quality impact evaluation, the evaluators ensure that the evaluation is conducted so that issues of specification, sampling, bias, etc., are specifically tested, analyzed, and addressed, either empirically or qualitatively (as is appropriate to the issue, program, and evaluation goals), and that they are disclosed for knowledgeable reviewers to understand.

The Measurement & Evaluation Protocols' Quality Assurance Guidelines presents a series of questions that needed to be answered as part of the evaluation analysis and reporting (pages 11-14). The Framework supports having these issues tested for, assessed, and addressed where needed for billing analysis. They include:

- Misspecification
- Non-random error term
- Non-random measurement error
- Autocorrelation

---

68 *A Guide to Econometrics*, 3rd edition. (Kennedy 1996), page 201. For an introduction to Bayesian statistics see: *Bayesian Statistics for Evaluation Research: An Introduction*. (Pollard 1986).

69 An initial assessment of the usefulness of using a Bayesian approach for impact evaluation is shown in "Impact Evaluation Accuracy and the Incorporation of Prior Information." (Violette 1991).

- Heteroscedasticity
- Collinearity
- Tests for exogeneity and endogeneity
- Influential data
- Missing data
- Triangulation
- Weather effects
- Engineering priors
- Precision
- Comparison group

Model misspecification

An easily overlooked regression problem is misspecification of the model. A critical assumption in regression analysis is that the model is correctly specified, i.e., that it represent the underlying process. In some sense, regression assumes that the regression model being tested is the one and only true representation of the process that determines the dependent variable. Using regression analysis for causality assumes that the independent variables cause the actions being measured in the dependent variable, not just a correlation. This assumption pertains to the variables, the mathematical form of the interaction, and the treatment of non-random error term effects. These are strong assumptions, and most practitioners realize that a regression model may be missing some variables and data (some of which may be important). It is important, however, not to get complacent about the imperfections, testing and correcting for misspecification. Typical categories of specification errors include the following: [70]

1. Omission of a relevant explanatory variable.
2. Disregard of a qualitative change in one of the explanatory variables.
3. Inclusion of an irrelevant explanatory variable.
4. Incorrect mathematical form of the regression equation.
5. Incorrect specification of the way in which the disturbance enters the regression equation.

It is hard to know when a model is misspecified. Specifying the model is generally based upon the relevant theory, experience, the literature on similar work, and testing alternative models. There are a variety of tests that can be utilized.[71]

---

[70] *Elements of Econometrics*. (Kmenta 1971), page 392.

[71] *A Guide to Econometrics*, 3rd edition (Kennedy 1996). There is also a 2003 edition that might prove useful with updating and additions from latest research.

Econometricians are trained in the theoretical underpinnings of the statistics and correct for a variety of non-random error term situations through statistical tests and corrections. An alternative approach to analysis has been recommended by Kennedy that it is important to consider. "Models whose residuals do not test as insignificantly different from white noise (random errors) should be initially viewed as containing a misspecification, not as needing a special estimation procedure."[72]

Misspecification can cause many problems. Although difficult to detect as a problem, model specification and alternative specifications should normally be assessed by the results they provide, the stability of the model, and the consequences in the regression diagnostic tests. It is important to recognize that energy savings estimates depend not on the predictive power of the model on energy use, but on the accuracy, stability, and precision of the coefficient that represents energy savings. The best evaluations show the alternative model specifications considered and how they performed.

Non-random error term

The most common critical threats to the validity of the statistical estimate of an estimator discussed in literature are the cases where OLS is used while one or more of its assumptions are violated.[73] (These assumptions were stated in an earlier part of this section.) As noted previously, if OLS is inappropriately used in cases where one of these assumptions is violated, it will produce an estimator with greater variance than GLS would have produced. GLS and other techniques can be used to correct for cases that violate the OLS assumptions.[74]

The ANCOVA model discussed above is one way used to address the problem that the error term is not truly random. It does so through measuring the covariance among categorical variables. Often these types of models are divided into random effects models (or variance components models) and fixed effects models. Much of the work in this field involves providing the appropriate estimators for differing circumstances or assumptions about the components and relationships of the error terms. (See the cited articles for more detailed discussions of this type of work and its applications.)[75]

As noted in the discussion above on billing analysis methods, efficiency billing analysis often involves cross-sectional time-series data. In other words, there are observations over time for an individual or firm (e.g., collecting utility bill data during pre- and post-installation periods), and there are data across individuals or firms. These types of

---

72 *A Guide to Econometrics*, 3rd edition (Kennedy 1996), page 77.

73 Recognize, however, that this may be the more recognizable problem while misspecification could easily occur most of the time without being recognized.

74 This is because the only assumed error structure in the generalized method is that the variance-covariance matrix of the error terms be multiplicative scalar and positive definite. *Econometric Models and Economic Forecasts*. (Pindyck and Rubinfeld 1981).

75 (Balestra et al. 1966; Maddala 1971; Maddala 1971; Hausmann 1978; Mundlak 1978; Hausmann et al. 1981; Lillard et al. 1981; Aigner et al. 1984; Aigner et al. 1985; Jasso 1985; Amemiya et al. 1986; Cornwell et al. 1988; England et al. 1988; Megdal, Haynes et al. 1993; Sumi, Oblander et al. 1993; Schutte and Violette 1994; * Megdal et al. 1995; Ozog et al. 1995).

datasets have the possibility of containing error distribution issues originating from either the time dimension or the participant dimension.

Non-random measurement error and Errors-in-Variables Bias

Each of the independent and dependent variables need to be appropriately measured to ensure proper results. If variables are measured with random error, the model has additional "noise" that may make ascertaining the correct model more difficult and lowers the precision achieved. If the variables are measured with nonrandom error that has any correlation with other variables in the model, then the regression coefficient can be biased. If an independent variable is measured with non-random error, then the results will be biased and inconsistent. If the dependent variable is measured with error, the variances of the estimates will be large. If it is measured with non-random error, then relationships being measured by the model are not those in the construct, and the results may be biased.

If the interest is the dependent variable, random measurement error in the independent variable is less important. The coefficient for the independent variable may be biased but the general error term picks up the random error in the measurement of the independent variable and the estimation of the dependent variable is still valid. Unfortunately, however, most billing analysis models are predicting consumption and the variable linked to the primary research interest, energy or demand savings, is often an independent variable. The measurement error in independent variables can bias their coefficients to zero and measurement error in multiple independent variables can provide bias in their coefficients of unknown direction and magnitude. This problem is often referred to as errors-in-variables bias.[76]

Heteroscedasticity

Another important error assumption is the constant error variance or homoscedasticity. There are many cases where this assumption may not be true. To not have a constant error variance or unequal variances, where the error pattern varies systematically for different types of individuals or firms (potentially a missing variable problem), is heteroscedasticity. This is quite likely when some unmeasured characteristic limits the variability for one group versus another group. For example, this assumption is not likely to hold for commercial analysis where there are small firms and a small group of very large firms. Large firms/buildings have the likelihood of having greater variance (and error variance) in variables with a greater potential size (i.e., size is correlated with the variance). This is quite likely because there are many more items and people that can vary in their actions in a larger building/facility than a smaller one. This is a classic example often used in regression textbooks, even outside the efficiency evaluation field, as a common problem for cross-sectional studies.

[76] A simulation that provides evidence of the accuracy problems caused by the errors-in-variables bias for energy efficiency program evaluation can be found in *Appendix Z Simulation Study: Comparison of Alternative Methods for Measuring the Gross and Net Energy Impacts of Demand-Side Management Program (with Addendum)*. (Schiffman and Engle 1993).

Given its common nature among cross-sectional studies involving facilities, it is recommended that billing analysis for commercial or industrial programs, at a minimum, test for heteroscedasticity. At the same time, there can be many different causes of heteroscedasticity (including income distribution) that would suggest that any high quality evaluation would at least examine this possibility. A relatively simple first test would be to plot and view the residuals against variables that are likely to be correlated with the error variance issues (e.g., facility size, income). There are also several formal tests available in most regression statistical software packages used that look for different relationships between the error terms and the independent variables, e.g., Goldfeld-Quandt (which looks at the magnitude issue in firm size), Breusch-Pagan (linear combination of known variables creating error disturbance), White test, etc.[77] Solutions are available for most problems. A common correction in GLS is to use weighted least squares, which is offered as an option in most regression statistical software packages.

Autocorrelation

Another common problem within this class of cross-sectional time-series models is autocorrelation.[78] Autocorrelation of errors is most common in time-series (due to the intrinsic relationship between the most recent prior period and the present measurement while unspecified variables are missing that would explain the underlying mechanisms for these changes). Autocorrelation, however, can also occur due to spatial autocorrelation (where firms located in the same area could have correlated errors due to an omitted variable for other factors influencing the firms that are correlated with their location).

The time-series nature of billing analysis makes it a prime candidate for problems with autocorrelation. There are, however, other logical ways in which billing analysis could generate autocorrelation problems. Let's suppose individuals living inland increase their air conditioner usage as temperature rises because either the heat is already in their buildings or due to their anticipation of how hot it may become inside the building. In contrast, individuals living closer to the coast are slower to respond to a temperature increase given an expectation of its temporary nature. If the weather variables cannot capture this complicated relationship and no geographic variable is used (and the model does not control for fixed effects, such as an ANCOVA model), then this circumstance could create a problem with autocorrelation given the non-random nature of the error term and its being correlated to geographic location.

Given that autocorrelation is relatively common in time-series models, a high quality evaluation would test for this. There are several tests and corrections available in common regression statistical software packages.

---

77 *A Guide to Econometrics*, 3rd edition. (Kennedy 1996), page 118.

78 Some econometricians believe that autocorrelation is a type of model misspecification.

Examples from Energy-Efficiency Evaluation

The first example is a large commercial and industrial impact evaluation that used billing analysis as one of its evaluation methods.[79] Regression diagnostics were performed on all the models regardless of how "good" the initial modeling results appeared. The regression diagnostics used were based upon testing the probability that the residuals were normally distributed; skewness measurement; kurtosis measurement; a Pearson's correlation coefficient between the residual and the lagged residual; and an examination of residual plots against the predicted values, the savings estimate, average heating degree days, average cooling degree days, and time.[80]

The initial office sector energy model had significant problems with heteroscedacity. The initial model had a probability of normally distributed residuals of 71 percent, a skewness measure of 2.43, and a residual plot showing one customer with consistently higher savings. This was solved by creating two models, a separate model for a large customer who had had much of its retrofitted space vacant in the post-period, and another model for the remainder of the sector. The correction of using two models allowed the sector model to have a probability of normally distributed residuals of 90 percent, and the skewness measure fell from over 2 to -0.8. These corrections found significant differences in the realization rates achieved for the savings estimates, proving the importance of this type of examination. One realization rate was created and used for the unusual customer and one for the rest of the office sector.

Within the same project, theoretical work was conducted on how trends in economic conditions would affect savings. This analysis showed that savings estimates would be overestimated (in terms of what otherwise would have occurred, or from the perspective of long-run expected realization rates) during recessionary trends. Savings estimates would be underestimated during strong periods of economic growth. The period being analyzed was known to be a strong recessionary period. This meant that, without considering this factor, savings would be overestimated.[81]

Since the analysis was conducted during a recessionary "trend," the evaluators hypothesized that the economy might be contributing to the "trend" problem seen in the autocorrelation tests. Autocorrelation was found; the residual and lagged residual had a Pearson's correlation coefficient of 0.59. A trend variable was used to correct the autocorrelation and as a proxy for the decreasing consumption trend seen in manufacturing over the long analysis period.

---

79 "The Importance of Using Analysis of Covariance (ANCOVA), Diagnostics, and Corrections within Billing Analysis for Large C&I Customers." (* Megdal et al. 1995), page 440.

80 Skewness is the measure of the departure from symmetry, with the median being different than the mean and skewness being defined whether the tail is elongated to the left or to the right. Kurtosis is the extent to which the sample is peaked (leptokurtic) or flatter (platykurtic) than a normal distribution. Pearson's correlation coefficient is one of the most common methods of assessing the nature of the relationship between two variables.

81 "The Changing Economy as Part of DSM Impact Evaluations: Evidence from a Large C&I Retrofit Program Evaluation." (* Megdal, Paquette et al. 1995).

The use of the trend variable did make a very large change in the savings estimate in the direction predicted by our initial theoretical economic analysis. The demand model without the trend variable showed a realization rate of 105%, while the model corrected for autocorrelation showed a realization rate of 55%. Here again, the importance of testing and then correcting for error assumption issues is shown. This work also points to the need to consider the issues associated with estimating realization rates when strong economic trends are occurring.

Another study shows the need for, and the use of, numerous regression diagnostics, corrections, critical examination, and methods discussed here and above.[82] The examinations conducted, corrections employed and comparisons made to assess reasonableness are too numerous to list here. Instead, a brief summary is provided. The evaluators found "the diagnostics for the time-series/cross-sectional models complex, but crucial." The format of their modeling immediately allowed an elimination of the first-order correlation in the pooled model. They used a comparison group created from several assessments to ensure a proper one that could control for exogenous factors. Engineering priors were created to provide an engineering estimate for evaporative cooler use for each customer in each time period (based upon cooling degree days in their area and floor space). This added more prior information to the model (a significant improvement over using a tracking system annual estimate). They created a correction factor so that the gross savings estimate from the participant/non-participant billing analysis did not underestimate gross savings by incorporating freeridership (which was estimated separately using a three-option nested logit approach). They found one regression diagnostic (DFBETA, a measurement of the influence of a particular observation) to be particularly useful in their analysis. Another important correction was the use of time-period fixed-effect dummies that otherwise indicated a substantial omitted variable bias.

The above example also points out an example of a pathway in the billing analysis prescriptive roadmap. The billing analysis roadmap shows that one of the methods has a side-arrow going from billing analysis methods to engineering analysis methods and back again to billing analysis. This method was used in the Samiullah[83] example. For this evaluation, new engineering priors were developed to be more accurate by being time and customer-specific.

Another study by Coito and Barnes conducted a more extensive engineering adjustment to engineering priors and one that was specifically done in relationship to regression diagnostics.[84] This study used a three-stage process for the impact evaluation. The first stage was a preliminary regression analysis that used telephone survey data and program tracking data along with the billing data. Regression diagnostics performed on this work were used to identify outliers that would then be the focus of on-site surveys and engineering analysis. The on-site surveys were used to (1) refine telephone survey data

---

82 "Bells, Whistles, and Common Sense: Billing Analysis of a Residential HVAC Rebate Program," (* Samiullah et al. 1996).

83 (* Samiullah, Goldberg et al. 1996).

84 "Improving Billing Analysis Results Using On-site Follow-up Surveys." (* Coito et al. 1996).

and (2) explicitly quantify non-program changes that altered electricity consumption and potentially masked program energy savings. The final regression model was a SAE Change Model with a vector of explanatory variables explaining change in consumption. The findings of this staged approach of engineering and billing analysis indicated that (1) the root mean square error was much lower than expected, (2) more observations were able to be included into a meaningful regression, (3) the coefficient for the on-site survey-based quantification of non-program factors provided evidence verifying the likely accuracy of this work, (4) the model was more stable with regard to the inclusion or exclusion of outliers, and (5) the program realization rate rose from 63% to 84%.

Billing analysis should examine outliers and their influence on the results, as shown in the above example. There are a variety of methods that can be used to treat outliers. These can include using a trimmed mean, weighting (bi-weight mean), and median, among others, as seen in Pigg and Blasnick.[85] Another potential method includes bootstrapping techniques.[86]

Summary Guidelines

The primary purpose of impact evaluation is to obtain the most accurate and unbiased estimate of energy and demand savings due to the program. A high quality billing analysis would produce these estimates in a manner that is defendable on the basis of current academic standards.

There are many types of errors, invalid assumptions, and threats to validity that can appear and may be very important to the final estimate of savings. A professional practice of examining the analysis with a critical "eye" and "knowing" its data are an important part of quality evaluation.

The best evaluations will include reasonable disclosure of potential biases, specification problems, and data limitations.

Engineering Analysis

Engineering methods use basic rules of physics to calculate estimates of energy and demand savings. The technical information required as inputs to engineering models generally come from manufacturers, research studies, and other general references combined with assumed or measured equipment operating characteristics.

In order to estimate savings via engineering methods, one must establish a ***baseline*** from which to compare the energy consumption and demand of facilities included in the

---

85 "Dealing with Outliers in Impact Evaluations Based on Billing Data." (Pigg et al. 1993).

86 "The Treatment of Outliers and Influential Observations in Regression-Based Impact Evaluation." (Schutte and Violette 1994). Another work that used bootstrapping and outlier correction, in this case mean absolute deviation estimator (MAD), can be found in "Billing Data Analysis of the C&I Sector: Application of Monthly Panel Models." (Ozog et al. 1995).

evaluation.[87] The program baseline may require a definition of the pre-program equipment or building characteristics and operations, as well as an estimate or measurement of pre-program energy consumption. The program baseline may consist of the following:

- For early equipment replacement (retrofit) programs, the pre-existing and still-functioning equipment replaced during program participation defines the baseline. Pre-program energy consumption may be adjusted to reflect changes in equipment or building operations not related to the program.
- For equipment that is being replaced at the end of its useful life (i.e., in all situations where the customer would have been replacing the equipment in the absence of the program), standard-efficiency new equipment defines the baseline. The program's purpose in these cases is to induce customers to do the replacement with a higher-efficiency alternative than they would have selected in the absence of the program.
- For operations and maintenance (O&M) programs (such as air conditioning tune-up or retro-commissioning programs), the existing condition of the equipment or existing O&M procedures define the baseline. Pre-program energy consumption may also be adjusted to reflect changes in equipment or building operations not related to the program.
- For new construction programs, the California Energy Efficiency Standards (Title 24) that define minimum standards for new construction, and Appliance Efficiency Regulations (Title 20) are used as the baseline. For program attributes not addressed by Title 20 or Title 24 (such as grocery store refrigeration systems), a "common practice" study may be conducted to establish the program baseline. Pre-program energy consumption data cannot be measured in new construction since the building does not exist. The energy implications of the baseline building characteristics are generally calculated using a building energy simulation program.

The Engineering Analysis Roadmap is presented in Figure 6.3. The various engineering analysis methods, literature references and examples are presented in this section. The net-to-gross methods, issues, and references are provided in a later section within this chapter.

[87] Engineering methods used to date estimate savings as the difference between baseline and post-program usage. This requires an estimation of baseline. Econometric methods can estimate baseline and post-program and use the difference to estimate energy savings. However, there are econometric methods that can directly estimate savings rather than this more indirect difference approach.

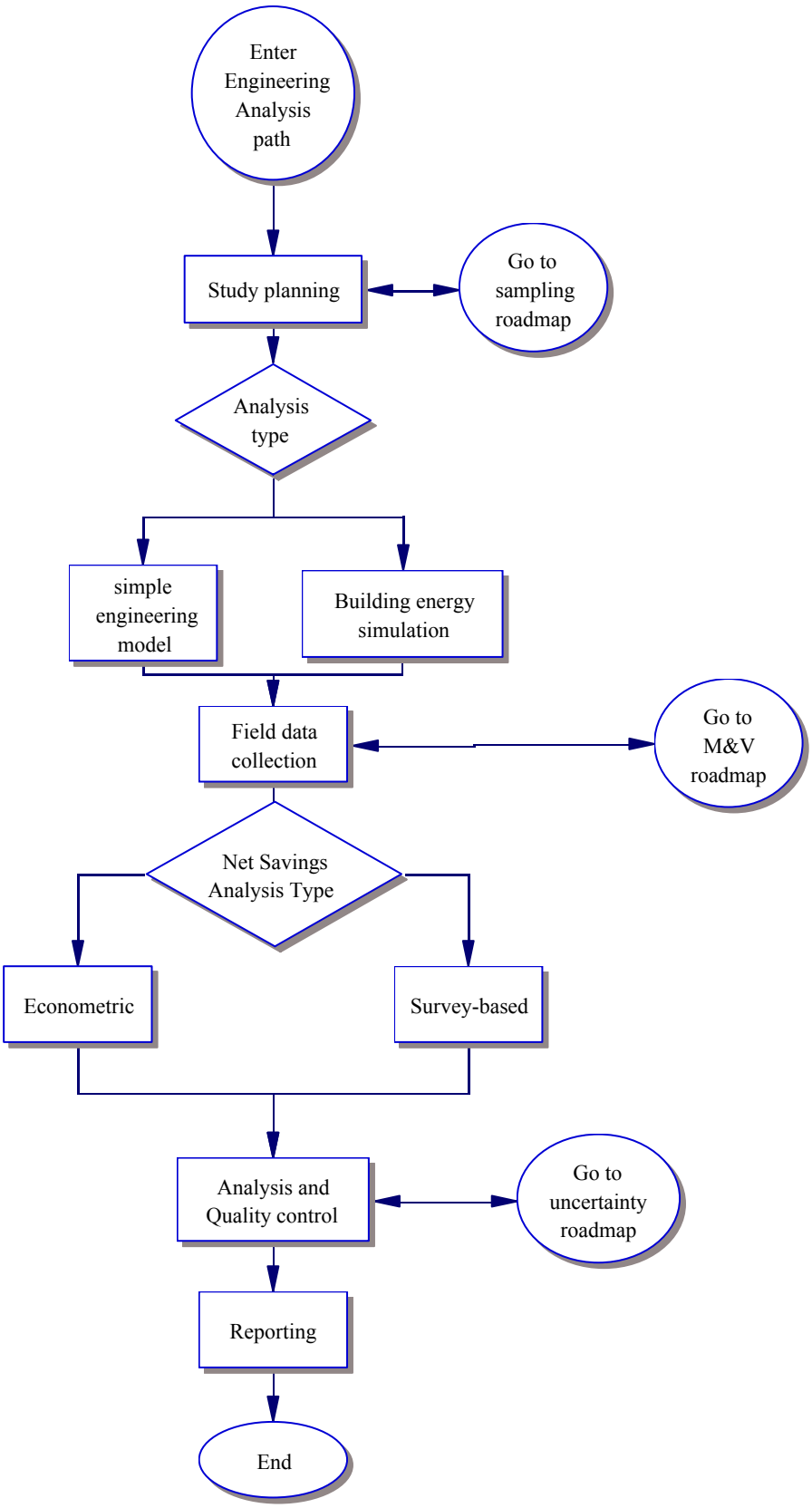

Figure 6.3: Engineering Analysis Roadmap

Engineering methods can be divided into two basic categories.

- Simple Engineering Model
- Building Energy Simulation Model

Simple Engineering Models

Simple engineering models and algorithms are typically straightforward equations for calculating energy and demand impacts of non-weather dependent energy efficiency measures, such as energy efficient lighting, appliances, motors, cooking equipment and so on. Simple engineering models are generally not used for weather dependent measures such as building envelope and HVAC measures; these measures are generally analyzed using building energy simulation models. HVAC interactions with internal loads such as lighting in conditioned spaces are generally calculated from a building energy simulation program. However, virtually all measures can be estimated using a simple engineering model, provided weather or load dependant parameters are developed by a more sophisticated method. Simple models also require knowledge of the efficiency and baseline operating conditions for the equipment that existed before the energy efficiency investment.

Simple engineering models are often incorporated into the program tracking system, to provide an initial savings estimate based on program accomplishments and building or project characteristics data collected during the course of the project. These estimates may also be used as part of a statistical billing analysis. See Chapter 13, Sampling, for more information on the use of program tracking data in the context of sample design.

The general form of the gross annual energy savings equation is as follows[88]:

$$\Delta kWh_{gross} = units \times \left[\left(\frac{kW}{unit} \times RLF \right)_{base} - \left(\frac{kW}{unit} \times RLF \right)_{ee} \right] \times FLH \times (1 + HVAC_c)$$

[88] Subscripts in the engineering equations are defined as:
base = baseline measure or technology
ee = energy efficient or enhanced measure or technology
s = summer
w = winter
d = demand
c = energy consumption
pk = peak
heat = heating system
cool = cooling system
Annual or seasonal average values will be denoted with a bar, such as $\bar{\eta}$, $\overline{EER}$, and so on.

where:

ΔkWh_{gross} = gross annual energy savings
units = number of technology units installed under the program
$\frac{kW}{unit}$ = nameplate power consumption of baseline and energy efficient units
RLF = rated load factor
FLH = full-load operating hours
$HVAC_c$ = HVAC system interaction factor for annual energy consumption

The general form of the gross summer demand savings equation is as follows:

$$\Delta kW_{gross,s} = units \times \left[\left(\frac{kW}{unit} \times RLF \times DF_s \right)_{base} - \left(\frac{kW}{unit} \times RLF \times DF_s \right)_{ee} \right] \times CF_s \times (1 + HVAC_{d,s})$$

where:

$\Delta kW_{gross,s}$ = summer gross coincident demand savings
units = number of technology units installed under the program
$\frac{kW}{unit}$ = nameplate power consumption of baseline and energy efficient units
RLF = rated load factor
DF_S = summer demand diversity factor
CF_S = summer coincidence factor
$HVAC_{d,S}$ = HVAC system interaction factor at utility summer peak hour

Terms for simple engineering equations
Units. Units refer to the number of individual technology units installed under the program, normalized to some convenient quantity, such as lighting fixtures, motor horsepower, tons of air conditioning, square feet of insulation, and so on.

KW/unit. This factor refers to the connected or "nameplate" load of the baseline or energy efficient technology, normalized to some convenient unit of measure. It is often referred to as the **unit demand**. The normalization parameter used should be consistent with the units defined above. In some cases, the number of units of technology installed at a particular location is different for baseline and energy efficient technologies (e.g., lamps or fixtures in overlit spaces may be removed during an energy efficiency upgrade). The value used for the baseline unit demand should include the power of all baseline technologies replaced by each unit of energy efficient technology.

The expression of the unit demand will depend on the type of technology evaluated. Lighting measures are characterized simply by the fixture input power. Motors are generally characterized by their nameplate horsepower, rather than input power. The unit of measure used in horsepower, and motor kW per horsepower is calculated as follows:

$$kW/hp = \frac{0.746}{\eta}$$

where:

hp = nameplate motor horsepower
η = motor efficiency (unitless)
0.746 = conversion factor (kW/hp)

Energy-efficient cooling equipment is characterized in a similar fashion. The measure quantity data is expressed in terms of tons of cooling capacity. The unit demand is expressed in terms of kW per ton of cooling, and is calculated as follows:

$$kW/ton = \frac{12}{EER}$$

where:

EER = energy efficiency ratio of the cooling equipment (Btu/Wh)[89]
12 = conversion factor (kBtu/ton-hr)

For measures involving increased levels of insulation, such as wall insulation or water heater insulation, the unit demand is calculated from the thermal load and the efficiency of the equipment meeting the load. When the units are expressed in terms of insulation surface area, the unit demand is:

$$kW/SF = \frac{U \times \Delta T}{\eta \times 3413}$$

where:

U = overall heat transfer coefficient (Btu/hr-ft^2-°F)
ΔT = temperature difference (°F)
η = efficiency of energy conversion device (unitless)
3413 = conversion factor (Btu/kWh)

For processes involving fluid flow, such as water heating and air infiltration, electricity demand is also calculated from the thermal load and the efficiency of the equipment meeting the load:

[89] The energy efficiency ratio (EER) as published by air conditioning equipment manufacturers refers to the efficiency of the unit under specific outdoor and equipment entering air conditions, which may not correspond to the peak or average load conditions in a particular climate region. Under peak conditions in the inland regions of California, outdoor temperatures are generally hotter and entering air conditions are generally drier than the standard conditions used to evaluate the equipment EER. Residential equipment is often rated with a seasonal energy efficiency ratio (SEER), which is intended to be representative of the average efficiency of an air conditioner over a national average cooling season. EER data may not be available for residential units, although the Consortium for Energy Efficiency (CEE) maintains a database of EER data for residential units. See www.cee1.org.

$$kW = \frac{\dot{m} \times c_p \times \Delta T}{\eta \times 3413}$$

where:

$\dot{m}$	=	fluid mass flow rate (lb/hr)
c_p	=	fluid specific heat (Btu/lb-°F)
ΔT	=	temperature difference (°F)
η	=	efficiency of energy conversion device (unitless)
3413	=	conversion factor (Btu/kWh)

Rated load factor. The rated load factor is defined as the ratio of the maximum operating power of a measure to its nameplate power. It is used to correct for discrepancies in the nameplate power input or capacity of a piece of equipment relative to its actual peak operating power or capacity. These discrepancies can come from:

- equipment over-sizing due to conservative design approaches used by architects and engineers
- margins of safety applied by manufacturers to the nameplate power rating

In the context of appliances and other plug loads, the rated load factor is also referred to as the usage factor.[90]

Demand diversity factor. The demand diversity factor is used to account for the fact that not all measures in all buildings are drawing full power at the same time. The demand diversity factor is defined as the peak demand of a population of units to the peak demand of an individual unit. It is estimated by comparing the peak demand obtained from load research or end use metering studies to the connected load inventory and rated load factor for all applicable loads.

$$DF = \frac{kW_{pop}}{\sum_{i=1}^{n} \left(kW_i \times RLF_i \right)}$$

where:

kW_{pop}	=	peak demand of population of units
kW_i	=	nameplate rating of unit i
RLF_i	=	rated load factor for unit i

[90] *The ASHRAE Handbook of Fundamentals*, Chapter 29 Nonresidential Cooling and Heating Load Calculation Procedures refers to the ratio of the nameplate load to the actual running load as the usage factor. (ASHRAE 2001).

Coincidence factor. The coincidence factor is used to account for the fact that peak measure savings may not be coincident with utility peak demands. The coincidence factor is defined as the portion of the technology, system, or end use demand reduction that is coincident with the system peak.

Full-load hours. Full-load hours are a measure of the annual equipment operating hours, and are used to calculate annual energy savings. Equipment full-load hours are defined as the total annual energy consumption divided by the equipment peak demand.

$$FLH_{eq} = \frac{kWh}{kW_{pk}}$$

For equipment servicing loads that do not vary with time (e.g., a motor driving a constant load), full-load hours are simply equal to the operating hours. However, the power requirements of many types of equipment (e.g., air conditioning) vary during their operating hours; thus, the equipment full-load hours are equal to the ratio of the hourly load to the maximum hourly load summed over all hours of operation.

Equipment full-load hour data are usually obtained from end use metering or load research studies. These data are a function of the system or process load requirements, ***and*** the efficiency of the equipment meeting those loads. When analyzing DSM measures with different efficiency characteristics than their respective baseline technologies, it is often useful to define the ***process*** full-load hours. When such equipment is replaced, the process load is not affected, yet the equipment full-load hours may change. The process full-load hours are related to the equipment full-load hours by:

$$FLH_{proc} = FLH_{eq} \times \frac{\bar{\eta}}{\eta_{pk}}$$

where:

$FLH_{eq.}$ = equipment full-load hours (calculated from load research or end use metering studies)

$\bar{\eta}$ = average equipment efficiency

η_{pk} = equipment efficiency under peak process load conditions

For example, the annual cooling load imposed on an air conditioner can be defined as the annual cooling load hours (CLH):

$$CLH = \frac{\text{Annual Cooling Load (Btu)}}{\text{Peak Cooling Load (Btu / hr)}}$$

Air conditioning equipment equivalent full-load hours are defined as:

$$\text{EFLH} = \frac{\text{Annual Cooling Energy (kWh)}}{\text{Peak Cooling Demand (kW)}}$$

Cooling load hours and equivalent full-load hours are related by:

$$\text{CLH} = \text{EFLH} \times \frac{\overline{\text{EER}}}{\text{EER}_{pk}}$$

where:

$\overline{\text{EER}}$ = Seasonal average cooling efficiency
EER_{pk} = Cooling efficiency under peak cooling load conditions

For many high-efficiency cooling machines, the ratio of the average cooling efficiency to the peak cooling efficiency is different than the baseline units replaced by the program. The process full-load hours (CLH) remain constant (assuming the building has not been altered in any other way), but the equipment full-load hours (EFLH) change as the ratio of the average to peak efficiency changes.

HVAC interaction factors. Certain measures, such as energy efficient lighting measures, directly influence lighting energy consumption and indirectly influence HVAC energy consumption. In certain climates, energy efficient lighting systems can reduce building internal heat gain, thus reducing cooling loads and increasing heating loads. Similar effects are produced by energy efficient appliances, cooking equipment and motors located in conditioned spaces. HVAC interaction factors can be defined for measures that have secondary influences on HVAC system energy. Separate HVAC interaction factors are defined for demand and energy savings calculations as follows:

- The HVAC interaction factor for ***demand*** is defined as the change in the HVAC demand per unit change in measure demand during the utility peak hour. Two factors are defined; the summer HVAC interaction factor generally considers the change in cooling demand, and winter HVAC interaction factor generally considers the change in heating demand.
- Similarly, the HVAC interaction factor for annual energy ***consumption*** is defined as the change in the annual HVAC electricity consumption per unit change in the annual measure electricity consumption. A single factor is defined which considers measure interactions with the annual heating, cooling, and/or fan electricity consumption.

There are many complicated interactions between internal gains, shell heat gains, thermal mass effects, HVAC system efficiency, and HVAC system control. Thus, HVAC interactions are best investigated with building energy simulation programs and for particular climate regions. Hourly simulation programs are best suited to calculating demand interaction factors. HVAC interactions are influenced by a number of factors, such as:

- Climate

- Relationship of building internal and solar loads to the overall envelope heat losses (balance point temperature)
- Coupling of internal load to HVAC load
- Building thermal mass
- HVAC system type
- HVAC system fuel
- HVAC system efficiency
- HVAC system controls

Thus, it is not appropriate to use a single HVAC interaction factor for all buildings, measures, and climates. HVAC interactions should be investigated over the range of climates and building characteristics listed above.

Simple engineering models can be calibrated through a combination of in-field metering and enhanced engineering analysis. Data collection is often carried out in a statistical framework, where the data collection and/or metering are conducted at a sample of sites. The results of the data collection are compared to the uncalibrated engineering estimate, and an adjustment factor is calculated.[91] The adjustment factor is then applied to the engineering model results for all participants. If the data collection is carried out according to a reasonable statistical sampling plan, tests of statistical validity can be applied to the results.

Building Energy Simulation Models

Building energy simulation models are computer programs that use mathematical representations of important energy and control processes in an attempt to realistically simulate the thermal and energy systems in a building. Energy calculations are carried out on an hourly or sub-hourly basis for a selected time period or more commonly for an entire year based on typical weather data for the selected building site. The programs are made up of a collection of mathematical models of building components, such as windows, wall sections, HVAC equipment, and so on. The individual component models are linked together to form a complete building simulation. The results predict the performance of the building structure and energy systems under given weather conditions at a selected geographic location. All building energy simulation programs have limitations that must be well-understood before applying the program to a particular energy estimation problem.[92] For example, most programs are limited to the simulation of common HVAC system types with a predetermined system configuration. Considerable latitude is given to the user with respect to describing system performance

[91] *Quality Assurance Guidelines for Statistical and Engineering and Self-Report Methods for Estimating DSM Program Impacts* (* CADMAC 1998) for more information on calibrated engineering models.

[92] For more information on building energy simulation programs, see *State-of-the-Art Review: Whole Building, Building Envelope and HVAC Component and System Simulation and Design Tools.* (Jacobs et al. 2002).

parameters, but the basic arrangement of the system component is fixed and defined by common practice in the building design and HVAC industries. This does not present a problem for most buildings and systems, but for complex custom HVAC configurations, the judgment and experience of the user is critical. In other situations, accurate weather data may not be available for the specific building site, and analogous weather data from a nearby location may produce some inaccuracy in the simulation results.

Below, a brief description of several building energy simulation programs commonly used in code compliance, evaluation, and building science research is presented.

DOE-2

DOE-2 is the most commonly used building energy simulation program for commercial buildings in the evaluation field. The DOE-2 is a ***public domain*** program, developed by the Lawrence Berkeley National Laboratory, with substantial funding from the U.S. Department of Energy. The first version of DOE-2 was released in 1979. The program has been modified and developed through 1999, when development was stopped in favor of the DOE-2.2 and EnergyPlus programs.

The final version of the DOE-2.1 program was DOE-2.1e. Within the DOE-2.1e version, there are 133 modifications or "releases." A new release of the program may be stimulated by a series of "bug fixes," or may include major new features. The California Energy Commission has adopted DOE-2.1e release 112 as the standard modeling program for non-residential Title 24 energy code compliance under the performance compliance option. This version is used in the EnergyPro Title 24 energy code compliance software.[93]

The DOE-2.2 program is built on the DOE-2.1e program. DOE-2.2 includes a number of upgrades and changes, including improved simulation of window heat gains, duct losses[94], heat recovery, and central plants in large buildings. DOE-2.3[95] includes the ability to model "refrigeration loops" that are made up of compressors, condensers, evaporators, and other components. This new functionality allows grocery store refrigeration and refrigerated warehouses to be simulated in detail.

EnergyPro

EnergyPro applies a user-friendly front end to the DOE-2.1e software engine to enable operation within a Windows environment. EnergyPro is certified by the California Energy Commission as a compliance tool for residential and non-residential buildings under California's Title 24 Building Energy Efficiency Standards. As required by the CEC, EnergyPro calculates energy use for heating, cooling and water heating for residential buildings; non-residential compliance calculations also account for lighting energy use. All forms necessary for California building permit submittal are output by EnergyPro. HVAC load calculations are available for both residential and non-residential buildings. EnergyPro is used as a modeling tool for the California statewide non-

93 (Gabel-Dodd/Energy Soft LLC 2001).

94 The duct loss model in DOE-2.2 includes infiltration into the return side.

95 Funding for the development of DOE-2.3 was provided by PG&E and SCE.

residential new construction program (Savings by Design) and is approved for California Energy Star Home qualification.

Micropas

Micropas is a detailed energy simulation program that performs hourly calculations to estimate annual energy usage for heating, cooling and water heating in residential buildings. Micropas is certified by the California Energy Commission as a compliance tool for California's residential Title 24 Energy Efficiency Standards. The program includes a load calculation for use in sizing heating and cooling equipment. All forms necessary for California building permit submittal are generated by Micropas. Micropas is approved for Energy Star Home qualification in California and nationwide.

Micropas has been in wide use in California since the early 1980s as a building energy code compliance tool and is growing in use elsewhere under the Model Energy Code. Up to 15 thermal zones can be defined. Heating, cooling, natural ventilation and water heating systems for residential buildings are simulated. No detailed modeling of heating and cooling systems is provided - seasonal performance values are used.

EnergyPlus

EnergyPlus is the next-generation building energy simulation program. Released in April 2001, the program was developed jointly by Lawrence Berkeley National Laboratory, the University of Illinois, the U.S. Army Construction Engineering Research Laboratory, Oklahoma State University and others, with support from the U.S. Department of Energy, Office of Building Technology, State and Community Programs.

EnergyPlus includes advanced simulation capabilities including time steps of less than an hour and modular HVAC systems simulation modules that are integrated with the building (zone) simulation. The architecture of EnergyPlus avoids several limitations inherent in the DOE-2 program. The response of the HVAC system to hourly zone conditions is greatly improved. HVAC systems of arbitrary design can be simulated, rather than relying on predefined templates. The program is seeing some use in the building science community, but has not seen wide use in the building design or evaluation community. Primary limitations are expanded data entry requirements, long run-times, and the lack of a user-friendly interface. Energy Plus is not certified for use as a Title 24 compliance tool in California.

The following table shows the relative applicability of each program to the analysis of energy efficiency measures in residential and commercial buildings. For more information, consult Jacobs and Henderson.[96]

[96] *State-of-the-Art Review: Whole Building, Building Envelope and HVAC Component and System Simulation and Design Tools.* (Jacobs and Henderson 2002).

Table 6.4: Comparison of Simulation Program Capabilities
(● indicates robust capability, ❍ indicates partial capability, blank indicates issue not addressed)

	MicroPas	DOE-2	Energy Plus
Air leakage sealing	●	●	●
CAD Interface		❍[1]	●
Code Compliance	●	❍[1]	
Cool Roofs	●	●	●
Cooling Towers / Fluid Coolers		●	●
Daylighting		●	●
Desiccant Dehumidification		●	●
Displacement Ventilation		❍	●
Duct Leakage	●	●[2]	
Economizers/Advanced Control		●	●
Exterior Shading	●	●	●
Ground Coupling	❍	❍	●
Heat Recovery Systems		●	●
High Performance Glass	❍	●	●
IAQ			●
Improved Unitary Equipment	❍	●	●
Interoperability			●
Lighting Design		❍	❍
Moisture Adsorption			●
Natural Ventilation	●	❍	●
Passive Solar Heat	●	●	●
Photovoltaics			●
Radiant barriers	●	❍	●
Radiant Cool / Heat			●
Refrigerant charge and airflow	●	❍	❍
Refrigeration		●	
Simplified Inputs		❍[1]	
Sunspaces, atria	●	●	●
System Sizing	●	●	●
Thermal Comfort		❍	●
Thermal mass	●	❍	●
Thermostatic expansion valves	●	❍	❍
Variable Speed Pumping		●[3]	●
VVT Systems		●	❍
Zonal HVAC systems	●	●	●

Note: 1. Simplified inputs, CAD interface and code compliance provided thorough third-party front-end programs
2. Full duct leakage modeling provided by DOE-2.2 only
3. Full variable speed pumping loop simulation provided by DOE-2.2 only

Building energy simulation programs should be calibrated with billing data and/or monitored data to improve the accuracy of the model. Data collection for model calibration should be used to improve the inputs to the model and to verify the response of the model. Data collection to improve model inputs generally focuses on the model

parameters with the largest influence[97] on the results and the greatest uncertainty. See Chapter 7 on Measurement and Verification for more information on simulation model calibration.

Building Energy Simulation Model Validation

The U.S. Department of Energy, through the National Renewable Energy Laboratory, has been working with the International Energy Agency and the American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) to develop standard methods of test for building energy analysis computer software. The Building Energy Simulation Tests (BESTEST) were developed to validate the response of building energy simulation programs. Several BESTEST protocols were designed to validate the simulation of building heating and cooling loads and the simulation of HVAC systems. ASHRAE recently published ANSI/ASHRAE Standard 140-2001, based on the BESTEST protocols.[98] DOE-2 and EnergyPlus have been validated under the ASHRAE 140-2001/BESTEST program.[99] Micropas has been validated under the building loads portion of the BESTEST procedure only.[100]

ASHRAE is considering an addendum to ASHRAE Standard 90.1 (*Energy Standard for Buildings Except Low-Rise Residential Buildings)* that would require any simulation program used to demonstrate ASHRAE 90.1 compliance to be validated using the tests contained in ASHRAE Standard 140. The authors believe that building energy simulation programs used for impact evaluation under this Framework should also be validated according to ASHRAE Standard 140. Simulation program validation should help identify weaknesses with the algorithms used within a program, thus removing an important source of bias from simulation results. Impact results developed from validated simulation programs that are calibrated to field data as described in Chapter 7 will provide the most reliable impact estimates.

Net-to-Gross Requirements and Method Options

The primary purposes of conducting a summative impact evaluation are (1) to provide reliable estimates of the energy and demand savings created by the program(s) for use in cost-effectiveness analysis, (2) to know how much resource can be depended upon, and (3) to incorporate savings in overall Public Goods Charge (PGC) fund estimates. In all of these cases what is really desired is the energy and demand savings induced by the program. In other words, the savings need to be "net" of what would have occurred in the absence of the program. Hence, we need to answer the question of what would participants (and non-participants) have done in the absence of the program that could affect their energy use level. This sounds like a simple question. But 15-20 years

---

97 Key model inputs are identified by calculating their influence coefficient. See (Spitler et al. 1989) for more information on influence coefficients.

98 (ASHRAE 2001).

99 "EnergyPlus Testing with ANSI/ASHRAE Standard 140-2001 (BESTEST)." (Henninger et al. 2003).

100 BESTEST validation reported on the US DOE Building Energy Software Tools directory: http://www.eere.energy.gov/buildings/tools_directory/software/micropas.htm

experience has proven it to be a very difficult question to answer with any assurance as to the (unbiased) accuracy or precision of the answer.

These doubts in the quality of these estimates are not for lack of trying. The increasing study and sophistication of evaluation efforts to improve upon this capability throughout the 1990s was far more extensive in California than anywhere else. Thorough analysis by expert econometricians has repeatedly found problems in each new generation of "solutions." Nevertheless, this section will summarize and reference several recent methods as possible choices.

Principles for Undertaking and Using Net-to-Gross Analysis

Often gross savings may be estimated by one method and net savings (the net-to-gross adjustment factor) estimated by another method. Sometimes, the analyses can be done hand-in-hand during billing analysis. Thus, a net-to-gross analysis should be conducted at the same time as gross savings impact evaluation. However, some of the methods can be costly, and evaluation resources can be limited. Accordingly, it is possible that net-to-gross analysis could be done less frequently than the gross savings impact evaluation, as long as a prior net-to-gross analysis is available at the time of the gross savings evaluation, and the program, market, and participants are suitably representative of those upon which the prior estimate is based. These factors need to be considered as evaluation resources are allocated in the evaluation planning process and as evaluations are designed to best meet the requirements of the evaluation goals and this Framework. (See Chapter 5, the Umbrella Roadmap, and the discussion of prioritizing evaluation resources. The allocation of resources between gross savings impact evaluation and net-to-gross analysis, both of which are used in the final net savings calculation, and the propagation of uncertainty calculations are discussed in Chapter 12, Uncertainty.)

The best available estimate of the net-to-gross ratio (NTGR) should always be used in program planning. The NTGR can be expected to vary depending upon the maturity of the equipment or service, the type of delivery in the program, the maturity of the program, and the customer sector. This means that the best NTGR estimate for program planning is the latest estimate for that program or a similar program. A deemed estimate based on one or more of these dimensions can be the fallback position for program planning.

The Framework recommends that deemed NTGR, however, never be used to report net savings within an evaluation (unless a net-to-gross method was attempted and no defensible estimate was able to be derived, i.e., the method failed and could not be corrected within the evaluation budget and priorities). The NTGR is an important input into cost-effectiveness analysis.[101] As such, we need to ensure that evaluations produce the most reliable and precise estimates possible. An assessment of the different factors that help determine the NTGR (as listed above) shows that, though very unlikely, an unscrupulous implementer could "game" an environment that used deemed NTGRs. A program could be run that was lower cost by providing services to the most predisposed

[101] See the *Energy Efficiency Policy Manual, Version 2* (* CPUC 2003), page 18.

participants who would have taken, or are in the process of taking, the actions anyway ("freeriders") and yet claim upwardly biased net savings due to a high deemed NTGR. Though this is unlikely, it is prudent that this evaluation framework be designed to protect the use of PGC funds from this type of abuse.

Selecting an Analysis Method

The primary question is what would have participants done in the absence of the program. An obvious simple solution would be to compare the equipment installation rate of participants with the equipment installation rate among non-participants. However, if a program offers technical or financial services to anyone who wishes to participate, then those planning to take these actions would receive benefits with very little additional cost to them (just the program transaction/hassle costs). Hence, most people likely to adopt would participate in the program. (These people are called "freeriders.") Similarly, if someone was going to install energy efficiency measures later, they might be inclined to install these measures early to take advantage of the incentives. (These people are called "partial freeriders.") This means that the non-participant group is not a good comparison group for the participant group. There is a self-selection bias in who is a participant versus a non-participant, i.e., participation is not random. (This selectivity issue (among others) will continue to show up as a problem through the econometric methods.)

A simple comparison can demonstrate the nature of the selectivity bias. In this example, let's assume that twelve (12) percent of non-participants are found to install the same efficiency measure. A simple comparison might assume then that 12 percent of participants are freeriders. However, recognize that some participants participated because they were going to make the installation anyway and wanted to receive "free incentives" for their planned activity, resulting in fewer non-participants in the non-participant pool installing measures. The proportion of "installing non-participants" is lower and the proportion of freeriders is higher than would be indicated by comparing to the non-participant sample.

The impact evaluation roadmap categorizes NTGR methods very simply into those that are econometric methods (comparing participant and non-participants and adjusting for selectivity biases through econometric models) and those that are survey-based (asking participants what they would have done).[102] There are a variety of econometric methods that have been used with varying advantages and disadvantages. There are also a variety of more and less sophisticated survey-based methods that can include file review, review of letters in business decision process, payback exams and comparisons with other decision-making in the participating firms, engineering modeling, etc.

Econometric methods are sometimes considered the most accurate type of method, and they are preferred in California in situations where there are enough participants and comparable non-participants, and when the program is large enough to justify the

[102] At the same time, many of the econometric methods depend upon survey data from participants and non-participants to be used in the econometric model.

expense of the method. However, there are a few program types that make certain NTGR methods unfeasible. Programs with either a very small number of participants or non-participants (e.g., in sectors where there are only participants and former participants) or where comparability is a severe problem (e.g., programs that only work with large manufacturers) are not amenable to these econometric methods. This means that this small subset of programs needs to rely on a survey-based method. These could include participant surveys, vendor interviews, and/or record review methods.

An overview of the NTGR method options and examples of their use are provided in the discussions that follow.

Survey-Based Methods

Survey-based stated intentions, or "self-reports," is a method of estimating freeridership by asking participants directly a series of questions on what they would have done in the absence of the program. Generally, the best use of this method has involved asking a series of questions with each question allowing a scale of responses. The initial question asks the participant, if the program had not existed, would they have installed the same equipment? The scale of responses for this question is whether they "definitely would have," "probably would have," "probably would not have," or "definitely would not have" installed the energy efficient equipment. This scale, rather than a yes/no response, is thought by many to allow greater apparent accuracy in the estimate.

The program can also have an effect on the level of efficiency of the efficient equipment installed (when they say they would have installed the same efficient equipment without the program), when a participant installs efficient equipment, and how many they install. This is the element of partial freeridership. The continuum of freeriders is presented in Figure 6.4.

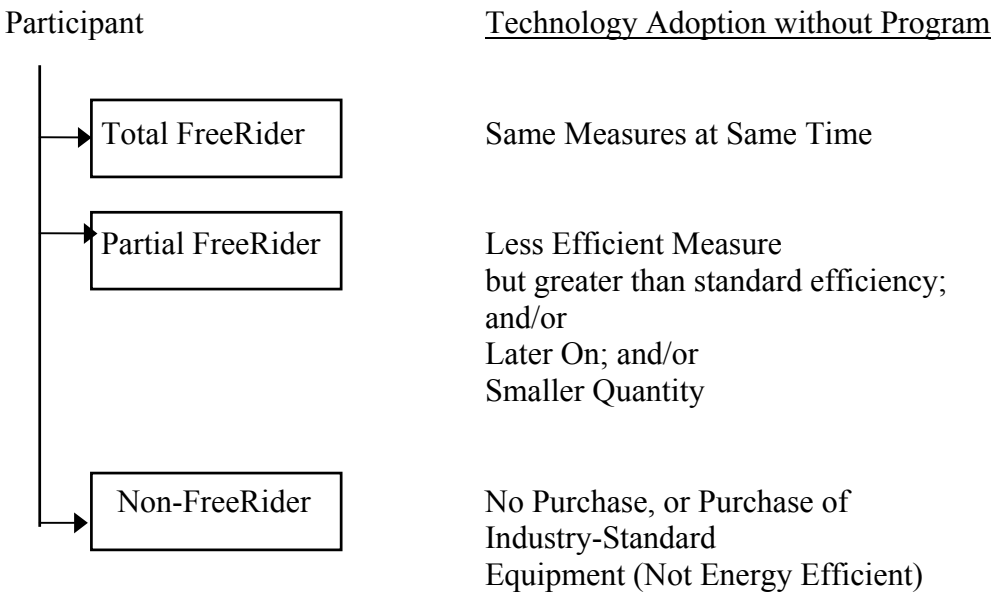

Figure 6.4: Continuum of Free Riders

The partial freeridership issue presented in Figure 6.4 is generally addressed through additional questions in a series of questions about stated intentions that are contingent on the response in the primary question, i.e., the questions are only asked of those that probably or definitely would have taken the actions. Examples of partial freeridership questions include the following:

1. If you had installed measures without the program, do you think the equipment would have been: "definitely as efficient," "probably as efficient," "probably not as efficient," or "definitely not as efficient."
2. If you had installed without the program, do you think the [measure] would have been installed at the same time?
3. If not installed at the same time, when would it have been installed? (Such as: earlier or later, or within three months, six months, next year, etc.)
4. Without the program, would you say the quantity of high efficiency measures you purchased would have been: "definitely greater," "probably greater," "the same," "probably less," "or definitely less."

Each of the responses in the series of questions is assigned a probability for the expected net savings. These estimates are then combined (additively or multiplicatively) into a participant estimate. The participant estimates are subsequently averaged (or weighted averaged given expected savings) to calculate the overall freeridership estimate. The assignments of the probabilities are critical in the final outcome. At the same time, there is little evidence of what these should be and they are often assigned and justified given a logical argument. With this, however, a multiple number of different probability assignments have been shown to be justified and accepted by various evaluations and

regulators. However, this can make the comparability and reliability of survey-based estimates problematic.

The survey method is the most straight-forward method of freeridership estimation, and the lowest cost method. It does, however, have its disadvantages in potential bias and with accuracy. One method for improving this estimate is to add a consistency check question and adjust the individual's estimate accordingly. For example, asking the participant how important the incentive was in their decision to select the energy-efficiency equipment. If the answer from this question and the overall freeridership estimate from the prior questions are completely inconsistent, then the estimate for that participant is either modified, or additional questions are asked to clarify and obtain a better estimate for that participant, or the participant is removed from the analysis.[103]

Early methods to handle responses of "Don't Know," missing data, and inconsistent answers involved assuming a 35% or 50% freeridership rate for these participants (as they might be less likely to have taken actions if they hadn't thought about it or made opposing reactions). These methods, however, were found to create a centrality tendency in the overall freeridership estimate, i.e., driving it towards 35% or 50%.

For example, a study in 1993 that created a standardized survey method for the New York utilities used these replacement assumptions. Then in 1997 the centrality bias was found to create NTGRs that were not reasonable in a study conducted for the Long Island Lighting Company (LILCo). This work created a NTGR estimate based upon removing the large number of inconsistent respondents from the analysis, rather than the prior procedure of substituting a central estimate for them, a revision to the standard method approved by the New York Public Utility Commission.

Due to this centrality tendency, the Framework recommends dropping respondents with inconsistent responses rather than assuming central estimates for them. Nevertheless, using additional survey questions to clarify and provide better estimates for these participants would be significantly preferable to losing these participants in the NTGR analysis. (Computerized surveys could be designed to check for the inconsistency and then add appropriate survey questions as needed to try and salvage a reasonable estimate for that respondent.)

Other questions have been used to help provide backup or adjustments for the freeridership estimates. For example, one can evaluate the decision-making process and identify where the participant was in that process when they learned about the program. This method is particular useful for customers that have difficult and time-consuming processes, such as government and large commercial customers. There have also been other studies examining information dissemination, other decision priorities, and general criteria for commercial investments.[104]

---

[103] *DSM Free Ridership Study* (Cambridge Systematics Inc. et al. 1994).

[104] An example of using alternative questions and comparing their responses can be found in: *DSM Free Ridership Study* (Cambridge Systematics Inc. and Freeman Sullivan and Company 1994); and "FreeRidership Estimation in the New Construction DSM Market." (* Tolkin et al. 1993).

Generally, the simplest and lowest cost NTGR method is using the survey-based stated intentions method with a telephone survey for data gathering. Unfortunately, there has been significant research that this method can easily provide biased results.[105] One type of bias is that respondents can overestimate in reporting what they would have done, either because they see that response as socially desirable (social desirability bias, potentially a type of "testing" internal validity threat), or because respondents do "intend to" save energy, but there are many things people can intend to do that never happen or are postponed significantly. These problems can also be increased depending upon the wording of the survey questionnaire, location of the survey questions (response effects' bias), and the length of time involved in recall (the telescoping problem).[106] As a result, one study concluded that the survey-based method is most likely to be biased.

> *"This is the most important component of an estimate's validity but can not be measured from statistical confidence intervals. Inexperienced analysts can easily be deceived that their answer is quite accurate when their survey instrument or assignment strategy is biased."*[107]

Survey-based stated intentions is a classic case of potential bias in measurement, non-random measurement error or an issue with construct validity (i.e., we are not properly measuring the construct – free ridership - we are trying to measure). Coming at the question of what the participant would have done in the absence of the program from a variety of different perspectives (directly asking, decision-making criteria, where they were in the process, etc.) and assessing these together is one way the survey methods have used to triangulate on the correct construct.

A few studies have specifically examined how we can improve the validity of our surveys and minimize their potential biases. One study found that strict, factual, and clearly defined terms appear to minimize social desirability bias.[108] Another study tested for the ability of survey design to improve the accuracy of self-reports by using graphics in WebTV surveys and assessing false positives and false negatives. The evaluator reported that there were circumstances when the graphics were quite useful and others where they were difficult to employ, or without enough advantage to justify other issues that must be considered e.g., depending upon WebTV surveys in cases where that might not be appropriate.[109]

---

105 *Net-to-Gross Ratios for PG&E's CIA Rebate Program* (Xenergy et al. 1993); and *DSM Free Ridership Study* (Cambridge Systematics Inc. and Freeman Sullivan and Company 1994).

106 The terms provided in parentheses are from research and survey design. Experience with these elements and this literature are necessary for designing a quality evaluation study that uses surveys for data gathering.

107 (Cambridge Systematics Inc. and Freeman Sullivan and Company 1994).

108 "Sure You Do. Un-Huh': Improving the Accuracy of Self-Reported Efficiency Actions." (* McRae 2002).

109 "Effects of Using WebTV Graphics on the Accuracy of Self-Reports." (* Peters 2002).

We can never know the "true" freeridership rate.[110] Survey methodology can provide a measure of relative precision that appears good based upon large sample sizes.[111] This is not the same thing as being confident that the answer is accurate. (See the Chapter on Uncertainty for a discussion of precision versus bias.)

Finally, survey-based methods may be appropriate for smaller programs when this is the only reasonable (affordable) NTGR method to use.

Given the ease with which the survey instrument and assignments can create varied NTGR estimates, and therefore, the potential for this to be "gamed" to obtain underestimated NTGRs, the Framework recommends that the CPUC consider supporting an overarching evaluation study to develop a standardized survey method for use in California. A carefully constructed survey with checks and triangulation methods might offer a more "accurate" measurement of a particular program's freeridership, particularly given differences across programs. A standardized survey method may not provide greater "accuracy." However, it can allow a freeridership survey method to be cost-effective for small programs to undertake, and it ensures consistency of measurement technique and application across these programs. This can help facilitate meaningful comparisons across programs for portfolio analysis, reporting, and planning purposes.

Massachusetts regulators recently asked for a study to derive a standardized freeridership survey method to be used by all the Massachusetts utilities for all their program evaluations. This study was completed in mid-2003, and the first program evaluations using it are just being completed.[112] The CPUC may want to use this study or conduct a similar study for California programs.

There are two improvements to the Massachusetts study that the Framework recommends for consideration if a standardized method were to be developed for use in California. First, a method needs to be developed for using contingency questions and for dropping responses from inconsistent respondents from the analysis if the inconsistency cannot be corrected (rather than substituting a central measurement). Second, testing and analysis is needed to calibrate the survey against a mix of econometric methods as discussed below.

---

[110] There have been comparison studies between survey-based methods and nested logit methods that pointed to the differences as illustrating the bias in the survey-based methods. However, bias could also exist in the nested logit method and concerns with this method have been discussed in later work. These earlier comparisons can be found in (Xenergy and Cambridge Systematics Inc. 1993) and *DSM Free Ridership Study* (Cambridge Systematics Inc. and Freeman Sullivan and Company 1994).

[111] The confidence level is providing precision that the sample is measuring the same answer as would be received if the full population had responded. The problem is not one of sampling but of construct validity. That is, do the specific questions asked and the response assignments, weights and combining of question accurately measure what would have occurred in the absence of the program.

[112] *Standardized Methods for Free Ridership and Spillover Evaluation – Task 5 Final Report (Revised)* (* PA Consulting Group Inc. 2003).

Examples of Survey-Based Methods with Record Review and/or Engineering Components

Asking a series of questions on what participants would have done in the absence of the program is problematic in complex and dynamic sectors, such as the non-residential new construction market, industrial projects, and large customized projects. In some cases, such as with new construction, the decision process involves many steps, there is no comparison point before action is taken, and it is difficult for the participant to imagine what "would have occurred." In other cases, participants are hesitant to provide unbiased information given perceptions about their firm, confidentiality on how investment decisions are made, or perceptions of how their answers might affect their relationship with the program, the utility or future funding of the program. This complication has been dealt with through careful, and sometimes customized, interview design and analysis methodologies.

For example, Goldberg and Scheuermann present work where an interview script was designed to be administered by an evaluation engineer familiar with the project and able to apply the inquiry with a discussion of project details on-site for large and custom projects. The project detail discussions can help the participant focus in on each component of the decision for a complex project or decision-making process.[113]

In another study, information was gathered for a net-to-gross analysis of an industrial program which included a review of program files, operations staff surveys that were used to create customized freeridership questions for a decision-maker survey, a decision-maker survey, an on-site survey, and a vendor survey. This multi-step process allowed customized questions to be created to better target and disaggregate the complex projects and the decision-making being examined.[114]

There are advantages of using multiple methods as either back-up or triangulation to obtain the most defensible NTGR estimate, particularly in cases with complex projects and decision-making. Differing circumstances can also significantly affect the accuracy of a NTGR methodology.

For example, in an evaluation of California's non-residential new construction program, Savings By Design, two different methods were used. One method was based on site visit data, baseline study, and engineering models, creating an engineering-based difference of differences between participant and non-participants (i.e., comparing participants versus non-participants on a baseline model estimate versus as-built model estimate). The second method included self-reported intentions soon after construction for both participant and non-participants and combined them with engineering modeling. The evaluators noted that increasing market effects and potential impacts from the California energy crisis might have been showing up as non-participant spillover. With this in mind, they reported a problem with the difference of differences methodology. This method left out non-participant spillover and underestimated the NTGR from

[113] "Gross and Net Savings Analysis for Unique Projects." (Goldberg et al. 1997).

[114] "A Combined Engineering and Decision-Analysis Methodology for Evaluating Spillover and FreeRidership in PG&E's 1995 Industrial Energy Efficiency Program." (Reid et al. 1997).

freeridership. The second method, using self-reports and engineering data, provided a more defensible (where anyone might believe) NTGR in the presence of non-participant spillover for the second year.[115]

Econometric-Based Methods

The purpose of the econometric-based methods for estimating the NTGR is generally to use a comparison with non-participants and control for differences and the selectivity bias inherent in the participation decision. This section's discussion focuses upon selectivity bias. It is important to recognize that using a non-participant comparison group controls for many other occurrences other than whether the measure would have been adopted or not, such as: changes in the economy, behavioral trends, fads, and other historical changes.

Some impact billing analysis models attempted to measure net savings by performing a billing analysis on both participants and non-participants with one dummy variable to indicate the post-retrofit time period for both participants and non-participants (as a control for exogenous effects on both) and a second dummy variable for only participants during the post-retrofit period. The coefficient on participation was meant to measure "net savings." (This variable is zero for non-participants. This means that the participant post-retrofit change is allocated between the post-time period dummy where all the non-participant post data exist and the post-participation dummy.) This type of analysis does generally lower the savings estimated by the billing analysis compared not including this selectivity correction (except in times of booming economic growth or due to historical issues). Yet, this does not mean that it is a reliable estimate of net savings. The post-period change for participants is pulled downward by the non-participant data being within that variable, but there is still a selectivity bias that exists. Given the selectivity bias in participation, the savings from this analysis is an overestimate of net savings (as the reduction in the post-period for non-participants is less than the participants would have seen as some greater proportion of participants would have adopted the actions).[116,117]

A common correction methodology for standard selectivity bias in econometrics is the Heckman correction factor by using an inverse Mills ratio. Heckman first proposed this use in 1976 in labor economics when examining the effect of education on wage equations with truncated data, i.e., a selectivity that if a laborer's wage to be received is lower than their "reservation wage" (the lowest wage for which they are willing to work), no wage would be observed for them.[118] The method was then used in other areas

---

115 "Measuring Accomplishments of Energy Efficiency in California's Nonresidential New Construction Market." (* Brost et al. 2002).

116 An example of a simple participant/non-participant comparison adjustment factor that incorporates these issues can be seen in "Matching Methodologies and Program Types: An Evaluation Retrospective." (Conant et al. 1993).

117 This argument was presented in "Why Discrete-Continuous Billing Models Mis-Estimate Net Savings of DSM Programs." (* Paquette 1996).

118 "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models." (Heckmann 1976).

beginning with employment and job training. The method has also been refined and modified across a variety of applications.[119,120]

The Heckman method of using an inverse Mills ratio was then used in energy efficiency analysis. Yet, it has since been shown that the type of selectivity bias that these methods address are not the same as the participation selectivity bias found in energy efficiency program evaluations.[121] For example, in the above original use of the correction factor, the selectivity is caused by truncated data; those without work have no wage data and are more likely to have the lowest levels of education. This logic is equivalent to the fact that customers more likely to install the measures are more likely to participate in the program (and they cannot participate in the program without installing). This is the primary selectivity issue discussed above. In addition, however, there is a second correlation of concern for the econometric modeling. Those with greater savings over what they would have done otherwise (those with lower freeridership percentages) are more likely to install and participate in the program. (It is this second factor that led to the use of the double inverse Mills ratio, as discussed below.)

A third method used to correct for selectivity is a two-stage regression where the first stage estimates the probability of participating in the program, which would then become a variable in the second stage that ran the billing analysis for net savings. Given that the probability to participate is a discrete choice (you either participate or not), a discrete choice analysis is needed to estimate this regression. The discrete choice model includes variables hypothesized to lead to the decision to participate in the program. The billing analysis then contains variables related to energy usage plus the probability of participation (estimated from the discrete choice analysis). There have been a number of impact evaluations estimating net savings through this method (termed discrete-continuous billing analysis).[122] This approach yields higher estimates of freerider savings and lower estimates of net savings than the basic participant/non-participant billing analysis, which is what would be expected with a correction for this type of selectivity bias. However, the discrete-continuous method only corrects for one of two decisions that may differ between participants and non-participants, the decision to participate. There may still be a selectivity bias due to differences in the decision to install the equipment. Hence, the method may be increasing non-random error and uncertainty.[123]

Nested logit models have been used to estimate naturally occurring conservation investments or actions to derive net-to-gross ratios. The theoretical basis for this work pointed out that the nested logit could address the joint decision-making between installation and program participation. Nested logit attempted to solve this problem in

---

119 *Time Series Techniques for Economists*. (Miller 1990).

120 "Sample Selection Bias as a Specification Error." (Heckmann 1979).

121 *Net Savings Estimation: An Analysis of Regression and Discrete Choice Approaches*. (* Xenergy 1996).

122 An example can be found in "Combining Monitoring, Engineering Analysis, and Billing Analysis to Evaluate PG&E's Commercial Retrofit Incentive Program." (Buller et al. 1993).

123 "Why Discrete-Continuous Billing Models Mis-Estimate Net Savings of DSM Programs." (* Paquette 1996).

that it jointly estimates the probability of installations and the probably of participation given that the measure will be adopted.[124] Nested multinomial logit (NML) analysis examines three possible outcomes: (a) do not install the measure, (b) install the measure and do not obtain a rebate, or (c) install the measure and obtain a rebate.[125] The nested logit model explicitly tests the joint decision-making between the decision to participate and the decision to adopt the technology.[126] This structure can be seen in Figure 6.5.

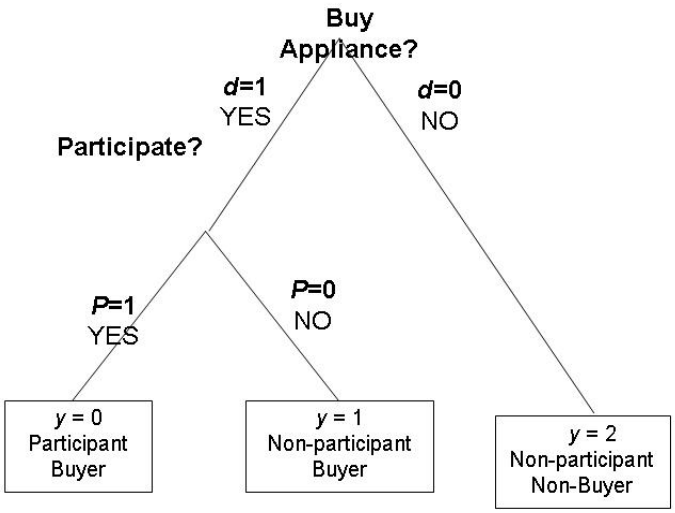

Figure 6.5: The Nested Logit Model for Adoption and Participation Decisions

Source: "Instrumented Decomposition: A Two-Stage Method for Estimating Net Savings."[127] Similar graphics available in several of the other publications cited concerned use of the NML method.

An alternative econometric method developed was to use two inverse Mills ratios - the first was applied to the entire sample and a second was applied to only the participants. This method is known as the Double Mills Ratio method. It was developed to address the particular type of selectivity bias found with energy efficiency programs that is not properly addressed by the Heckman inverse Mills Ratio method.[128] A subsequent study

---

124 "Estimation of Net Savings from Energy-Conservation Programs." (Train 1994). Another comparison was made with Monte Carlo experiments with simulated data sponsored by the Wisconsin Center for Demand Side Research as tested by Goldberg and Kademan: "Is It Net or Not? A Simulation Study of Two Methods." (Goldberg et al. 1995).

125 "Estimation of Net Savings for Rebate Programs: A Three-Option Nested Logit Approach." (* Train et al. 1994); "Modeling Technology Adoption and DSM Program Decisions." (* Buller et al. 1994).

126 See also: "Freerider and Freedriver Effects from a High-Efficiency Gas Furnace Program," (* Seiden et al. 1999); "Developing Confidence in Your Net-to-Gross Ratio Estimates," (* Mast et al. 1996); "A Comparison of Two Net Analysis Methods Using Data from PG&E's Nonresidential New Construction Program," (* Heitfield et al. 1996); "Do Central Air Conditioner Rebates Encourage Adoption of Air Conditioning?" (* Samiullah et al. 2002).

127 (Kandel 1999).

128 A similar method was developed for a similar problem by James Heckman as published in "Alternative Methods for Evaluating the Impact of Interventions," (Heckmann et al. 1985). Thank you to Adrienne Kandel for provided us this information and other clarifications within this section.

tested this method in an evaluation of a commercial HVAC program,[129] and another study has used this method in a market effects study.[130]

In 1995/1996, a comparative analysis of the discrete choice method (NML) and the double inverse-Mills-ratio approach was conducted. The study confirmed the use of NML for rebate programs. It also found the double inverse Mills ratio could properly correct for self-selection bias given specific underlying population distributions. At the same time, the results and potential bias for using the double inverse Mills ratio could be worse than omitting the self-selection correction at all for some distributions.[131]

The distribution issue for using a double inverse Mills ratio led to the development of "instrumented decomposition" (ID) as a method that did not need to assume any particular error distribution.[132] This method used NML to estimate the correlation of errors and then used an instrumental variable in a two-stage least-square model (2SLS). The ID method has been tested with data from an evaporator cooler program[133] and an air conditioning rebate program.[134] This method continues to be refined.

Most econometric methods for NTGR require survey information. The more they rely upon self-report type data, intentions, and psychographic data, the more they are likely to have some of the same measurement issues as the survey-based approach. The ID method seems promising. At the same time, there may be extensive data requirements for this method that might limit its applicability to programs that either have or can justify large data collection costs.

Summary of NTGR Method Options and Roadmap Choices

In summary, there are many econometric methods and survey methods that can be applied to the question of estimated NTGR. Most, if not all, contain issues with potential bias, limiting assumptions, or lack of outside verification and testing. On the other hand, many methods may provide defensible estimates for a wide variety of specific programs. The Framework recommends that a method be selected in the evaluation planning process that is consistent with the type of program, its program theory, knowledge of the data collection effort that can be accomplished, and a critical approach to measurement and assessing potential problems.

---

129 "How Many Mills Ratios Does it Take to Estimate Net Savings?" (* Randazzo et al. 1996).

130 "Self-Reports and Market Transformation: A Compelling New Approach," (Cavalli et al. 1999).

131 *Net Savings Estimation: An Analysis of Regression and Discrete Choice Approaches*. (* Xenergy 1996)

132 "Instrumented Decomposition: A Two-Stage Method for Estimating Net Savings." (Kandel 1999).

133 "Evaporative Cooler Rebate Program Cuts Load Significantly, and May Overcome Class Barrier." (Kandel 1999).

134 "Instrumented Decomposition: A New Method to Estimate the Net Energy Savings Caused by Efficient Appliance Rebate Programs." (Kandel 1999)

Chapter 7: Measurement and Verification

Preface

The Umbrella Framework and Impact Evaluation (Chapters 5 and 6) provide some guidance on the role of measurement and verification (M&V) in the context of conducting an impact evaluation. This chapter provides more details on conducting M&V studies. As such, it can be useful for policy-makers and implementers needing an overview of M&V methods and issues.

During the period of 1994 to 1997, California's measurement and evaluation protocols (sometimes referred to as the CADMAC Protocols) referenced the National Association of Energy Services Companies (NAESCO) standards for measurement and verification in Appendix H as a resource for M&V activities. The NAESCO protocols were the precursors to the International Performance Measurement and Verification Protocols (IPMVP)[135] established by the US Department of Energy in 1996. The California Energy Efficiency Policy Manual,[136] first published in 2001 and revised in 2003, references the IPMVP, directing evaluators to "reference the appropriate IPMVP option" and "state any deviations from (the) IPMVP approach" when developing evaluation plans for programs by the Public Goods Charge (PGC). This chapter provides additional information on the IPMVP, attempts to clarify the intent of the various IPMVP options as they relate to impact evaluation studies, and describes how these activities support the overall evaluation process.

M&V and Process Evaluation

Measurement and verification studies can provide important information to program implementers regarding measure installation verification, installation quality, manufacturing defects, measure use and operation, equipment maintenance procedures, and in-situ measure efficiency that can improve program design. While billing analysis is useful in providing energy savings, it generally does not provide measure-specific performance information or insights into why measure performance might be less than expected. Process evaluation and M&V activities conducted early in the program implementation cycle can provide important early feedback on potential program or measure problems to implementers at a point where effective corrective action can be taken. Thus, M&V can serve an important and complementary role with process evaluation studies.

[135] (International Performance Measurement and Verification Protocol 2001).
[136] (* CPUC 2003).

Skills Required for Measurement and Verification

Simple engineering equations and simple instrumentation such as run-time data loggers can be understood and used by most people with a general science background.[137] Large-scale data acquisition systems and complicated analysis techniques generally require someone with a graduate degree in mechanical or architectural engineering.

This chapter provides an overview of the methods used in M&V. It provides background and references for policy-makers, evaluation managers, and evaluators new to this field but with some knowledge of M&V issues. It is not meant to provide a "how-to" manual. The discussion here is meant instead to provide an overview and a perspective on how M&V methods are applied in the task of evaluating energy efficiency programs. References to the literature are provided for more information on M&V.

Introduction and Key Issues

The purpose of M&V activities is to verify that measures promoted by a program were actually installed and to measure the gross electricity savings from the measure installation relative to some baseline pattern of use. M&V, in this context, refers to data collection, monitoring and analysis activities associated with the calculation of gross energy savings from ***individual customer sites or projects.*** Program level gross and net impacts will be guided by the impact evaluation chapter, which also contains a sampling section to guide decisions about individual site selection. The M&V chapter is a subset of the overall impact evaluation process. M&V activities will primarily support program impact evaluations. Protocols for data collection to support other evaluation activities will be covered in their respective sections.

Performance contracting programs, such as the Standard Performance Contract (SPC) program, require M&V as a component of the program offering. Incentives are paid to implementers based on measured energy savings. The M&V activities conducted for the program may also provide information for program evaluation.

The Role of the International Performance Measurement and Verification Protocol

The IPMVP is the basis of the M&V activities covered in this chapter. The IPMVP was first published in 1996 as the National Energy Management Verification Protocol (NEMVP) and has since gone through several revisions. The current version (as of this writing) was published in the last few years in three volumes:

- ***IPMVP Volume I: Concepts and Options for Determining Savings (2002 Edition).*** Volume I defines basic terminology useful in the M&V field. It defines general procedures to achieve a reliable and cost-effective determination of

[137] The Association of Energy Engineers (AEE) offers a certificate for a Certified Energy Manager (CEM). The material covered in the CEM program is good background for understanding energy engineering concepts addressed by measurement and verification.

savings. Verification of savings is conducted relative to the M&V Plan for the project. This volume is written for the general application of measuring and verifying the performance of projects improving energy or water efficiency in buildings and industrial plants. This volume is the primary reference for the Framework.

- ***IPMVP Volume II: Concepts and Practices for Improving Indoor Environmental Quality (2001 Edition).*** Volume II reviews indoor environmental quality issues as they may be influenced by an energy efficiency project. This volume focuses on measurement issues, project design, and implementation practices associated with maintaining acceptable indoor conditions for an energy efficiency project, while providing guidance on key related elements of M&V and energy performance contracts. Issues discussed in this volume have limited relevance to impact evaluation, but may be useful when quantifying indoor environmental quality improvements. See Chapter 11 on Non-Energy Effects for more information.

- ***IPMVP Volume III: Applications (due to be published in 2004).*** Volume III will provide guidance on application-specific M&V issues. Individual chapters on *Concepts and Options for Determining Energy Savings in New Construction* and *M&V for Renewable Energy Technologies* are available.

The IPMVP was originally designed as a protocol to verify energy-savings projects implemented by energy services companies (ESCOs) under a shared-savings type contract or a guaranteed savings contract. It has since found applications to a broad variety of energy and water conservation projects throughout the world, including recent work on the verification of energy-savings projects for climate change mitigation.[138] The CPUC's *Energy Efficiency Policy Manual, Version 2*[139] references the IPMVP protocols as a key component of an overall Evaluation, Measurement and Verification (EM&V) plan.

There are several related documents that reference and complement the IPMVP:

Federal Energy Management Program (FEMP) M&V Guidelines.[140] This document is based on the 1997 version of the IPMVP, and provides more detailed guidance on the application of different M&V protocol options for specific energy conservation measures.

ASHRAE Guideline 14-2002 – Measurement of Energy and Demand Savings.[141] ASHRAE Guideline 14 provides additional detail on implementing M&V plans within

---

138 IPMVP is listed as the preferred choice for monitoring and evaluating energy efficiency projects for climate change mitigation in "Guidelines for the Monitoring, Evaluation, Reporting, Verification and Certification of Energy-Efficiency Projects for Climate Change Mitigation." (Vine and Sathaye 1999).

139 (* CPUC 2003).

140 Federal Energy Management Program (FEMP) M&V Guidelines: Measurement and Verification for Federal Energy Projects, Version 2.2. (* FEMP 2000).

141 ASHRAE Guideline 14-2002 – Measurement of Energy and Demand Savings. (* ASHRAE 2002).

the IPMVP framework, including detailed HVAC system examples. It also contains detailed information on quantifying and minimizing uncertainty in M&V plan design, data acquisitions system selection and design, sensor selection, sensor placement, and calibration.

Statistical Context for M&V Studies

M&V studies are generally not conducted at every individual customer site within a program. The selection of M&V study sites should be done according to a sampling plan that allows the individual site findings to be expanded to the full program with a quantifiable level of statistical precision. Similarly, it may be impractical and costly to apply M&V activities to all measures installed at a particular site. Measures within a particular site should be sampled in a manner that allows the results to be expanded to the full set of measures installed at the site with a quantifiable level of statistical precision. See Chapter 13, Sampling, for more information on sampling for M&V activities.

Upper Level Roadmap

The upper level of the M&V roadmap is shown in Figure 7.1. The appropriate M&V option is discussed in the next section.

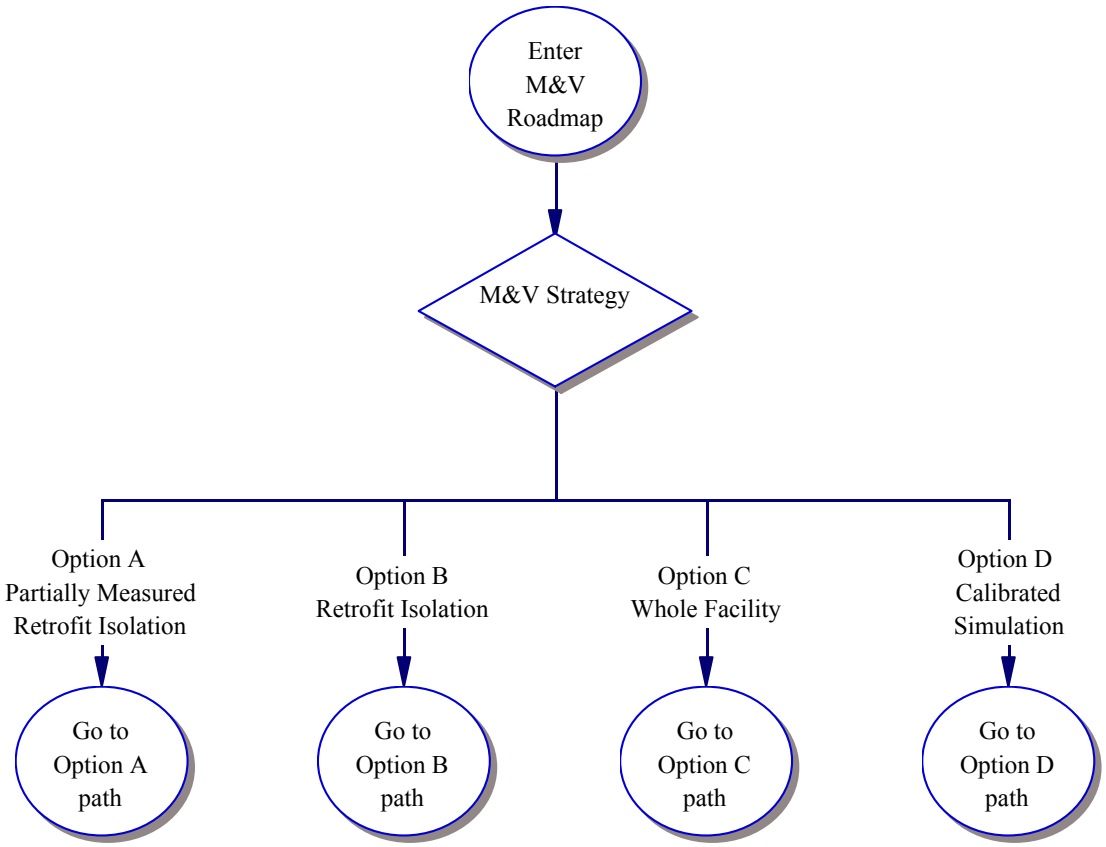

Figure 7.1: Upper Level M&V Roadmap

Measurement and Verification Options

This chapter provides several options for M&V studies. The options follow the terminology used in the IPMVP.

Option A - Partially Measured Retrofit Isolation. Savings under Option A are determined by partial field measurement of the energy use of the system(s) to which an energy conservation measure (ECM) was applied separate from the energy use of the rest of the facility. Measurements may be either short-term or continuous. Partial measurement means that some parameter(s) affecting the building's energy use may be stipulated, if the total impact of possible stipulation error(s) is not significant to the resultant savings. Careful review of ECM design and installation will ensure that stipulated values fairly represent the probable actual value. Stipulations should be shown in the M&V Plan, along with analysis of the significance of the error they may introduce. Savings are estimated from engineering calculations using short-term or continuous post-retrofit measurements and stipulations. A typical application of Option A is a lighting retrofit, where pre/post fixture watts are stipulated from a standard fixture wattage table, and operating hours are derived from short-term measurements of fixture run-time.

Option B - Retrofit Isolation. Savings are determined by field measurement of the energy use of the systems to which the ECM was applied separate from the energy use of the rest of the facility. Short-term or continuous measurements are taken throughout the post-retrofit period. Savings are estimated from engineering calculations using short-term or continuous measurements. A typical application of Option B is a variable frequency drive applied to a constant speed pumping application. Pre-retrofit power consumption is measured with a hand-held power meter (or short-term metering to confirm constant power draw), and post-retrofit power consumption is short-term metered along with some relevant independent quantity (such as fluid or ambient temperature). The relationship between power and the independent variable is used to project long-term post-retrofit energy consumption from the short-term measurements.

Option C - Whole Facility. Savings are determined by measuring energy use at the whole-facility level. Short-term or continuous measurements are taken throughout the post-retrofit period and compared to 12 to 24 months of pre-retrofit data. Savings are estimated from an analysis of whole-facility utility meter or sub-meter data, using techniques ranging from simple comparison to regression analysis. This approach is very close in concept to a billing analysis, but may contain baseline adjustment factors that are specific to each building addressed under this option.

Option D - Calibrated Simulation. Savings are determined through simulation of the energy use of components or the whole facility. Simulation routines should be demonstrated to adequately model actual energy performance measured in the facility. This option usually requires considerable skill in calibrated simulation. Savings are estimated from energy use simulation, calibrated with hourly or monthly utility billing data and/or end use metering.

Choosing the M&V Option

The M&V option employed will depend primarily on the impact analysis chosen and the overall precision goals for impact results as defined in the impact evaluation plan. In general, Option A is the least accurate and least costly option, whereas Option B or D, depending on the measure, are generally the most accurate and most costly options. All measures can be evaluated using Options A, B, and D, but the accuracy of the estimates provided under Option A decreases as the measure complexity increases. Option C is limited to projects where the expected savings exceeds the metered energy consumption by at least 10%.[142] A list of general considerations for selecting an M&V approach is shown in Table 7.1. Specific criteria for option selection are described in the section covering the details of each option.

Table 7.1: M&V Option Selection Criteria

Consideration	Option A	Option B	Option C	Option D
Ability to determine individual measure savings	Yes	Yes	No	Yes
Ability to adapt to unanticipated baseline adjustments	Yes, revise savings calculation procedure	Yes, if changes don't affect measure performance	Can be complex	Yes, most situations can be readily simulated
Measure type	All	Could be difficult for shell measures	All, as long as impact > 10% of total consumption	All
Ease of method understanding by non-technical reviewers	Easy to understand	Easy to understand, as long as baseline adjustments are straight-forward	Easy to understand, as long as baseline adjustments are straight-forward	Difficult
Special skills required	General science or engineering	Metering and monitoring expertise	Statistics or math background required.	Many - including simulation application, metering and monitoring, statistics.
Level of measure interaction	None, unless part of the engineering equation	None	Full interactions present in billing data accounted for	Full interactions present in model accounted for

[142] 10% is the minimum savings criterion established in ASHRAE Guideline 14. Depending on the variability of the data, a greater energy savings fraction may be required for a successful billing analysis.

Table 7.1: Continued

Consideration	Option A	Option B	Option C	Option D
Length of post-retrofit period	Representative period, depending on measured parameter	Representative period	At least one year	None, if model calibrated to pre-retrofit data
Applicable to new construction project	Yes	No	No	Yes
Accuracy	Low, subject to uncertainty calculation	Can be very good, subject to uncertainty calculation	Can be very good, subject to uncertainty calculation	Moderate to good, subject to modeler ability
Cost	Estimate 1% - 3% of annual measure cost savings	Estimate 3% - 15% of annual measure cost savings	Estimate 1% - 10% of annual measure cost savings	Estimate 3% - 10% of annual measure cost savings

Selection of the M&V option represents a balance between accuracy and cost. Approaches for striking this balance vary in the literature. For example, ASHRAE Guideline 14[143] takes a quantitative approach, where the risk in the uncertainty is calculated from the energy value of the difference between the savings estimates at the upper and lower ends of the confidence interval. Improvements to the M&V approach are introduced iteratively, with the incremental M&V costs compared to the reduction in savings risk. The literature on greenhouse gas trading policy encourages more comprehensive M&V by applying savings discount factors tied to the M&V option.[144] The USEPA Conservation Verification Protocols[145] direct the evaluator to report verified savings at the low end of the confidence interval, thus encouraging more precise estimates.

M&V Plan

A measurement and verification plan should be developed for each site included in the M&V study. These individual site plans should be filed with the final EM&V report at the conclusion of the project. The purpose of the plan is to identify the data needs and analysis procedures prior to collecting the field data. The overall components of the M&V plan are listed below:[146]

---

143 (* ASHRAE 2002).

144 A discussion of other approaches is found in "International Greenhouse Gas Trading Programs: A Discussion of Measurement and Accounting Issues." (Vine et al. 2003).

145 Savings are reported with 75% confidence at the low end of the confidence interval. See (USEPA 1995).

146 The M&V plan outline is taken from the 1999 ASHRAE Handbook, Applications Volume, Chapter 39 (ASHRAE 1999). The IPMVP gives less precise guidance on M&V plan content than what is listed here.

Identify Goals and Objectives

The goals and objectives of the M&V activity at a particular site are stated in the plan and may include the following:

- Verify measure installation
- Verify proper operation of measures
- Measure specific parameters required for energy savings calculations
- Meter energy consumption and demand for energy savings calculations

Specify Building Characteristics

Building characteristics are listed in the plan to help future users of the data understand the context of the monitored data. The building characteristics description should include:

- General building configuration and envelope characteristics (particularly the energy-related characteristics)
- Building occupant information (number of occupants, occupancy schedule, activities)
- Internal loads (lighting, appliances, plug and process loads)
- Type and quantity of energy-using systems and control setpoints
- Changes in building occupancy or operation during monitoring period that may affect results.

Besides the general building or facility description, a description of the energy conservation measures and their respective projected savings should also be included. A typical form used to describe a building is shown in Table 7.2. This form is shown as an example only, and may be too detailed or too simplistic for any particular M&V plan.

Table 7.2: Building Characteristics Reporting Form

General Information	
Building Description	3-Story Municipal Office Building
Building Type	According to CEC load forecasting conventions:
Building Vintage	New construction
Building Location	Davis, California
Building Owner	City of Davis
Occupancy Schedule	6 a.m. to 6 p.m. normal business days
Weather Station	Davis Airport
Architectural Features	
Total Conditioned Area	49,000 net SF [per architectural program] / 75,000 gross SF
Layout	3 floors above grade, plus basement
Floor to floor height	15 ft
Floor to ceiling height	10 ft
Overall Opaque Wall U-value	0.054
Overall Roof U-value	0.039
Window U-Value (with frame)	0.68
Window Shading Coefficient	0.60
Window to wall area ratio	23
Internal Load Densities	
Lighting Power Density – All Areas	1.57 W/sf 0.2 W/sf task lighting in office areas
Plug Load Density – All Areas	1.0 W/sf
Occupant Density – All Areas	175 people

Table 7.2: Continued

HVAC Description	
HVAC System Type	Two Packaged VAV with Gas Heating Sections and Return Fans
Fan Flow Control	Inlet vanes
Terminal Heating	Fan Powered boxes w/ Electric Coils
Cooling/Heating Setpoint	74° F Cooling / 72°F Heating
Supply Air Rate	40,000 cfm per unit
Outside Air Ventilation Rate	2,625 cfm (15 cfm per person)
Control Sequence	Night Setback/Economizer
Cooling System	9.47 EER DX Coils
Measure Description	
DEER 2003 Measure ID Number(s)	D03-002 lighting power reduction D03-025 high performance glass and daylighting controls
Measure capacity and efficiency information	D03-002 – 40% LPD reduction D03-025 – Standard glass and no daylighting controls
Operating hours	Follows building occupancy
Installation date	New construction
Baseline capacity and efficiency information	Title 24 requirements by space type

The building description data should be tailored to the scope and needs of the project. It may not be desirable to gather all of the information about the building during a phone survey confirming measure installation. However, future use of the data may depend on a reasonable understanding of the context of the equipment and related operating conditions. Descriptive data should be coordinated with the needs of related "overarching" studies, such as updates to the Database on Energy Efficiency Resources (DEER)[147] and measure potential studies, as described in Chapter 15, Overarching Evaluation Studies.

Specify Data Products and Project Output

The end products of the M&V activity should be specified. These data products should be referenced to the goals and objectives on the project, and include a specification of the data formats and engineering units. To maximize future use of M&V project results, the results should be normalized and reported according to a standard format. For example, the format followed in the DEER Update Study[148] is shown in Table 7.3.

[147] (Xenergy, ADM Associates et al. 2001).
[148] (Itron Incorporated 2004).

Table 7.3: Example Units for Measure-Specific Savings Reporting from DEER

Technology	Measure group	Unit	Comments
Lighting	Lighting	kW_{pre} connected	Measures expressed as % reduction
Lighting controls	Lighting	kW_{pre} controlled	Measures expressed as % reduction
Plug loads	Plug loads	kW_{pre} connected	Measures expressed as % reduction
Roof insulation, cool roof	Shell	SF roof	
Air curtain	Shell	Each	
Glass solar heat gain	Shell	SF glass	Show pre/post SHGC
HP glass plus daylighting controls	Shell	SF glass	
Chillers	HVAC	Ton	Show pre/post kW/ton
Chilled and hot water distribution loop temperature control	HVAC	SF building	
Chilled and hot water distribution loop flow control	HVAC	Hp pump	
HVAC system conversion	HVAC	SF building	
VAV fan control	HVAC	Hp fan	
Evap cooling of makeup air	HVAC	Ton	
Ventilation rate	HVAC	SF building	
Heat recovery	HVAC	Each	
Economizers	HVAC	Ton	
HVAC maintenance	HVAC	Ton	
Heat rejection system	HVAC	Ton	
High efficiency furnace	HVAC	KBtu/hr	Show pre/post efficiency
Large boilers	HVAC	SF building	Show pre/post efficiency
HVAC system control	HVAC	SF building	
High efficiency packaged HVAC	HVAC	Ton	Show pre/post efficiency
High efficiency motors	motors	Hp	Show pre/post efficiency
Water heater tank insulation	Water heating	SF building	
Water heater efficiency	Water heating	SF building	Show pre/post efficiency

Specify M&V Option

The M&V option chosen for the project should be specified according to the options described in this Chapter:

- Option A - Partially Measured Retrofit Isolation
- Option B - Retrofit Isolation
- Option C - Whole Facility
- Option D - Calibrated Simulation

For any particular building, a combination of options may be used. The M&V option applied to each measure should be described.

Specify Data Analysis Procedures and Algorithms

This is a key component of the M&V plan. Often, data are collected without a clear understanding of the later use for the data. This can result in either extraneous data collection and/or missing data during the data analysis step. Fully specifying the data analysis procedures will help ensure that an efficient and comprehensive M&V plan is presented.

Specify Field Monitoring Data Points

The actual field measurements made are specified here. This is analogous to an energy management system "points list," where the sensor, location, measurement and engineering units are specified. For example:

- For measuring the run-time of a boiler, the field data point description would be: *"Accumulated run-time of draft fan serving boiler number 1, using an inductive run-time logger mounted on the draft fan motor."*

- For measuring air conditioner supply air temperature, the field data point description would be: *Duct air temperature (in degrees F) using a sheathed thermistor sensor located in the supply duct three feet downstream from AC-1.*

- For measuring chilled water temperature, the field data point description would be: *"Chilled water supply temperature measured with a probe-type thermister inserted in a thermowell."*

Estimate Data Product Accuracy

All measurement systems have error, expressed in terms of the accuracy of the sensor and the recording device. The combined errors should be estimated using a propagation of error analysis, and the final data product accuracy described. For example, a thermistor sensor and recording device may have an accuracy of ± 0.5°C (± 0.9°F). If the data product delivered is heat supplied by a boiler, then the combined error of two sensors and the flow measurement device on the overall calculation should be supplied. In situations where the data product is calculated from a temperature difference, the temperature measurement error can represent a large fraction of the error in the final data product. All measurement device accuracy specifications assume the device has been recently calibrated and the sensor is placed correctly.

Specify Verification and Quality Assurance Procedures

Data analysis procedures should include checks to identify invalid data. For example, temperature differences that are negative can indicate a problem with a sensor or with sensor placement. Energy data that exceed the product of the nameplate rating and the

monitoring duration are also invalid. Any data verification and quality assurance procedures planned for the project should be described.

Specify Recording and Data Exchange Formats

Data formats should be described, so others attempting to use the data can interpret them correctly. For example, monitoring equipment may store data in a proprietary format, requiring the use of proprietary software to read and analyze the data. The use of non-proprietary data and software is recommended. Where proprietary data is collected, simple solutions often exist for making the data available to the public: e.g., deleting the name and/or address of a building. Some monitoring systems allow export to ASCII or Microsoft Excel formats. These should be specified as part of the monitoring plan.

Measure Installation Verification

All M&V options generally include measure installation verification as a component of the overall M&V plan. The objectives of measure installation verification are to confirm that: (1) the measures were actually installed, (2) the installation meets reasonable quality standards, and (3) the measures are operating correctly and have the potential to generate the predicted savings. Verification activities are generally conducted during on-site surveys of a sample of projects. Phone and mail surveys may be used for very simple measures (such as CFL replacements), but on-site inspection is preferred.

Option A

Savings are determined by partial field measurement of the energy use of the system(s) to which an ECM was applied separate from the energy use of the rest of the facility. Measurements may be either short-term or continuous. Partial measurement means that some but not all parameter(s) affecting the building's energy use may be stipulated, if the total impact of possible stipulation error(s) is not significant to the resultant savings. Careful review of ECM design and installation will ensure that stipulated values fairly represent the probable actual value. Stipulations should be shown in the M&V Plan along with analysis of the significance of the error they may introduce. Savings are estimated from engineering calculations using short-term or continuous post-retrofit measurements and stipulations. A typical application of Option A is a lighting retrofit, where pre/post fixture watts are stipulated from a standard fixture wattage table, and operating hours are derived from short-term measurements of fixture run-time.

The overall Option A M&V path is shown in Figure 7.2.

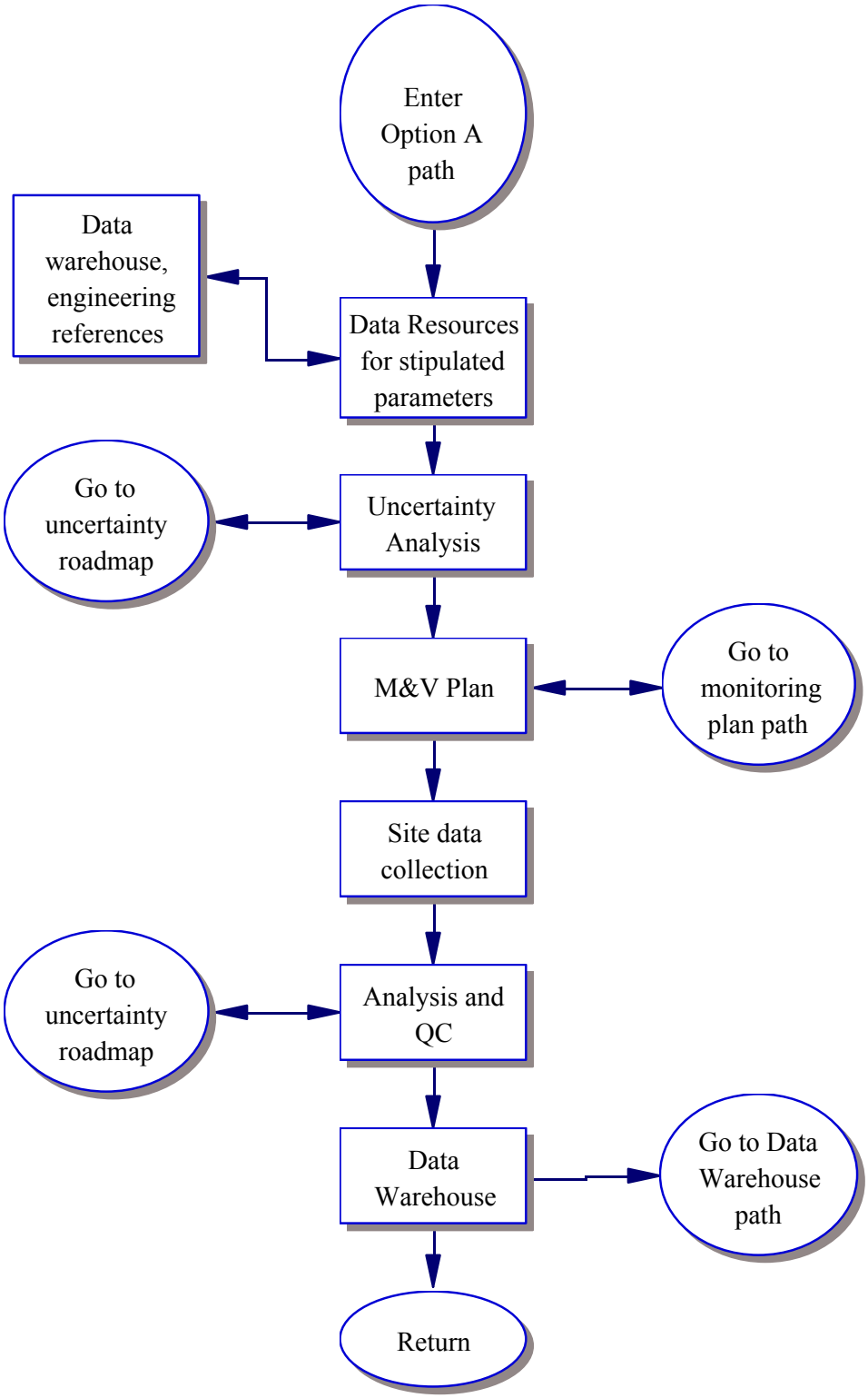

Figure 7.2: Option A Roadmap

The important issue relative to the application of Option A is that ***some*** measurement must occur and ***some*** parameter is stipulated. The number of parameters stipulated can vary significantly. For example, measurement of only one parameter with all other

parameters stipulated, or measurement of all parameters but one, with the remaining parameter stipulated are all covered under Option A. When M&V activities are conducted after measure installation, then Option A may be the only option available, since some aspect of the pre-retrofit condition may need to be stipulated.

The best applications for Option A include:

- Measures with constant loads
- Measures with small anticipated impact overall (low risk measure)
- Measures with small anticipated impact relative to the energy recorded at the billing meter
- Measures where interactive effects[149] are small and can be ignored
- Measures where baseline adjustments to whole-building data could be problematic
- Uncertainty in the deemed parameters is acceptable

All measures can be verified using Option A, but the accuracy of this option is generally inversely proportional to the complexity of the measure. Typical measures analyzed using Option A include:

- Lighting efficiency
- Lighting controls
- Constant load motors
- HVAC efficiency improvements

Some typical measured and stipulated parameters for measures are shown below.

Table 7.4: Stipulated and Measured Parameters for Lighting Measures

Engineering Parameter	Stipulated data source	Measured data source
Fixture counts	Contractor records	Pre-retrofit field audit Post-retrofit field audit
Fixture wattage	Standard fixture wattage tables	Spot measurements of fixtures or groups of fixtures
Operating hours	Standard values by building type Customer questionnaire Assume measured pre equals post or vice versa	Short-term monitoring of fixture on/off status Short or long-term monitoring of fixture circuits
Interactive factors	Ignored or stipulated based on engineering analysis	

[149] Interactive effects include the interactions between the measure and a non-measure-related end use (e.g. efficient lighting generally reduces cooling loads) and interactions between packages of measures that can cause the sum of the measure package savings to be less than the sum individual measure savings.

Table 7.5: Stipulated and Measured Parameters for Motor Measures

Engineering Parameter	Stipulated data source	Measured data source
Motor count	Contractor records	Pre retrofit field audit Post retrofit field audit
Motor wattage	Based on nameplate hp, efficiency and oversizing assumption	Spot motor kW measurement
Operating hours	Standard values by building type Customer questionnaire, Assume measured pre equals post or vice versa	Short-term monitoring of motor run hours Short or long-term monitoring of motor circuit current or kW
Interactive factors	Ignored or stipulated based on engineering analysis	

Table 7.6: Stipulated and Measured Parameters for HVAC Measures

Engineering Parameter	Stipulated data source	Measured data source
Unit count and capacity	Contractor records	Pre retrofit field audit Post retrofit field audit
Unit efficiency	Based on nameplate EER or SEER	In-situ efficiency measurement
Operating full-load hours	Standard values by building type Assume measured pre equals post or vice versa	Short or long-term monitoring of unit current or kW

Data Resources

Assumptions for stipulated parameters used in Option A come from a variety of sources, including:

- Evaluation data warehouse and/or DEER
- Lighting fixture wattage tables from the Standard Performance Contract program[150]
- Express Efficiency workpapers[151]
- Engineering references, such as the ASHRAE Handbook
- Manufacturers'data
- General evaluation literature

Data resources used to justify stipulated value assumptions should be documented as part of the M&V plan.

[150] The Standard Performance Contract (SPC) program has published a table of fixture watts for standard fixture types. (Pacific Gas & Electric 2001).

[151] Workpapers filed in support of the measure savings assumptions for the Statewide Express Efficiency commercial retrofit program are another source of engineering data and measure performance assumptions for deemed savings estimates. See www.cpuc.ca.gov.

Uncertainty Analysis

The level of uncertainty in savings estimates calculated under Option A is a function of (1) the variance in the deemed values relative to the actual measure performance parameters, (2) the magnitude of any measure or end use interactive effects, (3) the uncertainty in the measured parameters, and (4) any errors introduced when a sample of measures is used. The M&V plan should account for the potential errors in the analysis and develop a strategy to reduce the most influential errors.

Initial estimates of engineering parameter uncertainties should be used to provide an estimate of the overall uncertainty in the savings calculations. Initial estimates of parameter uncertainty values are typically based on professional judgment, due to a general lack of data on the uncertainty of stipulated engineering parameters.[152] The uncertainty estimate should address both instrument error and variations due to differences in equipment schedule and performance. The uncertainty of the savings estimate will likely be reduced by site-specific data collection. An up-front estimate of the uncertainty in the savings calculations should be used to guide the development of the M&V plan.

Option A Example

This example describes an uncertainty analysis applied to a lighting retrofit project. The annual energy savings from a commercial lighting retrofit program are estimated using an engineering analysis. Initial estimates of the uncertainty of each input parameter are shown in Table 7.7 below:

Table 7.7: Data for Lighting Retrofit Program

Parameter[153]	Value	Uncertainty	Error
Number of fixtures replaced	10,000	± 1%	± 100
Efficient fixture watts	105W	± 1%	± 1.1
Baseline fixture watts	160W	± 7%	± 11.2
Lighting operating hours	2,860 hr	± 25%	± 715
Demand diversity factor of efficient fixture	0.8	± 5%	± 0.04
Demand diversity factor of baseline fixture	0.75	± 5%	± 0.038
HVAC interaction factor	0.24	± 15%	± 0.036
Energy savings	1.28 GWh	± 40%	± 0.51

See the Appendix to this chapter for the calculations associated with this example. The uncertainty in the engineering estimate of savings prior to any M&V activity is ±40%.

[152] Developing a data resource for engineering parameter uncertainty could be an important role for the Data Warehouse, as described later in this chapter.

[153] See the Appendix to this Chapter for a definition of these parameters.

The last column in the previous table is an indication of the relative influence of each parameter on the overall uncertainty of the estimate. The priorities for data gathering for estimating annual energy savings are as follows.

Table 7.8: Data Gathering Priority Ranking for Gross Annual Energy Savings Estimates

Rank	Parameter	Parameter Uncertainty	Contribution to Overall Uncertainty[154]
1.	Operating hours	± 25%	39%
2.	Baseline fixture watts	± 7%	34%
3.	Demand diversity factor of baseline fixture	± 5%	17%
4.	Demand diversity factor of efficient fixture	± 5%	9%
5.	Efficient fixture watts	± 1%	< 1%
6.	Number of fixtures replaced	± 1%	< 1%
7.	HVAC interaction factor	± 15%	< 1%

To reduce the uncertainty in the overall estimate, data gathering activities should focus on reducing uncertainty in the parameters that have the greatest contribution to the overall uncertainty. The priorities for data gathering for reducing uncertainty are monitoring fixture full-load hours, connected load, and demand diversity factors.

Option B

Savings are determined by field measurement of the energy use of the systems to which the ECM was applied; this is separate from the energy use of the rest of the facility. Short-term or continuous measurements are taken throughout the post-retrofit period. Savings are estimated from engineering calculations using parameters derived from the short-term or continuous measurements. A typical application of Option B is a variable frequency drive applied to a constant speed pumping application. Pre-retrofit power consumption is measured with a hand-held power meter (or short-term metering to confirm constant power draw), and post-retrofit power consumption is short-term metered along with some relevant independent quantity (such as fluid or ambient temperature). The relationship between power and the independent variable is used to project long-term post-retrofit energy consumption from the short-term measurements.

The overall path through Option B is shown in Figure 7.3.

[154] See the Appendix to this Chapter for a calculation of the contribution of each parameter to the overall uncertainty.

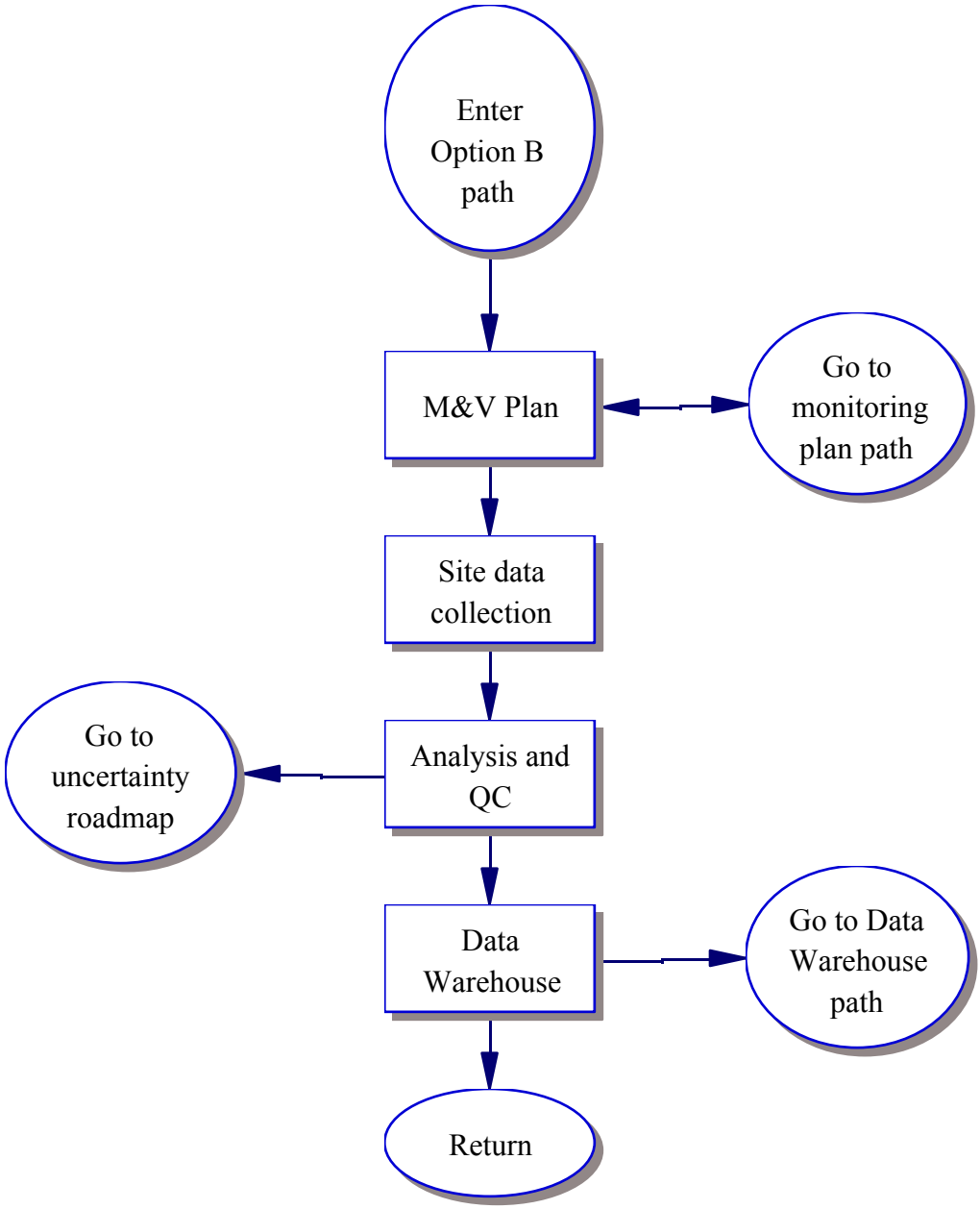

Figure 7.3: Option B Roadmap

Option B is very similar to Option A, but no stipulations are allowed. Full measurement is required. This option is analogous to the retrofit isolation approach described in ASHRAE Guideline 14, where full measurement is also required. The Option B measurements allow true electric power, or proxy measurements such as accumulated run-time. Although both of these strategies are allowed within Option B, both may not yield the same results in terms of accuracy.

Short- or long-term monitoring may be employed under Option B. Short-term or spot measurements are appropriate for measures with constant or predictable operating hours. Continuous metering may be installed at sites financed though a performance contract.

Long-term metering provides greater certainty in reported savings, but is more expensive. Secondary use of long-term or continuous measurements may also be used to enhance or maintain measure performance. The benefits of enhanced measure operation may offset the costs of additional metering.

All measures may be verified using Option B, but the degree of difficulty and cost goes up with measure complexity. Savings verification under Option B is likely to be more accurate than Option A, but more costly.

The best applications for Option B include:

- Measures with small anticipated impact relative to the energy recorded at the billing meter
- Measures where interactive effects are small and can be ignored
- Measures where baseline adjustments to whole-building data could be problematic
- Buildings where sub meters already exist to isolate the energy use of affected systems
- Situations where metering added under Option B would have additional benefit to the building operators, offsetting the cost
- Projects where measure level impact information is desired

Option B Example:

This example describes an application of Option B to a comprehensive HVAC system retrofit. A chiller plant improvement project at a municipal office building was subject to short-term pre/post monitoring. The retrofit consisted of chiller replacement, cooling tower replacement, conversion of constant volume air handlers to variable air volume (VAV) operation, and conversion of constant volume pumping to variable volume operation. The monitored data were used to build a simple regression model, where daily cooling energy consumption is predicted as a function of average daily ambient temperature, as shown in Figure 7.4. The annual savings were estimated by applying the pre/post regression model to long-term average temperature data.

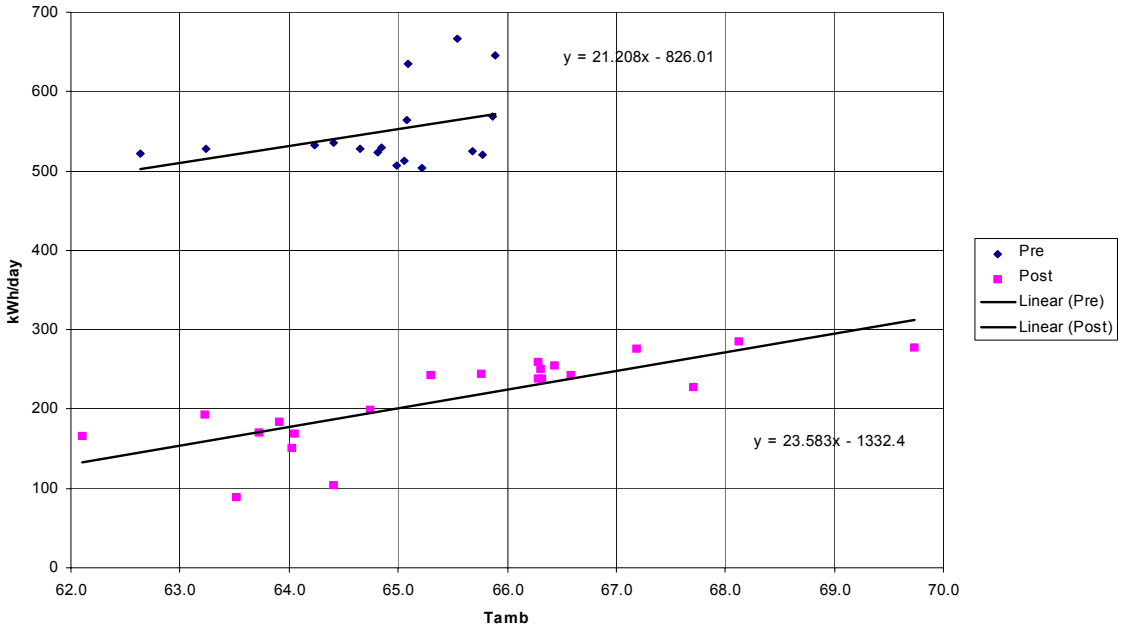

Figure 7.4: Example of Short-Term Monitoring of HVAC System Upgrade Using Option B

Option C

Under Option C, savings are determined by measuring energy use at the whole-facility level. Short-term or continuous measurements are taken throughout the post-retrofit period and compared to 12 to 24 months of pre-retrofit data. Savings are estimated from analysis of whole-facility utility meter or sub-meter data using techniques ranging from simple comparison of utility bills to regression analysis. This approach is very close in concept to the statistical billing analysis technique described in Chapter 6, and many of the issues addressed in Chapter 6 also apply to Option C. A billing analysis conducted under Option C may contain baseline adjustment factors that are specific to a particular building addressed under this option, rather than factors applied to a population of buildings.

The overall roadmap through Option C is shown in Figure 7.5.

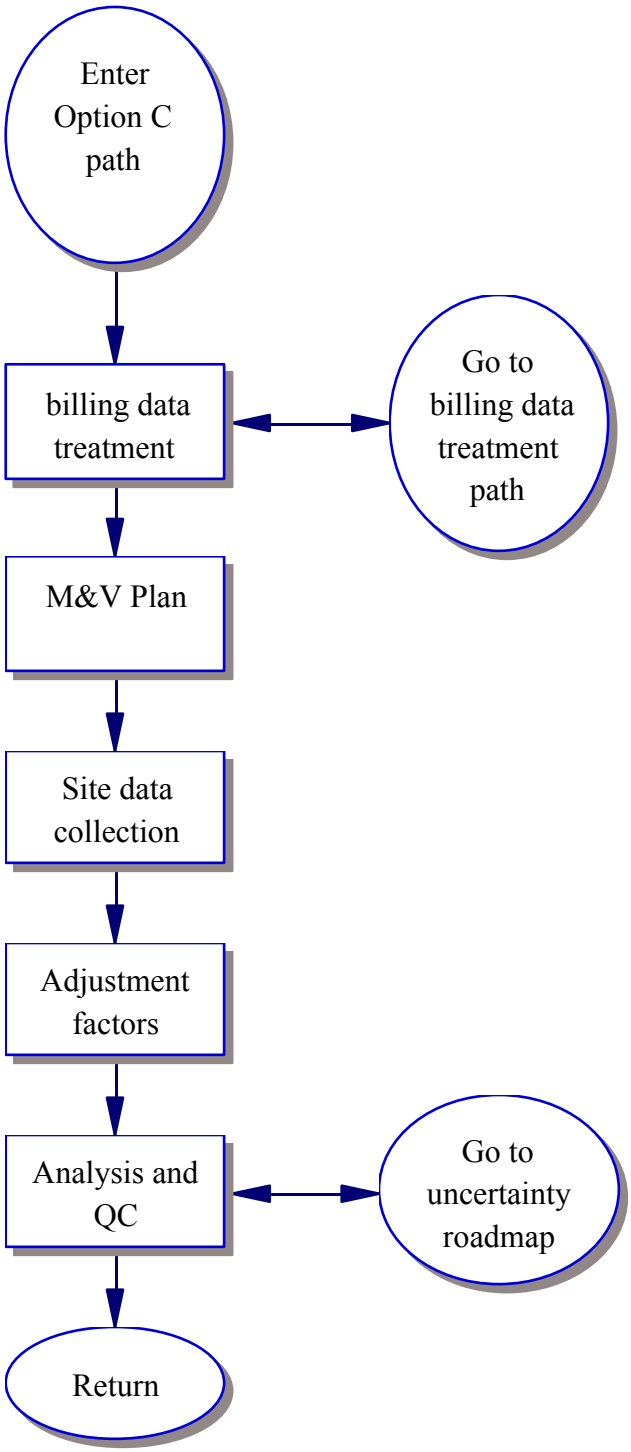

Figure 7.5: Option C Roadmap

Billing Data Collection

Issues with Option C are similar to those encountered when conducting a statistical billing analysis:

- Data may need to be normalized to account for meter read dates
- Missing data or estimated billing may confound the analysis
- Interval demand data from demand-recording meters may be available from the serving utility, but special permission and billing data release permission from the customer or a consultant non-disclosure agreement with the customer's utility will likely be required
- Account information and billing addresses may not match the site studied. During onsite verification activities, recording the meter numbers of all meters affected by the project will help identify the correct billing record

Comparison Models

Simple comparison models look at the monthly billing data (corrected for meter read dates) during the pre- and post-retrofit period, and derive savings as a simple subtraction of the pre- and post-period data. These comparisons are appropriate only for non-weather-dependent measures where the hours of operation and other factors influencing energy consumption besides the energy efficiency measure remain constant during the pre- and post-retrofit period.

Adjustment factors

Adjustment factors are generally added to the models to account for differences in the pre- and post-retrofit periods that can affect energy consumption outside of the impact of the installed measures. Adjustments are usually made based on weather, hours of occupancy, and building operating mode (e.g. heating or cooling seasonal operating mode). Adjustments for additions of new process loads (such as the addition of a new computer center to an office building), changes in process output (such as widgets produced or hamburgers sold), and occupied floor space may also be required. Weather adjustments may be based on heating and cooling degree days, humidity, and/or temperature.

Analysis Techniques

Simple regression analysis can be used with daily or monthly consumption data. Models developed to predict energy savings from interval data should use daily rather than hourly data to control the number of independent variables and reduce uncertainty. The interval of the weather data used to make the adjustments must be compatible with the billing data; e.g. daily weather data may be needed to calculate weather adjustments to monthly data to correspond to the billing read dates. Monthly models should use pre-retrofit data in full year increments to avoid bias (by capturing all seasonal effects): e.g., 12 or 24

months prior to retrofit. Regression models used for weather correction are usually one of the following types:[155]

Two-parameter

Two-parameter models are used when the energy consumption is linear with outdoor temperature, as shown in equation 1 and Figure 7.6.

$$E = C + B(T)_+ \qquad (1)$$

where:

E	=	predicted baseline energy consumption per period
T	=	outdoor temperature
C	=	regression intercept
B	=	regression model slope

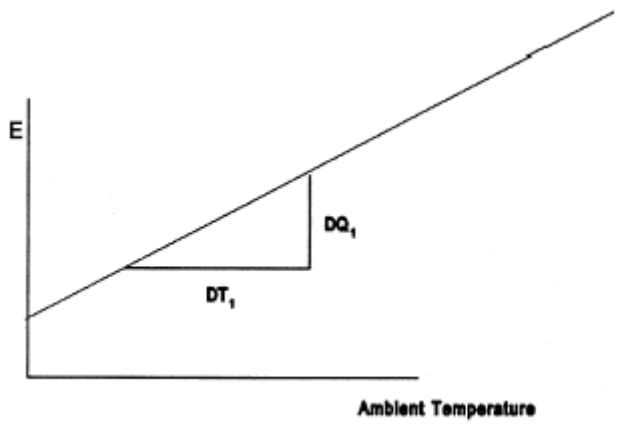

Figure 7.6: Two Parameter Model

Change point

Change point models are used when the energy consumption data show a constant plus temperature dependent behavior. Electricity consumption in air conditioned buildings with gas heating typically show this behavior, where electricity consumption during the heating season is fairly flat (e.g. not temperature dependent), and shows a temperature dependence in the cooling season. In this situation, a three-parameter change point model is used, as shown in Equation 2 and Figure 7.8.

where: $$E = C_C + B_C(T - T_{CCP})_+ \qquad (2)$$

[155] ASHRAE Guideline 14, Annex D provides detailed information on baseline model development. FORTRAN code to implement several of these models is available from ASHRAE as the Inverse Modeling Toolkit. See "Inverse Modeling Toolkit: Numerical Algorithms," (Kissock et al. 2003), and "Inverse Modeling Toolkit: Application and Testing," (Haberl et al. 2003).

E = predicted baseline energy consumption per period
C_C = constant (non-weather dependant) component of electricity consumption
T = outdoor temperature
T_{CCP} = cooling change point temperature
B_C = slope of the temperature dependent region

Note: the subscript + indicates that only positive values of the quantity in parentheses are used, otherwise, the quantity is set to zero.

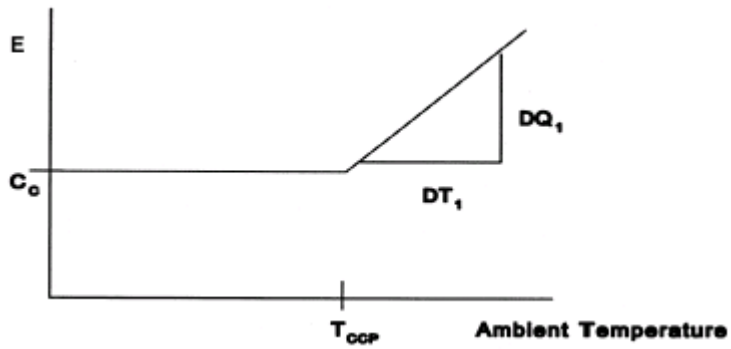

Figure 7.7: Three Parameter Change Point Model

In air conditioned buildings with electric heat, electricity consumption generally varies with temperature in the heating and cooling seasons, and is constant during the "swing" season (periods during the spring and fall when no heating or cooling is needed). In this case, a five-parameter change point model is used, as shown in Equation 3 and Figure 7.8:

$$E = C_{HC} + B_H (T_{HCP} - T)_+ \quad \text{if } T < T_{HCP}$$

$$E = C_{HC} \quad \text{if } T_{HCP} < T < T_{CCP}$$

$$E = C_{HC} + B_C (T - T_{CCP})_+ \quad \text{if } T > T_{CCP} \qquad (3)$$

where:

E = predicted baseline energy consumption per period
C_{HC} = constant (non-weather dependant) component of electricity consumption
T = outdoor temperature
T_{HCP} = heating change point temperature
B_H = slope of the temperature dependent region during the heating season
T_{CCP} = cooling change point temperature
B_C = slope of the temperature dependent region during the cooling season

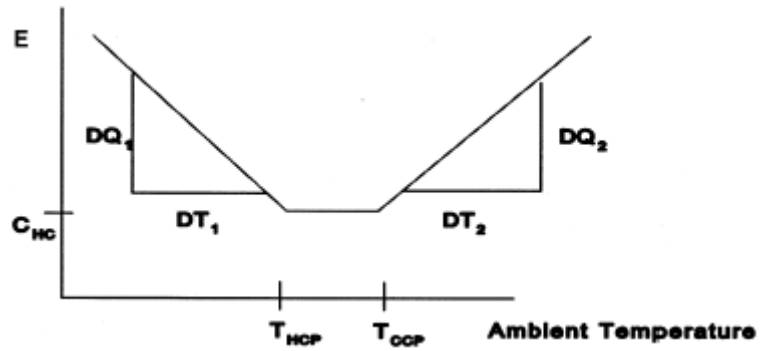

Figure 7.8: Five Parameter Change Point Model

Multi-parameter

Multi-parameter models are linear regression models that consider weather and other baseline adjustments (e.g., operating hours, production output etc.). Data on the baseline adjustments included in the model should be collected to determine the magnitude of the adjustments. If the data are not available, the magnitude of the adjustment factor can be stipulated.

Hourly or sub-hourly models

Prediction of baseline hourly or sub-hourly energy consumption may be of interest for peak demand savings estimation. Interval demand meters have been installed in many commercial buildings in California, and these meters represent a potentially rich source of whole-building energy consumption and demand data. Models suitable for predicting hourly or sub-hourly baseline energy consumption include regression[156] and artificial neural networks (ANN).[157] Evaluation of several types of hourly energy prediction models was investigated by ASHRAE under the "Great Energy Predictor Shootout."[158]

Models used to predict hourly whole-building demand may contain terms to account for day of the week, type of day (workday or weekend/holiday), hour of the day, hourly temperature and/or outdoor humidity. Several techniques for predicting baseline hourly demand in order to evaluate the effectiveness of direct load control (DLC) programs[159] were investigated by the California Energy Commission.[160] The report examined and tested the ability of several hourly regression models to predict hourly energy consumption in the absence of a DLC action. Hourly models tested included weather correction terms based on outdoor temperature, daily and hourly cooling degree days, and a temperature-humidity index.

---

156 A linear regression model scored 2nd place in the Great Energy Predictor Shootout II. See "Modeling Energy Use in Large Commercial Buildings." (Katipamula 1996).

157 Examples of ANN models include (Kreider et al. 1991; Anstett et al. 1992).

158 See the "Great Energy Predictor Shootout." (Kreider et al. 1994)

159 Although the efficiency programs currently funded through the CPUC specifically exclude DLC programs, the protocols described in this document may be useful if such programs are included in the future.

160 See "*Protocol Development for Demand Response Calculation – Findings and Recommendations*," (Kema-Xenergy 2003).

Model Accuracy Criteria

Regression models developed under Option C are generally evaluated for accuracy according to the three statistical parameters: 1) net determination bias, 2) the root mean squared (RMS) error and 3) the coefficient of variation of the RMS error (CV(RMSE)). These terms are defined below. See also Chapter 12 – Uncertainty for more information.

$$RMSE = \sqrt{\frac{\sum_{period}(M-S)^2}{N}} \tag{4}$$

$$MBE(\%) = \frac{\sum_{period}(M-S)}{\sum_{period} M} \times 100 \tag{5}$$

$$A_{period} = \frac{\sum_{period} M}{N} \tag{6}$$

$$CV(RMSE) = \frac{RMSE_{period}}{A_{period}} \times 100 \tag{7}$$

where:

M	=	the measured usage from billing data
S	=	the predicted usage from the baseline model
N	=	the number of periods in the baseline billing data
A	=	the average usage over the period

Baseline Model Selection Criteria

Criteria for developing and selecting baseline energy consumption models under Option C are:

- Extrapolation range - Apply data that are within 90% of the minimum value and 110% of the maximum values used to develop the baseline model.

- Expected savings should exceed 10% of the whole-building energy consumption

- Baseline period should span at least 12 months, and contain at least 9 data points. Data should be included in full-year increments (e.g. 12 months, 24 months or 36 months) to reduce weather-induced bias.

- The IPMVP specifications for Option C models require the Coefficient of Variation of the Root Mean Squared Error (CVRMSE) to be less than or equal to 20% on energy and 30% on demand.

Best applications of Option C include:

- Projects where the expect impacts are 10% or more of the whole-building consumption
- Projects where whole-building rather than measure-specific results are permissible
- Projects where the measures do not lend themselves to retrofit isolation – such as shell measures
- Projects where interactive effects need to be included.

Option D

Savings under Option D are calculated by simulating the energy use of components or the whole facility. Simulation routines should be demonstrated to adequately model actual energy performance measured in the facility. This option usually requires considerable skill in calibrated simulation. Savings are estimated from energy use simulation, calibrated with hourly or monthly utility billing data, and/or end use metering. The overall roadmap through Option D is shown in Figure 7.9.

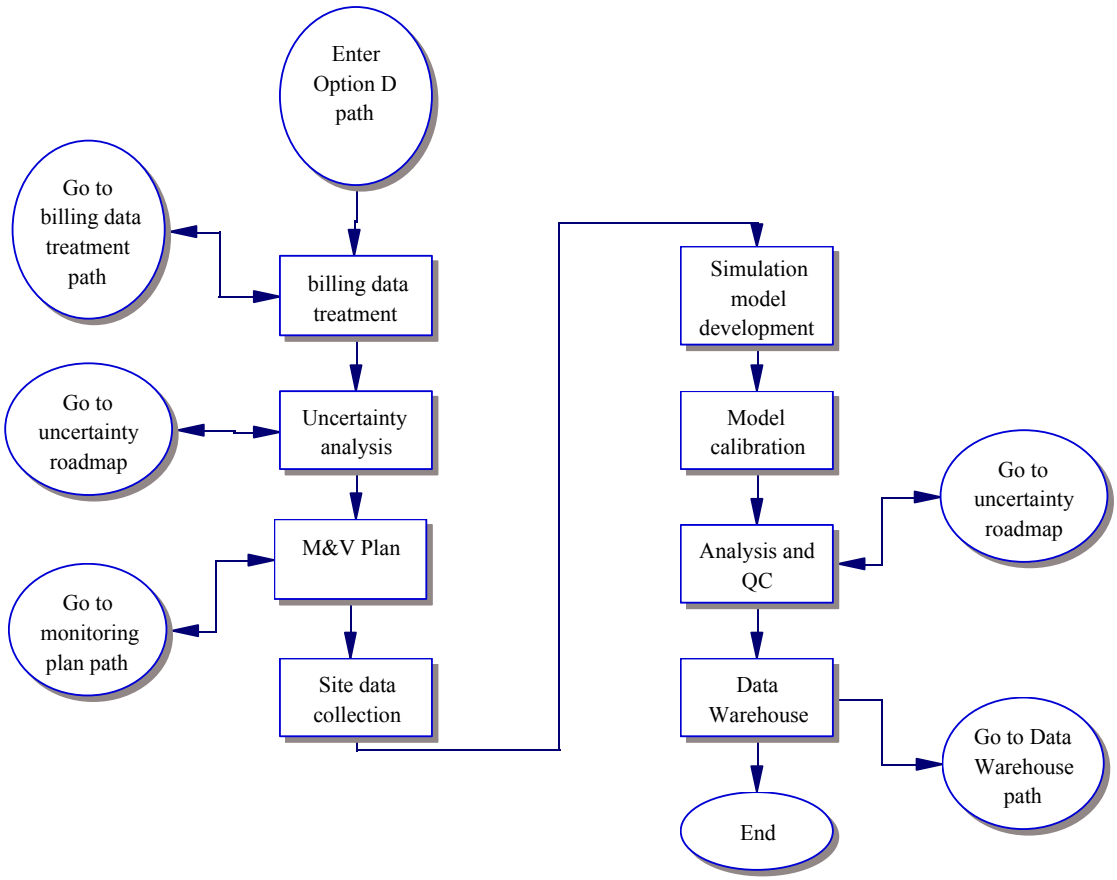

Figure 7.9: Option D Roadmap

Model calibration is done by reducing uncertainty in input variables through measurement and comparing model response to measured data. Certain model inputs are derived from measured data, while other model inputs are adjusted until reasonable agreement to measured data is reached.

There are dozens of model inputs that exert a major influence on the results. Any number of inputs can be adjusted to achieve the same final answer. The problem with model calibration to monthly billing data is that agreement may be reached ***for the wrong reason*** by randomly adjusting inputs. Accuracy of savings estimates calculated from an improperly calibrated model may be worse than those calculated from an uncalibrated model. The modeler should take advantage of all information available to eliminate as much uncertainty in the model inputs as possible, but should not adjust inputs randomly to produce a predetermined result. Data sources for model calibration are described below.

Billing Data

Historical utility billing data from a one or two year period can be used to check model results for gross errors. Billing data can give some insight into building energy use through a process of elimination, as described below.

- Estimate combined lighting, equipment and fan energy use from billing data for months not requiring mechanical cooling, primarily in systems equipped with economizers and non-electric heating.
- Estimate annual cooling energy from the difference between the non-cooling and cooling month energy consumption.
- Estimate annual heating consumption when heating is the only significant non-electric fuel end use.
- Estimate hot water consumption from billing data during the non-heating season when heating and hot water are served by a non-electric energy source.

Whole Building Hourly Demand Metering
Load research meters installed at larger commercial facilities can provide important calibration data for building energy simulation models. The whole-building load data can also be used to estimate building operating schedules.

One-Time Tests or Spot Metering
One-time tests can be useful to quantify performance of critical building parameters. These tests can be used in conjunction with short-term monitoring to improve model inputs and calibrate model response. They offer a "snapshot" of building performance, but may be useful to infer performance of the building on a long-term basis. One-time tests are best suited to measuring building parameters that do not change with time or change very slowly. Examples of one-time tests include:

Infiltration (Blower Door) Tests. The M&V plan may require a measurement of air leakage rates before and after weatherization actions are taken. Typically, a blower door is used to determine where leaks should be sealed. The blower door is attached to the house, and the leakage rate before any air-sealing actions are taken is established. Buildings already "tight enough" are not subject to any further leak sealing. Leaky buildings undergo sealing until some minimum leakage rate is achieved. Leakage rates before and after leakage sealing are recorded.

Glazing Transmissivity Tests. It may be difficult to determine the solar gain properties of windows in the absence of building plans and/or specifications. A simple test with a portable pyranometer will provide information on solar transmission, which can be used to estimate the solar heat gain coefficient of the glazing system.

Duct Leakage Tests. Duct leakage in residential and small commercial buildings can have a significant influence on the overall efficiency of the HVAC system. Duct leakage test methods are described in ASHRAE Standard 152-P.[161]

Refrigerant Charge and Efficiency Tests. The state-of-charge of residential or small commercial air conditioning systems can also affect the efficiency of the unit. Test protocols for evaluating adequacy of refrigerant charge are described in Wilcox et al.[162]

---

[161] ASHRAE Standard 152-P. (ASHRAE 2001).
[162] (Wilcox et al. 2001).

Data collected during refrigerant charge testing and adjustment can be used to estimate the in-situ efficiency.[163]

Furnace Efficiency Tests. Proctor[164] presents a method for determining in-situ furnace efficiency. This test method may prove useful for measuring the effects of furnace replacement programs. The method can be used to establish furnace efficiency for calibrating simulation models.

Fan Energy Tests. One-time tests of fan energy provide valuable data for model calibration. These tests are particularly effective on constant volume fans. Variable volume fan tests will require a manual override of the fan controls to obtain test data over the full range of operation.

Outside Air Tests. Outside air drawn into an HVAC unit to provide ventilation can have a big effect on total HVAC loads and peak demand. A one-time test of the quantity of outside air may be necessary to reliably establish this model input. Pitot tube traverses in large systems or flow grid measurements[165] in smaller systems can be used.

Figure 7.10: Airflow Measurements Using a Flow Grid

A flow grid is used to measure as-installed airflow rate. A series of flow grids are installed in place of the filters; the airflow rate through each flow grid is displayed on a digital manometer. Flow grids can also be installed in outdoor air intakes to measure outdoor airflow rate.

Short-Term Monitoring

Short-term monitoring of critical performance and behavioral parameters are useful for model calibration. Examples of short-term tests include:

---

163 See the website for the ACRx Service Assistant, www.acrx.com for more information. (Field Diagnostic Services 2003).

164 "The Development of a Field Furnace Efficiency Test: A More Accurate Prediction of Seasonal Efficiency." (Proctor 1991).

165 See "Field Evaluation of a New Device to Measure Air Handler Flow." (Francisco et al. 2003).

- Status or run-time monitoring of lighting circuits
- Status or run-time monitoring of fans
- HVAC diagnostic tests, including economizer operation, HVAC controls
- Short-term end use monitoring similar to tests conducted under Option B

An example of a short-term diagnostic test of an economizer is shown in the Figure below.

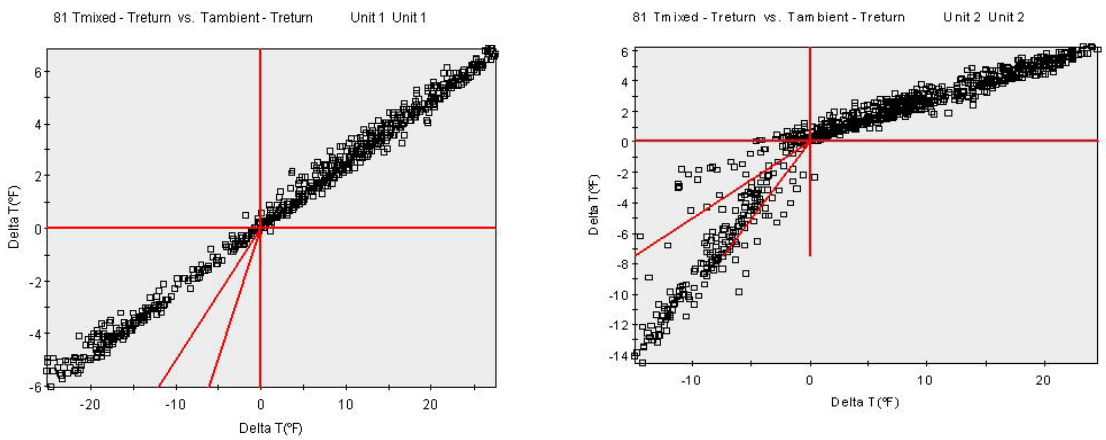

Figure 7.11: Economizer Diagnostic Plots

Short-term monitored data is used to help diagnose system problems. In Figure 7.12, the difference between the cooling coil entering (i.e. mixed) air temperature and the return air temperature (Tmixed - Treturn) on the vertical (Y) axis is plotted against the difference between the outdoor (ambient) temperature and the return air temperature on the horizontal (X) axis. The slope of the line is equal to the outdoor air fraction. Units with fixed outdoor air (no economizer) have a straight-line relationship between these data, as shown in the plot on the left. Units with economizers show a characteristic change in the slope of the line to the left of the vertical (Y) axis, as shown in the plot on the right. The slope in this region is equal to one, indicating a functioning dry bulb economizer allowing 100% outdoor air.

Calibration Accuracy

Model calibration accuracy is generally assessed by calculating the root mean square (RMS) error and the mean bias error (MBE) between the model output and the calibration data source. These quantities are calculated as follows:

$$RMSE_{period} = \sqrt{\frac{\sum_{period} (M - S)_{hr}^{2}}{N_{hr}}} \qquad (8)$$

$$MBE(\%) = \frac{\sum_{period} (M - S)_{hr}}{\sum_{period} M_{hr}} \times 100 \qquad (9)$$

where:

Mhr = the measured hourly subsystem average usage
Shr = the hourly average predicted usage from the building simulation
Nhr = the number of hours in the monitoring period

See Chapter 12 on Uncertainty for more information on these parameters.

According to ASHRAE Guideline 14,[166] model calibration to monthly whole-building utility data should continue until the RMS error is less than 15% and the net mean bias error is less than 5%. Calibration to hourly data should continue until the RMS error is less than 30% and the net mean bias error is less than 10%.

Best applications of Option D include:

- Projects where the expected impacts are greater than the expected modeling error
- Projects where measure-specific results are desired
- Projects where the measures do not lend themselves to retrofit isolation – such as shell measures
- Projects where interactive effects need to be included
- New construction projects, where the baseline must be simulated rather than measured
- Complicated HVAC control measures
- Commissioning and O&M programs

M&V Accuracy Requirements

M&V plans following Options A-D should comply with a minimum set of requirements. An example of the requirements listed in ASHRAE Guideline 14 is shown in the Table 7.9.

[166] (* ASHRAE 2002).

Table 7.9: M&V Accuracy Requirements (Adapted from ASHRAE Guideline 14)

Requirement	Option A	Option B	Option C	Option D
Net determination bias		< .005%	< .005%	
Maximum level of uncertainty	±50% at 68% confidence	±50% at 68% confidence	±50% at 68% confidence	±50% at 68% confidence
Baseline model uncertainty			CV(RMSE) < 20% on energy, 30% on demand	
Calibration accuracy				***Monthly data:*** RMS error ± 15% MBE ± 5% ***Hourly data:*** RMS error ± 30% MBE ± 10%

Metering and Monitoring

This section on metering and monitoring provides a brief overview of the types of instrumentation used to conduct M&V studies and their application to specific measurement problems. The overall metering and monitoring path is shown in Figure 7.12.

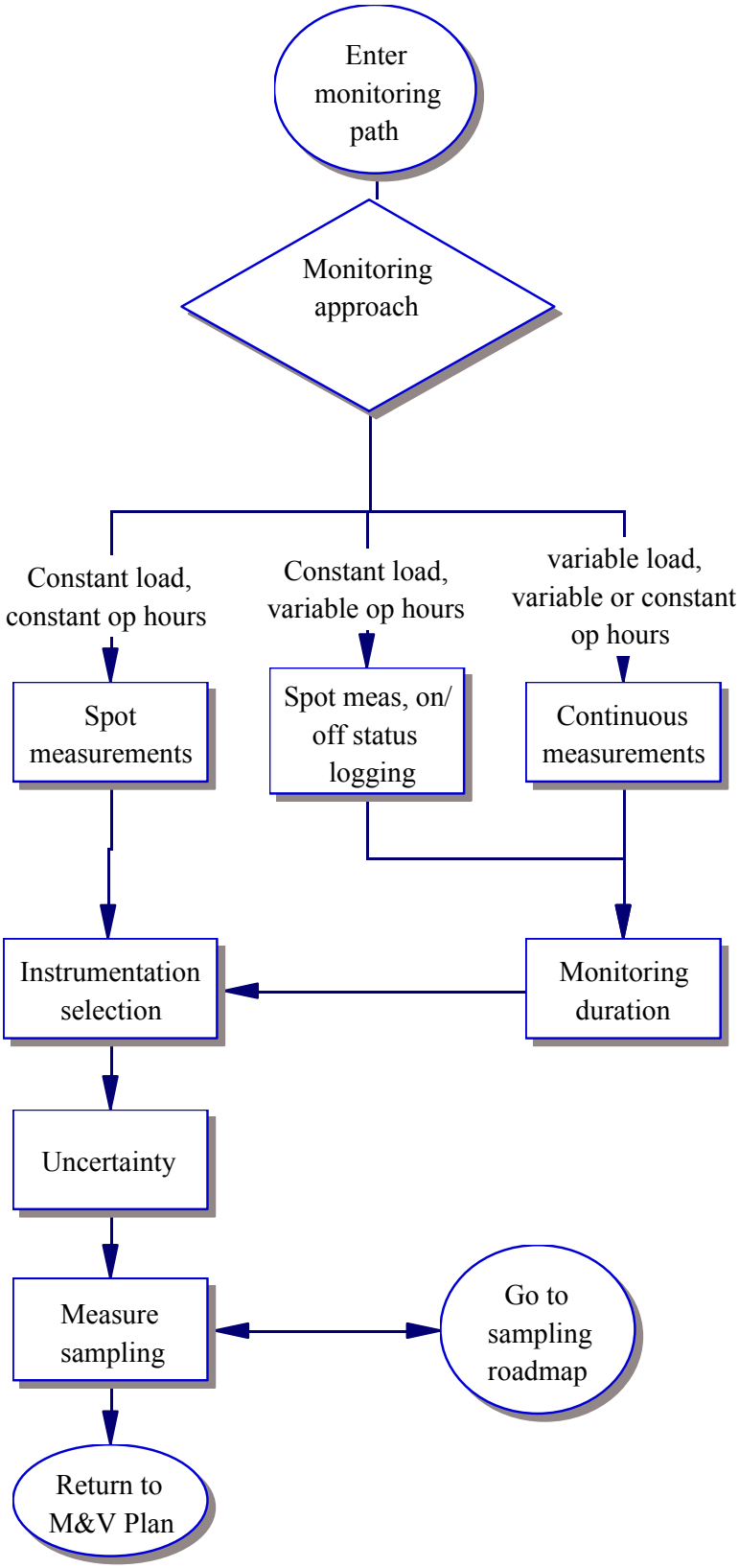

Figure 7.12: Metering and Monitoring Path

Energy-efficiency measures can be classified according to their operating characteristics. The characterization helps to determine an appropriate metering and monitoring strategy.

- ***Constant Load, Constant Operating Hours***. Constant load, constant operating hour equipment can be most economically measured with a combination of a one-time measurement of input power (spot watt) combined with a stipulated value for operating hours. Examples of constant load, constant operating hour measures include lighting, and motors driving constant volume fans.

- ***Constant Load, Variable Operating Hours***. Constant load, variable operating hour equipment can be most economically measured with a combination of a one-time measurement of input power (spot watt) combined with on/off status or accumulated run-time recorders. Examples of constant load, variable operating hour measures include lighting on/off controls, and motors driving constant volume fans controlled on demand.

- ***Variable Load, Variable Operating Hours***. Measurement of variable load, variable operating hour equipment generally requires time-series measurements of electric power or a proxy for electric power. One-time tests of electric power and current, combined with time-series measurements of current, can serve as a reasonable proxy for true electric power measurements, as long as the relationship between true electric power and current is fairly constant over the operating range of the equipment. Examples of variable load, variable operating hour equipment include daylight dimming controls, air conditioners and heat pumps, variable speed drives on pumps and fans, etc.

A summary of energy efficiency measure operating characteristics is shown in Table 7.10.

Table 7.10: Measure Classification by Performance Characteristics

End-Use	Measure Type	Load	Operating Hours	Examples
Lighting	Efficiency	Constant	Constant, Variable	Lamps, ballasts, fixtures
	Control	Constant	Variable	Occupancy sensors, sweeps, time clocks
	Control	Variable	Variable	Daylighting dimming controls
Motors	Efficiency	Constant	Constant, Variable	High-efficiency motors driving constant loads
	Efficiency	Variable	Variable	High-efficiency motors driving variable loads
	Control	Constant	Variable	Time clocks applied to constant load motors
	Control	Variable	Variable	Variable speed drives, time clocks applied to variable load motors
Water Heating	Efficiency	Variable	Variable	Heat pump water heaters
	Control	Constant	Variable	Time clocks, DLC applied to resistance water heaters
	Control	Variable	Variable	Time clocks, DLC applied to heat pump water heaters
	Load Reduction	Constant	Variable	High-efficiency (improved insulation), additional tank wrap, pipe insulation, low-flow fixtures, solar pre-heat applied to resistance water heater
	Load Reduction	Variable	Variable	Tank wrap, pipe insulation, low-flow fixtures applied to heat pump water heater
HVAC	Efficiency	Variable	Variable	High-efficiency compressors, packaged equipment, heat pumps
	Control	Variable	Variable	Economizers, time clocks, reset controls, optimal start/stop
	Load Reduction	Variable	Variable	Insulation, high-performance glass, air leakage sealing applied to buildings with heat pumps and/or air conditioners
	Load Management	Variable	Variable	Thermal energy storage applied to buildings with heat pumps and/or air conditioners
	Control	Constant	Variable	Time clocks, reset controls, optimal start/stop controls applied to buildings with resistance heating only
	Load Reduction	Constant	Variable	Insulation, high-performance glass, air leakage sealing applied to buildings with resistance heating only
	Load Management	Constant	Variable	Thermal energy storage applied to buildings with resistance heating only

Examples

A few examples of measurement strategies are described below. The FEMP M&V Guidelines and ASHRAE Guideline 14 provide extensive examples of monitoring strategies for many energy efficiency measures.

Motor Runtime Logging

The objective of the M&V project is to verify motor loading assumptions and customer-reported run hours as part of an evaluation of a high-efficiency motor program. A sample

of motors in constant-speed, constant load applications is selected for monitoring. Motor running load is measured using a portable watt meter. Magnetically-triggered run-time loggers are placed on the motor for a period of four weeks. Annual run hours are extrapolated from the short-term monitoring, and compared to customer-reported run hours.

Short-Term Lighting Monitoring

Short-term monitoring of building lighting systems is used to verify engineering assumptions of fixture power, demand diversity and full-load hours for a commercial lighting program evaluation. A series of portable, battery-powered data loggers are used to make time-series measurements of lighting circuit current over a two-week period. Circuits are selected at random for monitoring. Spot measurements of circuit kW and current are also made to measure in-situ fixture power and develop an empirical kW per amp conversion factor. This factor is applied to the time-series current data to calculate time-series kW values for each monitored lighting circuit.

Short-Term Monitoring of HVAC Systems

Short-term monitoring of a commercial building HVAC system is used to develop and calibrate a DOE-2 model. A series of portable, battery-powered data loggers are deployed over a four-week period to measure HVAC system parameters, such as temperature, humidity, and HVAC equipment current. Time-series current data are combined with spot kW measurements to calculate time-series kW data. The measured data are used to develop inputs to the DOE-2 simulation model and to calibrate the model response over the monitoring period. The calibrated simulation model is used to estimate the impacts of energy efficiency measures.

Industrial VFD Monitoring

The objective of the project is to develop an empirical model to calculate the energy savings resulting from the application of variable frequency drive (VFD) controls to plastic injection molding machines. A sample of machines is selected, representing a range of machine sizes and operating conditions present in the factory. Input power to each machine is monitored with a portable power monitor for 48 hours before and after installation of the VFD. Since the cycle times of the machines are short (less than one minute), monitoring over numerous cycles is possible during this period. An empirical model is developed to express energy savings in terms of percent idle time. The model is used to estimate energy savings for all VFDs installed at the site.

Monitoring Duration and Schedule

Monitoring duration depends on the measure use variability, climate dependence of the measure, weather variability during the monitoring period, impact estimation precision goals, and budget. Short-term monitoring activities conducted on measures with seasonal variations in use can provide biased data. Monitoring cooling equipment during the cooling season or heating equipment during the heating season are obvious examples. However, short-term monitoring during the hottest part of the cooling season may not

provide enough variation in outdoor temperature to draw meaningful conclusions about annual cooling energy performance.

An example of an experiment conducted on long-term metered data for a lighting system in a commercial building is shown in Figure 7.13.[167] The annual data record was divided into a series of two-week periods, and the results of the two-week data record were extrapolated to annual energy consumption. The deviation from the extrapolated and actual annual consumption for each two-week period is shown in Figure 7.13. In this particular example, errors resulting from data extrapolation from a two-week monitoring period ranged from –22% to 13%. The worst period for two-week short-term monitoring occurred during the December holidays.

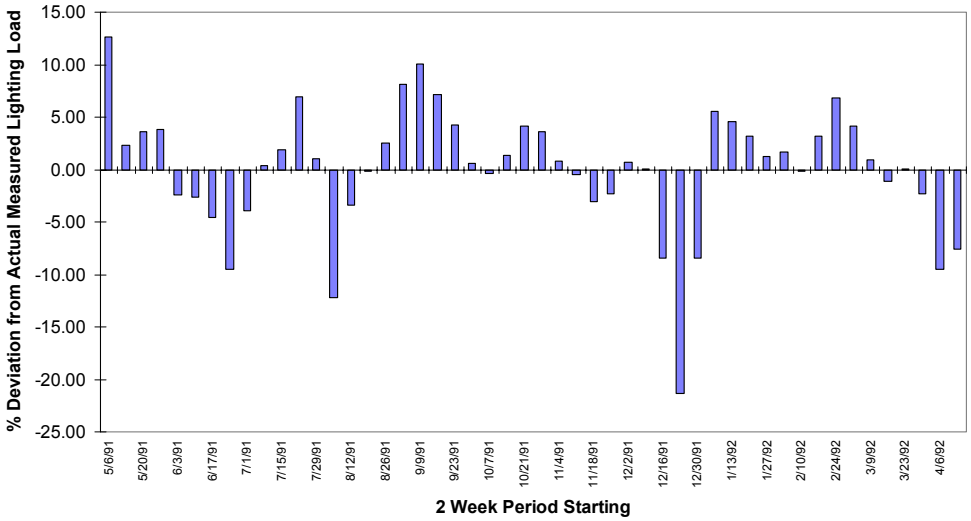

Figure 7.13: Errors Introduced by Short-Term Monitoring

The root mean squared error in the data extrapolation for lighting loads measured in several buildings as function of short-term monitoring period is shown in Figure 7.14.

[167] See (Amalfi et al. 1996).

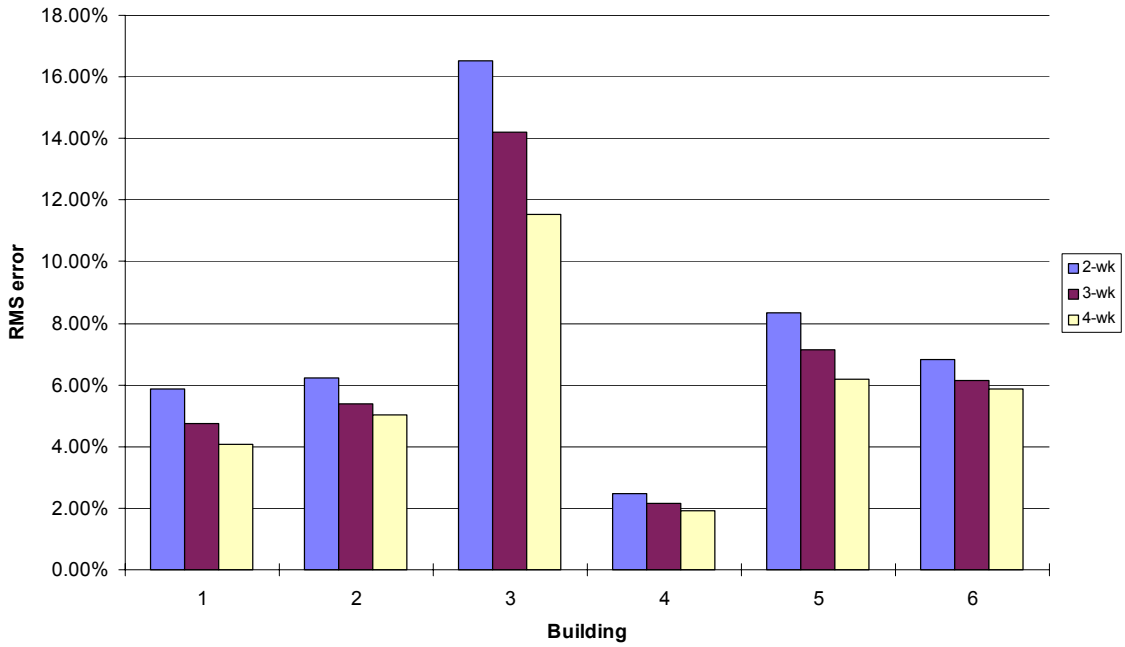

Figure 7.14: Extrapolation Error as a Function of Short-Term Monitoring Period

Increasing the monitoring period beyond two weeks can reduce the root mean squared error (RMSE) on the order of 10 to 30 percent. The majority of the costs involved in performing a short-term lighting study are contained in developing the monitoring plan, installation and retrieval of data loggers, and data analysis. Thus, increasing the study duration (assuming a single deployment of the data loggers) may not have a significant impact on costs, and can help to "smooth out" anomalous short-term behavior.

Instrumentation

A brief summary of instrumentation commonly used in M&V studies is provided below. For more information, consult the FEMP M&V Guidelines,[168] ASHRAE Guideline 14,[169] and Haberl et al.[170]

Hand-Held Watt Meters

Hand-held watt meters are commonly used to make instantaneous measurements of true electric power. In addition to true electric power, some meters also display voltage, current, kVA, and power factor. These directly measured values convey as-installed, or site-specific information about the equipment. This level of information can be helpful in estimating in-situ performance and reducing the variance between estimated and actual equipment performance due to field installation variances. For example, in lighting retrofit

[168] (* FEMP 2000).
[169] (* ASHRAE 2002).
[170] (Haberl, Sreshthaputra et al. 2003).

applications, it is common to find field-measured fixture demands that deviate 30% or more from the rated values available in manufacturers' catalogs. In-situ measurements of motor power often deviate from nameplate values, due to motor over-sizing.

Hand-held instruments are generally used to measure the power of constant load, constant operating hour equipment. When equipment power varies with time, time-series measurements using a recording meter may be necessary. Most hand-held instruments make single-phase power measurements. For three-phase loads, it is often necessary to make measurements on each phase and combine the measurements to obtain the total power. Good quality watt meters provide true power measurements. No adjustments for power factor need to be made to determine the demand because real-time measurement of Amps, Volts, and Power Factor are taken simultaneously, providing a true power measurement. Certain types of equipment, such as electronic ballasts, desktop computers, and variable frequency drives create harmonics on the power system. It is important to use a good quality meter capable of RMS power measurements when measuring power on these types of equipment.

Figure 7.15: Hand Held Watt Meter

Watt Transducers

Watt transducers are devices that provide an output signal in proportion to kW and/or kWh. These devices generally do not provide a visual readout, and are designed to be connected to a separate data acquisition or energy management system. Older style analog watt transducers use a Hall-effect device to measure true power. Newer digital watt transducers use high-speed sampling of voltage and current signals to calculate true power. In either case, be sure to examine the accuracy specifications, including the ability to handle harmonic distortion when specifying a watt transducer. Watt transducers can provide an analog DC voltage output, a 4-20 ma output, or a pulse output signal.

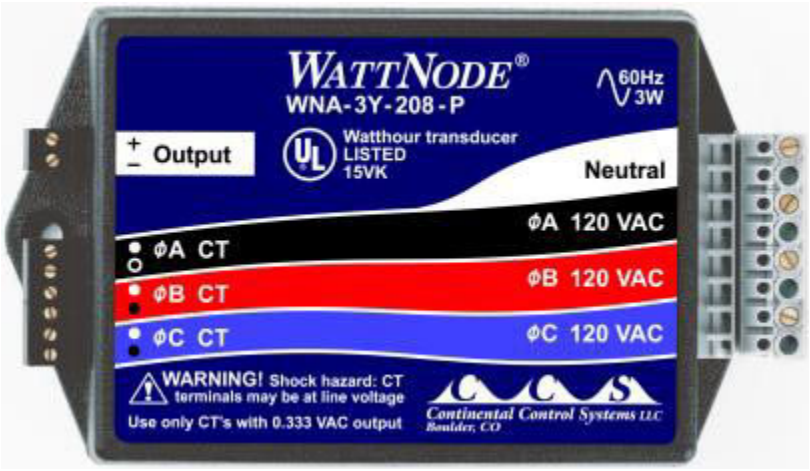

Figure 7.16: Watt Transducer

Data Acquisition Systems

Data acquisition systems are multi-channel instruments that are generally hard-wired into the building electrical system. They are generally wall mounted and connected to various sensors located throughout the building via instrumentation wire. Due to the installation expense, these instruments are used primarily for long-term monitoring projects. They are designed to accept signals from a large of variety sensors, such as temperature, humidity, flow rate, etc. Most data acquisition systems used in building energy monitoring also have built in power measurement capability.

Data acquisition system installation requires: (1) mounting the system near the electrical supply panel to obtain system power and reference phase voltages, (2) installing sensors that capture the engineering parameters of interest, (3) connecting the sensors to the system, and (4) calibrating and commissioning the system. Once installed, the system is periodically downloaded by a personal computer over a telephone modem or internet connection. An example of a hard-wired datalogger is shown in Figure 7.17.

Figure 7.17: Hard-Wired Datalogger

Portable Data Loggers
In recent years, small, battery powered data loggers have come onto the market. These devices have revolutionized the building measurement industry, providing the capability to make short-term, non-intrusive measurements of building performance at very low cost. Portable data loggers are generally used for ***short-term monitoring*** in ***temporary installations***. Multiple loggers are generally used, requiring little or no wiring between the logger and the sensors. Loggers are generally left in place for the duration of the monitoring period, then retrieved and downloaded off-site, rather than downloaded remotely.

On/off status and run-time loggers are used in projects involving constant load, variable operating hour devices. These units simply record whether or not a device is "on," and, combined with spot-measurements of input power, allow a calculation of accumulated kWh. The most basic form of these instruments simply record accumulated run-hours. Others record the date and time of each transition from on to off.

Optically Triggered Run-Hour Recorders. An optically triggered operating hour recorder is a small device with an optical photocell sensor. The most popular application of the device is monitoring lighting fixture run-hours. To discriminate between fixture and background lighting levels, most units provide an adjustable threshold that will trigger run-time logging whenever lighting levels exceed the threshold value. The device can be directly attached within the fixture or remotely mounted with an optical fiber sensor extender. The basic run-hour logger simply stores accumulated run hours.

Optically Triggered On/Off Recorders. Optically triggered recorders are available which generate a record for every change-of-state. Change-of-state simply means when the light is turned from "off" to "on" or vice versa. Onboard memory stores the time and conditional state into records that can be downloaded. This file can be used to simulate a time-series on/off lighting operating schedule. An example of an optically coupled on/off recorder is shown in Figure 7.18.

Figure 7.18: Optically Coupled On/Off Recorder

Current-Triggered Recorders. Current-triggered operating hour recorders are available which function similarly to the optically triggered recorders. Both accumulated run-time

and change-of-state recorders are available. Instead of utilizing a photocell to provide the control logic for run-hour accumulation, an internal magnetic field sensor or external current transducer is used. When the device with an internal sensor is placed on an electric motor, it accumulates run-time by sensing the magnetic field from the motor during operation. For devices with external current transducers, the device begins accumulating run-time when a minimum threshold current through the load supply wiring is exceeded. The current transducers are split-core, clothespin type devices attached around the electrical supply wiring, generally at a breaker panel or disconnect box.

Single-Channel Portable Data Loggers. Single-channel portable data loggers are small, single purpose loggers that make time-series measurements of quantities such as temperature, relative humidity, light level, current, etc. The loggers are called "single purpose" because they are designed to make only one type of measurement. They can be specified with integral or remote sensors. The loggers are quite small, battery powered, and can be configured for short-term or long-term deployment. The length of the monitoring period is constrained by the size of the memory and the measurement frequency.

Figure 7.19: Single Channel Datalogger

Multi-Channel Portable Data Loggers. Small, portable, multi-channel data loggers are available that make time-series measurements of quantities such as temperature, relative humidity, current, voltage, etc. These devices may be either single- or multi-purpose. Single-purpose loggers are designed with a fixed-channel configuration. Multi-purpose loggers can be configured to make any combination of measurements on each channel. Small multi-channel loggers are available with two to sixteen channels, but four channels is a popular configuration. Some models feature digital real-time displays and sensor excitation power. Monitoring periods are defined by measurement frequency, size of internal memory, and battery capacity. An example of a multi-channel portable datalogger is shown in Figure 7.20.

Figure 7.20: Multi-Channel Portable Datalogger

Measure Sampling

Measure sampling is a key issue in virtually every M&V study. In most cases, it is impractical to monitor every piece of equipment of interest to the study. Some sampling will inevitably be required. Some M&V studies likely to involve measure sampling include:

- Light-logger studies of fixture run-time
- Lighting circuit monitoring
- HVAC monitoring at a facility with multiple packaged air conditioning units.

A sampling plan, as described in Chapter 13, Sampling, should be developed and followed. This will allow quantification of sampling error. Monitoring should not be done only on equipment that is "convenient" to monitor.

An example of a light logger study conducted at a small commercial building is shown below. A total of twenty-six control points (light switches) were used in the space. Each control point is displayed on the horizontal axis, and the associated connected load (watts of fixture power) are displayed on the vertical axis. A stratified random sampling plan based on the connected watts of each control point was developed. Light loggers were installed in fixtures controlled by the sampled control points as indicated in Figure 7.21.

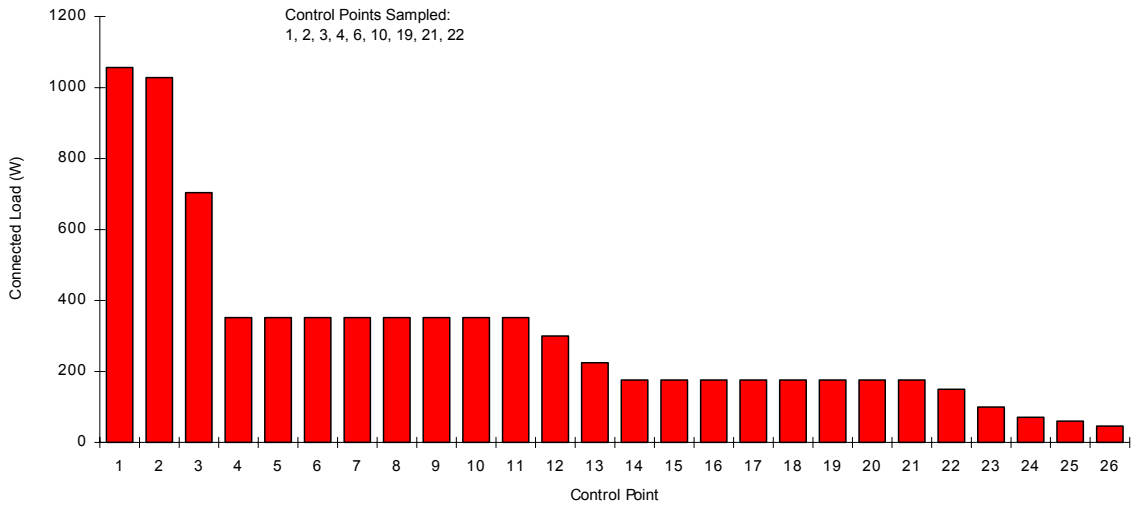

Figure 7.21: Control Point Sampling Strategy for Light Logger Study

Sensor Placement

Sensor placement is a key data acquisition issue. The most accurate sensor and recorder in the industry will not provide good data if the sensor is not placed appropriately. For example, the temperature recorded by a temperature sensor represents an equilibrium condition between the sensor and its surroundings. If a fluid temperature measurement is desired, then the thermal equilibrium temperature of the sensor should be as close to the fluid temperature as possible. In some cases, it may be possible to immerse the sensor in the fluid, or mount the sensor in a thermal well. In other cases, it may be necessary to attach a sensor to a pipe wall, and measure the temperature of the pipe surface as a proxy for the fluid temperature. In this case, a tight connection between the sensor and the pipe surface is required, with sufficient insulation around the sensor and pipe such that the sensor temperature is as close to the fluid temperature as possible. Other examples of sensor placement issues include:

- Air temperature measurements in ducts with non-uniform (stratified) air temperatures. This is especially important in the mixing section (where return air and outdoor air mix upstream of a heating or cooling coil) of an air handler or HVAC unit
- Flow meters installed too close to a flow obstruction
- Current transducers used in electric power measurements facing the wrong way
- Outdoor temperature sensors with inadequate shielding from direct solar radiation, or located in an exhaust air stream
- Optically coupled light loggers mounted where they could be triggered by daylight as well as electric light

Measurement Uncertainty

Total measurement uncertainty should be calculated using a propagation of error technique as described in Chapter 11, Uncertainty. Measurement uncertainty calculations should include the following factors:

- Sensor accuracy
- Recording system accuracy
- Data display or recording resolution
- Sampling error

See ASHRAE Guideline 14, Annex C for additional information on calculating measurement errors.

Quality Control

M&V quality control procedures include activities up and down the data collection and analysis process. The quality control plan should address the following issues as applicable:

- Sensor calibration
- Data quality control procedures, as outlined in the M&V plan
- Baseline model data fit criteria: mean bias error, RMS error, CVRMSE
- Checklist on data formatting and reporting requirements

Data Warehouse

A "data warehouse" for compiling and disseminating M&V study results should be established. This center can serve as a resource for the evaluation community, allowing the broad dissemination of evaluation data to improve deemed savings estimates and stipulated parameter assumptions used in M&V, program design, and portfolio planning. Use of quality M&V data resources can reduce the need for redundant data collection and focus resources on issues requiring additional study.

The overall roadmap through the data warehouse path is shown in Figure 7.22.

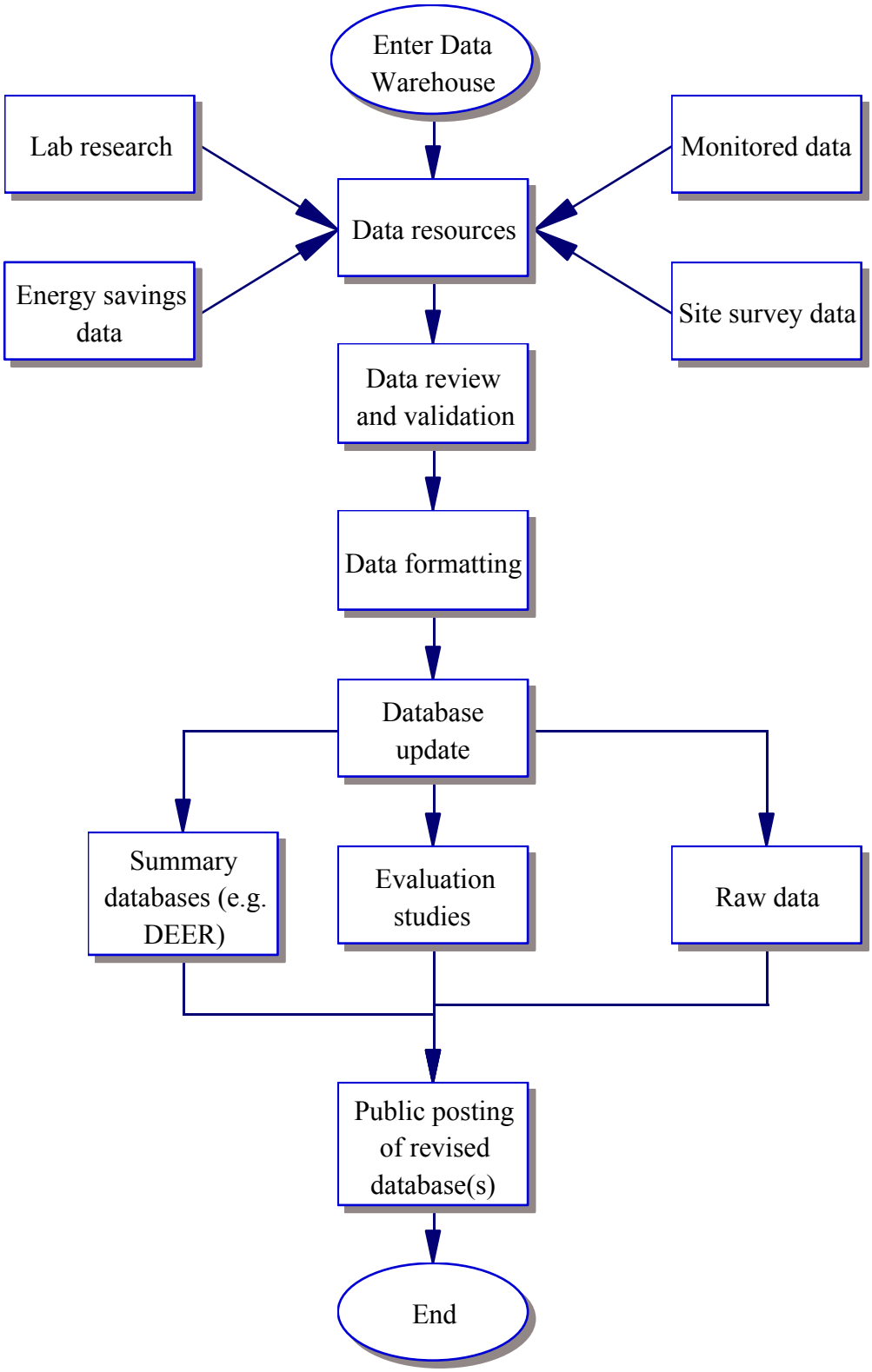

Figure 7.22: Data Warehouse Roadmap

The data warehouse can serve as a clearinghouse for engineering data related to measure performance. These data, or other data submitted and approved in the program implementation plan, can form the basis of the stipulated parameters described in the IPMVP Option A protocols. The relative precision of the parameters and the limits of applicability will form the basis of the uncertainty analysis performed during development of the M&V plan. The data warehouse will also contain data on building performance (total energy consumption) that can be used to examine the link between the presence of measures and the energy consumption of the building.

Data Resources

Data for the data warehouse can come from a variety of sources as appropriate. These are summarized below:

Laboratory Studies
Certain engineering data may be best developed under controlled conditions in a laboratory. For example, the sensitivity of the efficiency of an air conditioner to refrigerant charge and air flow variation may be studied in a laboratory, using instrumentation and test protocols that cannot be easily duplicated in field. Once this relationship is established, it can be applied to field measurements of refrigerant charge and air flow to estimate the impacts of correcting these problems.

Engineering Field Studies
Studies requiring more instrumentation and data analysis than can be reasonably supported under an M&V study can be separately funded and conducted. Examples include performance verification of emerging technologies.

Field Data
M&V data collection to establish engineering parameters for energy savings calculations and unit energy savings form the basis of the data warehouse. Data such as equivalent full load operating hours for HVAC equipment in various climates and applications, energy savings from daylighting controls in specific building types, and the variability of field measured versus nameplate efficiency for measures promoted by efficiency programs will be very useful data for both improving program design and focusing M&V resources appropriately.

On-Site Survey Data
Many program activities involving building energy simulation include whole-building surveys of building characteristics. These data can be a valuable resource for a variety of studies. One example of an existing database of survey information is the Nonresidential New Construction (NRNC) Database[171] a collection of 990 on-site building surveys covering commercial new construction in California. Development of similar databases of building survey information can be of great value for conducting efficiency potential studies, building codes and standards upgrade analyses, etc.

[171] Available at www.calmac.org. (RLW Analytics 2001).

Data Review, Validation, and Analysis

An authority (such as CALMAC) should be mandated and funded to maintain performance measurement data in the data warehouse. New studies reviewed and approved by the authority can be accepted into the database. This oversight will provide quality control and confidence in the quality of the data. The authority should periodically review gaps in the engineering data, and suggest specific data collection and analytical research projects to fill the gaps to reduce the uncertainty in engineering parameters.

Data Formatting

Unit energy savings and engineering parameters collected during the M&V analysis should include a reference indicating their source, uncertainty estimates when available, and limits of their applicability. These data should be delivered and stored in a standard format. The reporting format suggested in the M&V plan section above is designed to provide sufficient background information to accompany the measure performance data. Unit savings estimates, using the designated units shown in Table 7.2, along with other engineering parameters should be formatted and stored in the database. One promising standard for formatting whole-building survey data used in defining building energy simulation models is the gbXML[172] standard.

Database Update

The authority should update the M&V data database on a periodic basis, corresponding to the submittal of EM&V reports and the conclusion of special research studies. Other databases referencing the EM&V database such as the DEER[173] should be updated in a timely manner to provide feedback to program planning, efficiency potential studies, and subsequent EM&V studies. The data warehouse should also contain a collection of EM&V reports (similar to the current CALMAC website) and raw M&V data as appropriate.

Data Access

Access to the database should be unlimited for formatted data products such as DEER. Access to more technical data products may be restricted based on user registration and an appropriate disclaimer.

---

[172] gbXML was developed by Geopraxis (Petaluma, CA), with funding from the California Energy Commission under the PIER program. (Green Building XML (gbXML) 2003).

[173] (Xenergy, ADM Associates et al. 2001).

Appendix to Chapter 7: Uncertainty Calculation Example

This example describes an uncertainty analysis applied to a lighting retrofit project. The annual energy savings from a commercial lighting retrofit program are estimated using an engineering analysis. The gross annual energy savings are calculated from:

$$\Delta kWh = units \times \left[\frac{(Watts \times DF)_{base} - (Watts \times DF)_{ee}}{1000} \right] \times FLH \times (1 + HVAC_c) \qquad (10)$$

where:

ΔkWh = gross annual energy savings
units = number of units installed under the program
$Watts_{ee}$ = connected load of energy efficient unit
$Watts_{base}$ = connected load of baseline unit(s) displaced
FLH = full-load operating hours
DF = demand diversity factor
$HVAC_c$ = HVAC system interaction factor for annual energy consumption

Engineer parameter values and their associated uncertainties are listed below.

Table 7.11: Data for Lighting Retrofit Program

Parameter	Value	Uncertainty	Error
units	10,000	± 1%	± 100
$Watts_{ee}$	105W	± 1%	± 1.1
$Watts_{base}$	160W	± 7%	± 11.2
FLH	2,860 hr	± 25%	± 715
DF_{ee}	0.8	± 5%	± 0.04
DF_{base}	0.75	± 5%	± 0.038
$HVAC_c$	0.24	± 15%	± 0.036

From the data in the Table above,

$$\Delta kWh = 10{,}000 \times \left[\frac{(160 \times 0.75) - (105 \times 0.80)}{1000} \right] \times 2860 \times (1 + 0.24)$$

$$= 1.28 \text{ GWh}$$

The uncertainty in the estimate is calculated from a propagation of error analysis, as outlined in Chapter 11, Uncertainty. The calculations are reviewed here:

Consider an engineering equation with the general form:

$$R = f(p_1, p_2, p_3 \ldots) \tag{11}$$

where:

R	=	result
p_1	=	parameter 1
p_2	=	parameter 2
p_3	=	parameter 3, etc.

The equation for estimating the error in the overall calculation from the error in the individual parameters is shown below:

$$e_R = \sqrt{\left(\frac{\partial R}{\partial p_1} \times e_{p_1}\right)^2 + \left(\frac{\partial R}{\partial p_2} \times e_{p_2}\right)^2 + \left(\frac{\partial R}{\partial p_3} \times e_{p_3}\right)^2 + \cdots} \tag{12}$$

where:

e_R	=	error in results
e_{p1}	=	error in parameter 1
e_{p2}	=	error in parameter 2, etc.

The error in the gross annual energy savings calculation is determined from:

$$e_{\Delta kWh} = \left[\left(\frac{\partial \Delta kWh}{\partial units} \times e_{units}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial Watts_{base}} \times e_{Watts_{base}}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial Watts_{ee}} \times e_{Watts_{ee}}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial DF_{base}} \times e_{DF_{base}}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial DF_{ee}} \times e_{DF_{ee}}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial FLH} \times e_{FLH}\right)^2 + \left(\frac{\partial \Delta kWh}{\partial HVAC_c} \times e_{HVAC_c}\right)^2\right]^{\frac{1}{2}} \tag{4}$$

Using the data in Table 7.6, the partial derivatives are evaluated as follows:

$$\frac{\partial \Delta kWh}{\partial units} = \left[\frac{(Watts \times DF)_{base} - (Watts \times DF)_{ee}}{1000}\right] \times FLH \times (1 + HVAC_c)$$

$$= \left[\frac{(160 \times 0.75) - (105 \times 0.80)}{1000}\right] \times 2860 \times (1 + 0.24)$$

$$= 128$$

$$\frac{\partial \Delta kWh}{\partial Watts_{base}} = units \times DF_{base} / 1000 \times FLH \times (1 + HVAC_c)$$

$$= 10{,}000 \times 0.75 / 1000 \times 2860 \times (1 + 0.24)$$

$$= 26{,}600$$

$$\frac{\partial \Delta kWh}{\partial Watts_{ee}} = - units \times DF_{ee} / 1000 \times FLH \times (1 + HVAC_c)$$

$$= -10{,}000 \times 0.80 / 1000 \times 2860 \times (1 + 0.24)$$

$$= -28{,}370$$

$$\frac{\partial \Delta kWh}{\partial DF_{base}} = units \times Watts_{base} / 1000 \times FLH \times (1 + HVAC_c)$$

$$= 10{,}000 \times 160 / 1000 \times 2860 \times (1 + 0.24)$$

$$= 5.67 \times 10^6$$

$$\frac{\partial \Delta kWh}{\partial DF_{ee}} = units \times Watts_{ee} / 1000 \times FLH \times (1 + HVAC_c)$$

$$= 10{,}000 \times 105 / 1000 \times 2860 \times (1 + 0.24)$$

$$= 3.72 \times 10^6$$

$$\frac{\partial \Delta kWh}{\partial FLH} = units \times \left[\frac{(Watts \times DF)_{base} - (Watts \times DF)_{ee}}{1000} \right] \times (1 + HVAC_c)$$

$$= 10{,}000 \times \left[\frac{(160 \times 0.75) - (105 \times 0.80)}{1000} \right] \times (1 + 0.24)$$

$$= 446$$

$$\frac{\partial \Delta kWh}{\partial HVAC_c} = units \times \left[\frac{(Watts \times DF)_{base} - (Watts \times DF)_{ee}}{1000} \right] \times FLH$$

$$= 10{,}000 \times \left[\frac{(160 \times 0.75) - (105 \times 0.80)}{1000} \right] \times 2860$$

$$= 1.029 \times 10^6$$

The individual error terms are calculated and summarized below:

Table 7.12: Summary of Error Calculations

Parameter	$\frac{\partial R}{\partial p_i}$	e_{p_i}	$\left(\frac{\partial R}{\partial p_i} \times e_{p_i}\right)^2$
units	128	± 100	1.63×10^8
$Watts_{ee}$	-28,370	± 1.05	8.87×10^8
$Watts_{base}$	26,600	± 11	8.87×10^{10}
DF_{ee}	3.72×10^6	± 0.04	2.22×10^{10}
DF_{base}	5.67×10^6	± 0.038	4.53×10^{10}
FLH	446	± 715	1.02×10^{11}
$HVAC_c$	1.029×10^6	± 0.036	1.37×10^9

The overall error in the annual energy savings estimate is calculated as follows:

$$e_{\Delta kWh} = \sqrt{1.63 \times 10^8 + 8.87 \times 10^8 + 8.87 \times 10^{10} + 2.22 \times 10^{10} + 4.53 \times 10^{10} + 1.02 \times 10^{11} + 1.37 \times 10^9}$$

$$= 0.51 \text{ GWh}$$

Expressed as a fraction of the estimate:

$$\frac{e_{\Delta kWh}}{\Delta kWh} = \frac{0.51}{1.28}$$

$$= 0.40$$

The uncertainty in the engineering estimate of savings prior to any M&V activity is ±40%. The last column in the previous table is an indication of the relative influence of each parameter on the overall uncertainty of the estimate. To reduce the uncertainty in the overall estimate, data gathering activities should focus on reducing uncertainty in the parameters that have the greatest contribution to the overall uncertainty. The priorities for data gathering for estimating annual energy savings are as follows.

Table 7.13: Data Gathering Priority Ranking for Gross Annual Energy Savings Estimates

Rank	Parameter	Parameter Uncertainty	Contribution to Overall Uncertainty
1.	FLH	± 25%	39%
2.	$Watts_{base}$	± 7%	34%
3.	DF_{base}	± 5%	17%
4.	DF_{ee}	± 5%	9%
5.	$Watts_{ee}$	± 1%	< 1%
6.	units	± 1%	< 1%
7.	$HVAC_c$	± 15%	< 1%

Chapter 8: Process Evaluation

Preface

The process evaluation is an important tool in the evaluation toolbox. The process evaluation consists of in-depth examinations of the design, delivery and operations of energy programs in order to improve the ability of the program to achieve energy savings and accomplish other program goals. The process evaluation also provides a vehicle for sharing program design and operational improvements with other professionals in the field. When process evaluation results are shared with other energy efficiency professionals, these professionals can assess the relevance of the evaluation findings and recommendations to their policies, programs, and program portfolios. This is especially true for program designers and managers who may want to determine if the evaluation results can be used to improve the cost-effectiveness of their programs.

As with other chapters in this document, the process evaluation chapter is written for a wide range of stakeholders within the California energy program arena. This chapter should be read by regulatory staff and policy makers who need to understand the importance, purpose, scope and to a limited degree the tools associated with the process evaluation, and who are in a position to improve energy programs. This chapter is also intended for evaluation planners and designers as well as the staff who must conduct process evaluations. This is especially true for evaluators who are new to the process evaluation field. Program administrators who may be responsible for funding, or coordinating the process evaluation effort would also benefit from reading this chapter. Likewise, program managers who are requested to cooperate with the evaluation effort by providing access to program records and staff during the evaluation process would be helped by being familiar with this chapter. We also suggest that the evaluators of information and educational programs and evaluators who conduct market transformation evaluations read this chapter. These individuals may need to coordinate their evaluation efforts with the process evaluation efforts when similar types information are needed or when similar evaluation activities are conducted across their evaluation efforts.

This chapter presents a description of the skills needed to conduct process evaluations, a definition of process evaluation, and a discussion of the reasons why process evaluations are a component of the Framework. The chapter then provides a presentation of the typical evaluation tools, types of process evaluations and examples of evaluation approaches. The chapter ends with a presentation of the recommended decision steps associated with planning and conducting process evaluations within the Framework followed by a discussion on process evaluation budgets.

Skills Required for Process Evaluations

The investigative processes associated with designing, managing and conducting process evaluations focus on a wide range of researchable issues. These issues can range from evaluating the ability of a program's data management system to support the informational needs of the program, to assessing if the program is well-designed, managed, targeted, marketed and operated. As a result, the skills needed to conduct process evaluations are varied, suggesting the need for different skills for different types of process evaluations.

Evaluations that focus on the design and operation of program information systems, for example, need evaluators that understand how information management and information availability influence a program's operations. However, the evaluators should also be skilled at designing, developing and implementing information systems, in order to recommend changes that improve the program's ability to cost-effectively achieve its goals. Process evaluators who assess program satisfaction levels need to have the skills to identify and analyze different program characteristics that influence satisfaction and be able to identify those characteristics that can be changed to improve satisfaction scores. Likewise evaluators that focus on assessing program targeting, marketing and promotional operations need to have skills necessary to assess information flow, content and presentation effects as well as the skills associated with understanding how markets and market segments operate and can be influenced through different outreach and promotional efforts. These examples demonstrate the need to match the skills of the process evaluator with the research goals of the specific process evaluation.

It is equally important that process evaluation managers be trained and/or experienced with the tools used in the process evaluation. For example, if a telephone survey is needed, evaluators need to be knowledgeable and experienced in the field of survey research and instrument design. If focus groups are needed evaluators should be knowledgeable and experienced in the field of focus group design and operation.

Because of the diversity of researchable issues associated with conducting process evaluations and the diversity of skills needed to address these issues it is difficult to define a specific set of skills needed to conduct these evaluations. Instead, the Framework recognizes that a diverse set of program assessment and information analysis skills are needed across the various investigative issues on which these evaluations typically focus. However, in general the process evaluator should have the following knowledge and skills:

- Expert knowledge of a wide range of energy efficiency programs and a strong understanding of their operational designs, management practices and program goals,
- Expert knowledge of different process evaluation data collection methods and approaches, and a working knowledge of the process evaluation literature and how evaluation approaches have been applied in the energy efficiency program field,

- A strong analysis capabilities and an expert understanding of cause-and-effect relationships that impact the ability of energy efficiency programs to cost-effectively accomplish their goals,
- A strong understanding of statistical analysis approaches and analytical procedures appropriate for the process evaluation research goals,
- A strong understanding of sampling methods and approaches and the ability to identify potential biases in a sampling approach and to develop control strategies for mitigating levels of bias, and a
- High level of past experience in conducting process evaluations of energy efficiency program and in reporting the results of these studies.

Introduction and Key Issues

Definition

For the purposes of the Framework, a process evaluation is defined as: *a systematic assessment of an energy efficiency "program" for the purposes of (1) documenting program operations at the time of the examination, and (2) identifying and recommending improvements that can be made to the program to increase the program's efficiency or effectiveness for acquiring energy resources while maintaining high levels of participant satisfaction.*

This definition includes an assessment of the products and services provided by the program or the ways in which the program is designed, operated and delivered. This definition also includes an assessment of the program's approaches to engaging and interacting with the target markets or market actors for the purpose of improving the program's cost-effectiveness. However, it should be noted that this definition focuses on identifying improvements or modifications to a program that directly or indirectly acquires energy impacts in the short-term (as in resource acquisition programs) or over the longer-term (as in market transformation programs) as a result of program actions. This definition specifically excludes the assessment of energy programs for purposes other than increasing the efficiency or effectiveness of the program to acquire energy resources, either directly or indirectly. For example, this definition excludes conducting management audits or evaluations for the purposes of supplementing a financial audit of a program, unless these examinations, at least in part, are conducted for the purposes of reducing the net cost of acquiring the energy impacts.

Why These Two Goals are in the Definition

Essentially this definition has two components, documenting program operations and assessing programs for improvements that increase energy savings and improve cost-effectiveness. The primary purpose of the documentation component of this definition is to provide a baseline description of the program operations and processes and to compare the designed and operational practices with the program theory. It is also important for

the consumers of process evaluation reports to have a full understanding of the current program processes and their importance in the program theory relationships prior to considering recommendations for change. Without the documentation component in the definition, it is possible for a process evaluation to report recommended changes to the program without a solid presentation of how the program operates, how operations support the program theory, or a description that allows for a full understanding of what it is that will be changed if the recommendation is carried out.

While most process evaluations focus their effort on the second of the definitional objectives, the first objective is an important part of the evaluation process. The first objective is important because the formal program plans and operational descriptions and manuals seldom provide an accurate or detailed reflection of how a program operates in practice. This is usually because of one or more of the following conditions:
Program planning and presentation documents seldom describe program operations at a level of detail that allows others to understand the operational practices that make it successful or capable of improvement. These documents are typically prepared for the purpose of proposal presentation, for program delivery contracting, for providing documents for public consumption, or for sharing internally with program managers and staff for familiarization or training purposes. Seldom are these documents prepared for the purpose of sharing detailed operational procedures or systems with other stakeholders.

Most programs evolve as they move through the implementation process. Even if program documents were originally developed with detailed step-by-step implementation descriptions, they are typically outdated in terms of their content and accuracy. This is not a criticism of energy programs or of their descriptive documents. Rather it is an acceptance of the understanding that programs evolve as they move through the design and delivery process and this evolution occurs on an ongoing basis.
In some cases program documents will describe designs and operational procedures that are officially part of the program process, and may even be reflected in a detailed operational manual, but for one or more reasons these activities are not replicated in the implementation practice. These planned practices can be, and often are, inadvertently or purposefully dropped or modified along the way. In many cases program delivery problems develop that act to reduce a program's efficiencies or its effectiveness because program operations or implementation efforts do not match officially approved designs or procedures. This is different from the second issue stated above in that these are not purposeful evolutions in the program designs but are typically non-performance issues.

As a result of these three considerations the process evaluation serves as a valuable tool to describe the operational conditions, systems and procedures at the time of the evaluation that have a direct impact on the ability of the program to accomplish its goals. The process evaluation provides a method for historical preservation and presentation of the detailed operations of a program that are being assessed in a way that others can understand how the program works and how program modification are expected to influence program results.

While the documentation process is important for programs funded with Public Goods Charge funds, there is a second important benefit associated with the documentation process that directly serves the process evaluation function. The process of documenting the program's operations and processes also provides the foundation for comparing how the program is operated in relation to the concepts presented in the program theory. The program theory presents the cause-and-effect rational supporting the program's designs and operational practices.[174] As a result, the process evaluation should find that the program is designed and operated in a way that is consistent with the program theory. The documentation process provides the information needed to determine if the program's operational practices are consistent with the program theory or if there are significant disconnects between the program theory and the way the program is operated. When the process evaluation finds that a program has operational components that are inconsistent with the program theory, the process evaluation can assess if the program theory needs to be modified or if there are aspects of the program that do not appear to be an important or relevant component of the program. Some evaluators argue that this is the most important focus of the process evaluation, especially during the early implementation period when operational practices are being tested and refined.

Certainly, the primary reason for conducting process evaluations is to identify and recommend changes in a program's operational procedures or systems that can be expected to improve the program's efficiency or cost-effectiveness. These recommendations need to be developed so that they support the program or the program's operational practices consistent with the program theory or with a recommended change to the program theory. Indeed, this goal covers a lot of territory, and includes issues internal to the program's operations (such as management systems and procedures) as well as issues external to the program (such as issues associated with customer acceptance and levels of customer satisfaction). As suggested in the *2001 Framework Study,*[175] a process evaluation also examines the extent to which a program is meeting or can meet a customer or participant need. Programs that are not capable of meeting a customer need cannot be expected to be successful from the perspective of the participating customer and as a result, impact goals may not be achieved.

The process evaluation takes on the challenge of evaluating most, if not all aspects associated with the design or operations of a program in order to improve the energy resources acquired (directly or indirectly) by that program. The process evaluation can examine a variety of issues. These often include:

- Program design, goal setting, common vision, improvement process,
- Program staffing, staff skills, training, management and operations,
- Program information and information support systems,
- Program targeting, marketing and outreach efforts,
- Program goal attainment and implementation processes and results,

[174] See Chapter 4: Evaluation Overview and Issues, for information on the use of program theory in the evaluation process.

[175] *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*. (* Sebold, Fields et al. 2001). See the work written by Dr. Jane Peters.

- Program theories, theory assumptions and key theory relationships - especially their causal relationships,
- Program timing and timelines,
- Participant satisfaction (both overall and individual components that interact with the participant in order for these to be assessed),
- Quality control procedures and processes,
- Reasons for low participation rates,
- Reasons for overly high free riders, or to low a level of market effects, free drivers or spillover,
- Use of new practices or best practices, and
- Intended or unanticipated market effects, among others.

The Need for Process Evaluations

The need for process evaluations of energy programs began to be recognized in the late 1970s and early 1980s when the United States Department of Energy (USDOE) began developing energy conservation and energy efficiency programs to help stretch U.S. energy supplies. Several of these early programs were experiencing customer participation and acceptance problems. As a result, the USDOE and the state agencies implementing these programs began to conduct evaluations that focused on the reasons associated with low participation or customer acceptance problems. Examples of these problems included energy efficiency loan programs administered through established leading institutions that were targeted at customers with poor credit or who do not believe in borrowing, or energy audit programs that provided audit reports that were not user-friendly and were, as a result, not read or acted upon by program participants. The USDOE and state agencies learned that they needed to research program designs and operational systems to make sure they were appropriate for the market in which they were targeting. In these early years, energy program process evaluations were typically incorporated into other evaluation activities such as impact evaluations or market research studies about a particular product or service. As the energy program field matured, program designers, managers, policy makers and regulatory managers began to see the value in conducting process evaluations to improve program designs, to acquire more participants and increase the attainment of cost-effective energy savings. In the mid 1980s utility companies began to offer a wide range of demand side management (DSM) programs. These program providers quickly recognized the benefits of the process evaluation and began to focus evaluations specifically for the purpose of assessing the design, operations, performance, and customer interaction and acceptance characteristics of their DSM programs. In many cases these evaluations were in response to regulatory agencies needing to be assured that programs were well-designed and implemented.

By the early 1990s process evaluations had become a widely accepted and desired evaluation research tool within the energy program industry. In recognition of this fact the Electric Power Research Institute (EPRI) published the *DSM Process Evaluation Guidebook to Current Practice.*[176] This publication presented and described how the

[176] *DSM Process Evaluation: A Guidebook To Current Practice.* (* EPRI 1992).

process evaluation fit within an integrated program evaluation system and demonstrated how process evaluations were instrumental in the program design and redesign process leading to higher quality, more effective programs. According to the EPRI "Guidebook" the primary purposes of the process evaluation are to:

- Improve program performance with respect to internal administration, promotional practices, program delivery, incentive levels and data management,
- Provide information to regulators and other interested parties that [energy programs] are being implemented effectively and modified or refined as necessary,
- As a means of improving customer satisfaction and identifying market threats and opportunities, and
- As a means of contributing to industry-wide knowledge in order that other providers many improve their programs.

These goals are consistent to the process evaluation goals presented in an earlier published document by Ben Bronfman and Jane Peters[177] in 1991. In this document the goals of the process evaluation are to:

- Improve program implementation efficiency,
- Assess market segments and targeting of specific segments,
- Improve the quality of the measures installed,
- Identify program design issues,
- Providing an accounting of program progress, and
- Examine special issues (measure life, program comprehensiveness, etc.).

These goals are just as important today as they were in 1992. In California, energy resource suppliers, policy makers and regulators recognize the need for reliable program evaluation information as a tool to be able to more fully understand the potential energy supply and supply capabilities available through the State's energy programs. The State of California has adopted a policy decision that energy efficiency and renewable energy resources will contribute to lowering future per capita demand increases in California. As a result, the process evaluation has become an important tool in helping California policy makers understand the ability of energy programs to provide reliable energy resources and understand which programs and program operational and design practices can most effectively meet this goal. In California, energy programs are essentially a part of a statewide effort to make sure the State has reliable cost-effective methods of meeting the energy needs of its households and businesses.

[177] "Process Evaluations of DSM Programs." (Bronfman et al. 1991), p. 139.

Process Evaluations are In-Depth Studies

The process evaluation must be conducted at a sufficiently detailed level of investigation to be able to make clear and specific recommendations pertaining to what aspects of the program's management, structure, function, and operations need to be changed. The process evaluation should provide managers the information they need to make improvements to their programs, ranging from major improvements to the program's product or product lines, program structure, program service, program operations, and program marketing, to making small fine-tuning adjustments to products, services or operational methods.

The nature of the process evaluation is that it is an investigation into the detailed operations of the program, and often includes investigating the customer's levels of interaction with the program, the product and the service provided. For example, investigations into program operations focus on the structure, function and operational details associated with the program. Investigations into customer satisfaction focus on customer acceptance and use of a product within the three primary characteristics that influence satisfaction, including the customer-product relationship, the product-provider relationship and the customer-provider relationship. While program managers may track overall satisfaction scores for management tracking purposes, a process evaluation will investigate much deeper into the satisfaction ratings and investigate the characteristics of the program, product or delivery methods that influence satisfaction and identify specific program, product or delivery approach changes that can increase satisfaction and at the same time cost-effectively accomplish the energy savings goals.

Evaluation Timing

Process evaluations can be conducted at any time within the program design and implementation process. There are, however, some key considerations for the timing of these evaluations associated with the Framework. These include:

1. Design Feedback. In many cases, process evaluators can help programs be more effective or have more efficient operational systems before they are placed in the market. When process evaluation experts work with program designers and managers during the early development and implementation period, the evaluation staff can help identify potential problems associated with early program designs or operational practices. In addition, the process evaluation professional can work with program designers and managers to make sure they set up data management and operational systems that can operate to provide necessary reporting and evaluation data, and therefore be successfully evaluated during the process evaluation efforts. Evaluation experts have become accustomed to hearing the phrase "*we did not design the system to collect or trace the information you need to conduct your evaluation*" when they are not involved in the program design and planning process. Program designers and managers that work with their evaluation experts are better able to track and report their program progress and their programs are more easily, more efficiently and more economically evaluated. This means that incorporating process evaluation feedback into the

early program planning and implementation phases can be a very good investment for the program. It is almost always more expensive and more time consuming to modify program designs to collect the information needed to support the program evaluation effort after program data collection and tracking systems have been designed and are in operation.

2. Early Implementation Feedback. Early program evaluation feedback to the program designers and managers is an important component of any evaluation. If the process evaluation expert is brought into the program early, the evaluation expert can work directly with the program designers and managers to identify and agree upon an early feedback system that allows managers to be aware of early evaluation findings, and take corrective actions where appropriate before the evaluation report is finalized. These early feedback systems provide the program designer and manager with a method of ongoing interaction with the evaluation professional to make sure evaluation results are communicated to program management to allow for program improvements as early as possible.
3. Scheduling the Evaluation Efforts. Process evaluations should be conducted for all significant energy programs early in the implementation process so that improvements to the programs can be identified and initiated over a period of time in which the results from the evaluation can influence the program being evaluated. While a process evaluation conducted near the end of a program funding cycle may help future programs (including extensions of current programs), the evaluation should be conducted in time to help the current program more effectively or more efficiently achieve its goals. This means that the process evaluation may need to be conducted after the program's start-up issues have been identified and dealt with by program management, when the program is considered to be in a steady-state mode of operation. Typically a new program requires about four to six months to move from the early planning and organization phase, into the early market entry phase and on to a more normal, steady-state operational phase. Typically this means that an in-depth process evaluation can be initiated in about the sixth month following program rollout. However, there may be a need to schedule the process evaluation over multiple years if the evaluation budget does not allow for a full process evaluation during the first year or if the program is implemented in phases such that process evaluations focus on each phase of the rollout. Likewise, the process evaluation can be scheduled to support a planned program redesign process or a mid-course reassessment. A part of this effort may be to compare the program's design and operations to the program theory and logic models. This comparison can assess if a program design or operational change can cost-effectively improve goal attainment or to see if the program theory or the logic model needs adjusting to reflect program operations. In this case the process evaluation may be conducted early enough to inform the redesign process and then later to assess the success of the redesign effort.

Process evaluations can also focus on and address different researchable issues during different phases of the program implementation and evaluation cycle. It is not necessary to plan process evaluations for all aspects of the program to occur at the same time. The

benefit of scheduling and conducting process evaluations over the course of a program also have financial benefits in that the evaluation can be budgeted over the program funding cycle, rather than budgeted all at once. For example, the first year evaluation effort may want to focus on:

- Program design and operational systems,
- Program tracking and information management systems,
- Program delivery organization and staffing,
- Skill levels needed to implement the program,
- The methods and procedures used to target the outreach efforts,
- The marketing materials used to promote the program,
- Program operational efforts and their relationship to the program theory and logic model,
- The outreach efforts and the structure and content of these efforts, and
- Early program satisfaction and customer service experiences.

These issues can all be examined during the first year of the program operations. However, some aspects of the process evaluation may be more successfully assessed during the second or third year of a program cycle. For example, evaluations conducted after the first year might focus on:

- Assessing the relationship between the current program services and the needs of the market or participant,
- The program implementation system and its influence on customer perception of the program,
- The influence of the program on customer behavior and actions, and
- Field practices and their effects on energy savings achieved.

Most programs do not need a process evaluation every year of their implementation cycle. As stated earlier, new programs may want to undergo a process evaluation in the first year and involve the program evaluation staff early in the design process. For this reason, programs may want their process evaluation team on board and engaged during the early development efforts and for conducting the formal process evaluation initiated within the first year. However, if there have been previous process evaluation efforts for the program, the planning and conducting of the process evaluation needs to be informed by the past evaluation history. If the program has already had a rigorous process evaluation of the program's key components and the program has not significantly changed since that time, there may not be a need for a process evaluation for that program. The process evaluation decision path discussed later in this chapter provides guidance for when and under what conditions process evaluations of energy programs should be considered.

Process Evaluation Activities

Process evaluations involve a wide range of activities in order to investigate the researchable issues associated with energy programs. These activities include, but are not limited to the use of:

a. Reviews and tests of records, materials, tools, etc.,
b. Interviews and discussions with program management and staff, implementing contractors, subcontractors, and field staff,
c. Interviews and discussions with policy makers, key stakeholders and market actors,
d. Interviews, discussions, surveys and/or focus groups with participants and non-participants,
e. Collection and analysis of relevant data available from third-party sources (e.g. equipment vendors),
f. Field observations, measurements and examinations, and
g. Other activities as needed to address researchable issues.

In the following paragraphs we elaborate briefly on each of these items.

Examinations and tests of records and materials
Examination of program records and materials will give the evaluator a better understanding for how the program is implemented in practice and allows for testing of the content and accuracy of the program records. As part of this assessment, a sample of customer files can be examined to test for content and accuracy of the information collected and to assess the effects of incomplete, inaccurate or uncollected information on the ability to effectively manage and operate the program. This assessment can include a review of both participant and non-participant data as well as an assessment of the ability of the records to support management and operational needs.

Program electronic tracking systems and reports are important tools that program managers generally use to keep track of program activities and make sure that program activities, both internal and external, are proceeding on a reasonable schedule in a way that will achieve the program goals. An examination of these systems and reports can help the evaluator assess the strengths and weaknesses of the program operations and can also lead to the development of recommendations on how the tracking system and system reports can be improved. Likewise, this evaluation can examine the design of the database and database systems to assess the ability of the database to be modified, expanded, or upgraded and assess the ability of the tracking system reports to meet the needs of the program operations.

The program database is often a very valuable source of information. The database can be used to track, manage and report program marketing contacts, numbers of customers

at different stages in the program participation process, and customers who start to participate in the program but then drop out. Databases can also be examined to determine the time that elapses between each stage of program implementation for a customer in order to identify barriers or roadblocks. Finally, program databases are a key source of names to draw from for subsequent evaluation interviews, focus groups and surveys.

Reviews of management and program operations
The process evaluation can conduct a review of program operations and management efforts. For example, this review can include discussions with program staff detailing their efforts, activities and responsibilities and the steps they take to keep the program on track, and discussions with program managers and supervisors to review operational responsibilities and activities and identify monitoring efforts. The process evaluation can then chart the management and operational systems and compare those to the program theory and logic models to see if program management and operational efforts are consistent with the primary program goals. Such reviews can help the evaluator assess what should be considered to improve the management and operations of the program.

Examinations of customer handling and service delivery efforts
Customer handling and service delivery are core components of most energy programs. How customers are handled and the way in which services are delivered can lead directly to implemented projects and significantly influence satisfaction. Often customer handling and service delivery efforts directly affect the ability of the program to obtain participants who take the actions needed to provide the energy saved. A review of these procedures can include discussions with project staff, observation of interaction with the customer or participant, and conducting interviews or surveys with a sample of participants.

Assessments of operational tools and procedures
For many programs the ability to accomplish its intended goals is influenced by the tools and equipment that are required to operate the program, including the use of software and hardware. A process evaluation can assess the program tools to determine if the tools and the operational procedures relating to the use of the tools are consistent with the ability of the program to accomplish its goals while maintaining strong customer satisfaction ratings.

Interviews and discussions with program management and staff, implementing contractors, subcontractors, and field staff
Program management and staff can be a valuable source of information, as they typically know the program better than anyone. Interviews with lead program planners and managers, their supervisors, and a sample of program staff, including both central staff and field staff, can help the evaluator assess the program design and operations in order to make recommendations for changes that will improve the ability of the program to cost-effectively obtain energy savings. In some cases it may also be useful to interview

the administrative staff such as the staff responsible for program databases and for processing of incentive payments in order to support the evaluation.

These interviews can examine all aspects of the program's operations. This may start with each party's understanding of the program's goals and objectives. It is important to know whether different program personnel or stakeholders have a common vision for the program or if there are significant differences that may impact the program's success. The interviews can focus on the processes and activities of the key members of the program team and how their efforts support the program's goals and objectives. The interviews can be used to obtain detailed information on the program processes as practiced compared to the planned processes. Other subjects important to cover across different personnel include communications within the program, communications with customers, and communications with stakeholders. In addition the interviews can obtain impressions of the program's strengths and weaknesses and perceptions of the program's successes and the quality of work that can be compared and contrasted with those perceptions from stakeholders and participants. These interviews provide an opportunity to gather recommendations for program improvements from the interviewed personnel.

Interviews and discussions with policy makers, key stakeholders and market actors
In addition to program staff, many other individuals are involved in a program including policy makers (e.g. CPUC staff), utility managers, key stakeholders (e.g. trade associations and tenant groups), and a variety of other market actors such as product manufacturers, distributors, and installation contractors. It can often be useful to interview a sample of these key players in order to obtain their insights into what the program is doing well, and what can be improved.

Interviews, discussions, surveys and/or focus groups with participants and non-participants
One purpose of virtually all process evaluations is to understand the experience of the customers participating in the program in order to design program improvements resulting from those experiences. Program participants have valuable perspectives on what aspects of the program work well and what aspects represent barriers to participation or satisfaction. Detailed feedback from participants is also important for determining whether the customer's perceptions of specific program attributes and delivery procedures conflict or mesh with program design and management perceptions. Beneficial detailed feedback can include levels of satisfaction with the program and their participation experience and satisfaction with various elements of the program including satisfaction with the product(s), the organization, scheduling the service, educational services, quality of work performed, attitude of site staff, responsiveness to questions/concerns, level of savings achieved, etc.[178]

[178] Hall and Reed suggest that a process evaluation include three levels of satisfaction measurements, focusing on the relationships between the participant and the product or service received, the relationship between the participant and the organization implementing the program, and the relationship between the product or service provided and the organization providing the program. (Hall et al. 1997).

In addition, customers who are familiar with the program but elect not to participate (service rejecters) are also an important source of information. These customers can talk about barriers to participation and methods for overcoming those barriers. Depending on the number of program participants and the depth of the information to be collected, data from participating and non-participating customers can be collected with in-depth interviews, focus groups, and/or structured surveys. Generally, the smaller the number of participants and the more detailed the information collected, the more likely that in-depth interviews should be used. Questions that can be addressed using CATI systems typically focus on issues that can be addressed using simple questions leading to a "check the box" type of response. When more complicated or more detailed information is needed, a professionally conducted in-depth interview may be the best choice. Regardless of whether interviews, focus groups or surveys are used, care must be spent developing the questions that will be asked in order to provide objective responses at the level of depth needed to develop detailed process findings and recommendations.

There is a great deal of literature on developing and conducting high-quality surveys, interviews and focus groups. Unfortunately, it is easy for inexperienced people to think that anyone can design a survey, interview guide, or focus group guide. However, there is a large field of survey design research that quickly points to many easily overlooked problems in instrument and question design that can decrease the validity and reliability of the information collected lowering the usefulness of the information collected. Many of these issues become the basis for the different threats to validity discussed in Chapter 12, Uncertainty. Evaluators should be familiar with this literature and the state-of-the-art practice to be able to design meaningful and reliable evaluation studies. (Recommendations for using these tools and their associated methods can be found in most graduate level survey research design and survey development design textbooks.[179])

Collection and analysis of relevant data available from third-party sources

Data to support a process evaluation may be available from third-party sources or from sources of secondary data. For example, data on equipment stocking or equipment sales may be available from manufacturers or local distributors, or information on how markets work or how customers make decisions may be available from dealers and other venders. Likewise, census data or market information collected or maintained by other organizations can be helpful in conducting process evaluation assessments. For example, data from the federal ENERGY STAR® program can be helpful in understanding the markets for appliances that may be pushed by a California energy program. Other sources to support a process evaluation may include purchased market information from one of the for-profit companies that sell market or customer data. Customer, market or program information collected by other energy program evaluations may also help support process evaluation assessments.

[179] Examples include: *Applied Social Research*. (Rubin 1983); Chapter 10 covers an introduction for Survey Research; *Focus Groups: A Practical guide for Applied Research*, 2nd edition. (Krueger 1994); and *Focus Groups as Qualitative Research*, 2nd edition. (Morgan 1997).

Field observations, measurements and examinations
As noted above, sometimes it is useful for the process evaluator to go into the field with or without program staff to observe program operations, customer interaction practices, and customer education procedures and techniques used. Programs that have substantial on-site interaction with the participant involving measure installations, for example, may need to employ on-site examinations to assess the quality of the work provided. On-site observations can also be used in assessing the on-site approach and procedures associated with the program's interactions, relationships and job performance.

Sometimes important data is not collected by the program or by third-party sources, requiring the evaluator to collect data in the field. An example might be that the evaluator needs to conduct a quick survey on the retail price of eligible equipment at a sample of retail stores in the program's target area, or may need to observe the installation of equipment installed by program staff to assess the installation process.

Other activities as needed to address researchable issues
While the efforts discussed above can provide a wealth of evaluation information, sometimes other research activities or tools will be needed. For example, to review the adequacy of field efforts associated with the installation of insulation it may be appropriate to acquire infrared thermal imaging equipment to confirm the amount of installation installed or the adequacy of the installation. Likewise, it may be desirable for the evaluation team to hire and train interviewers from the target neighborhood who have pre-established relationships with members of the target group and who can gain access to the target group. The evaluation team may find itself needing to be flexible and creative in identifying approaches for conducting the process evaluation that meets the data collection and analysis needs of the research effort. The process evaluation professional must select the right set of activities to match the researchable issues to be addressed, the available budget and other available resources to the evaluation.

Access to Program Materials, Resources and Personnel

In conducting the process evaluation, it is typically necessary for the evaluation team to have unrestricted access to the program management and staff as well as a wide range of records, reports and materials associated with the program implementation. Program managers need to understand that the process evaluation will consume a small amount of their time and potentially the time of their staff for interviews and discussions and to help locate and acquire the materials that need to be reviewed during the evaluation. Evaluation professionals need to understand that the program management and staff need to incorporate these needs into their daily operational schedules and commitments. It is often necessary for the evaluation team to schedule events with program managers and staff in advance, so that program staff can manage their program responsibilities to incorporate the needs of the evaluation team. Cooperation and support by the evaluation team, the administrative team, and the implementation team are conditions that are a part of every program evaluation effort.

In most cases where information is needed from the program, the evaluation staff can prepare a formal information request identifying the data needed. These requests typically detail the exact data needed and the purpose for requesting the data.

Making Decisions and Selecting Methods: The Process Evaluation Roadmap

This section presents the evaluation decision roadmap, a tool to help determine if a process evaluation is needed for a specific program. These decision procedures are presented as steps in the process evaluation roadmap.

In using the process evaluation roadmap, decision makers can move through each of the twelve steps described in the roadmap. The decision steps in the roadmap can be used to guide the development of the evaluation plan and help it focus on the issues to be evaluated. For example, the first step in the roadmap is associated with a determination if any of the program's designs or operational and delivery characteristics are new to that program. If the program employs new designs or operational practices, the process evaluation can focus on the effectiveness or efficiency of these "new" practices and assess their influence on the program's operations and their consistency with the program theory. Once the roadmap decisions are made the evaluation plan can be structured to focus on the aspects of the program that extend from the roadmap's step-by-step decision process. In some cases, the roadmap's decision process may identify research needs that are greater than what can be conducted in a single process evaluation or fit within the available process evaluation budget. When this occurs, the focus of the evaluation plan will need to be prioritized and perhaps funded over multiple years within the program funding cycle, or delayed until a following cycle if necessary. In other cases the roadmap may suggest that a process evaluation is not needed within the current funding cycle. As indicated earlier, the process evaluation roadmap is constructed to be used to plan the evaluation effort within a single program funding cycle.

The process evaluation roadmap is designed to help program administrators and others make decisions about when a process evaluation can be expected to provide benefits to the program. In moving through the roadmap decision process it is important to keep in mind that program conditions change over time and as program conditions change it may be necessary to re-examine the roadmap and update the process evaluation decisions. For example, if a program is considered to be cost-effective or moving in a cost-effective direction during the early implementation period, the roadmap suggests a process evaluation may not be needed. As a result, the program administrator may not plan or budget for a process evaluation. However, if that program becomes less cost-effective as it moves through the implementation cycle, this decision may need to be reassessed based on the most current program information. The roadmap is designed to be used in an annual reassessment process.

In moving through the process decision roadmap it may be appropriate for a decision associated with a specific step to be overridden or revised by an organization other than

those associated with the administration or implementation of the program. The CPUC staff or their representatives may identify a program condition that justifies the implementation of a process evaluation to deal with a specific issue. In these cases the CPUC staff will notify the program implementer or administrator as appropriate, if conditions relative to their program warrant a process evaluation. In these cases the program administrators should plan a process evaluation to address the specific issue(s) identified by the CPUC staff.

Steps in the Process Evaluation Roadmap

The following figure (Figure 8.1) provides the process evaluation roadmap to help guide readers through the decision-making steps about when process evaluations can provide valuable information for California energy program policy decisions and to identify potential program improvements.

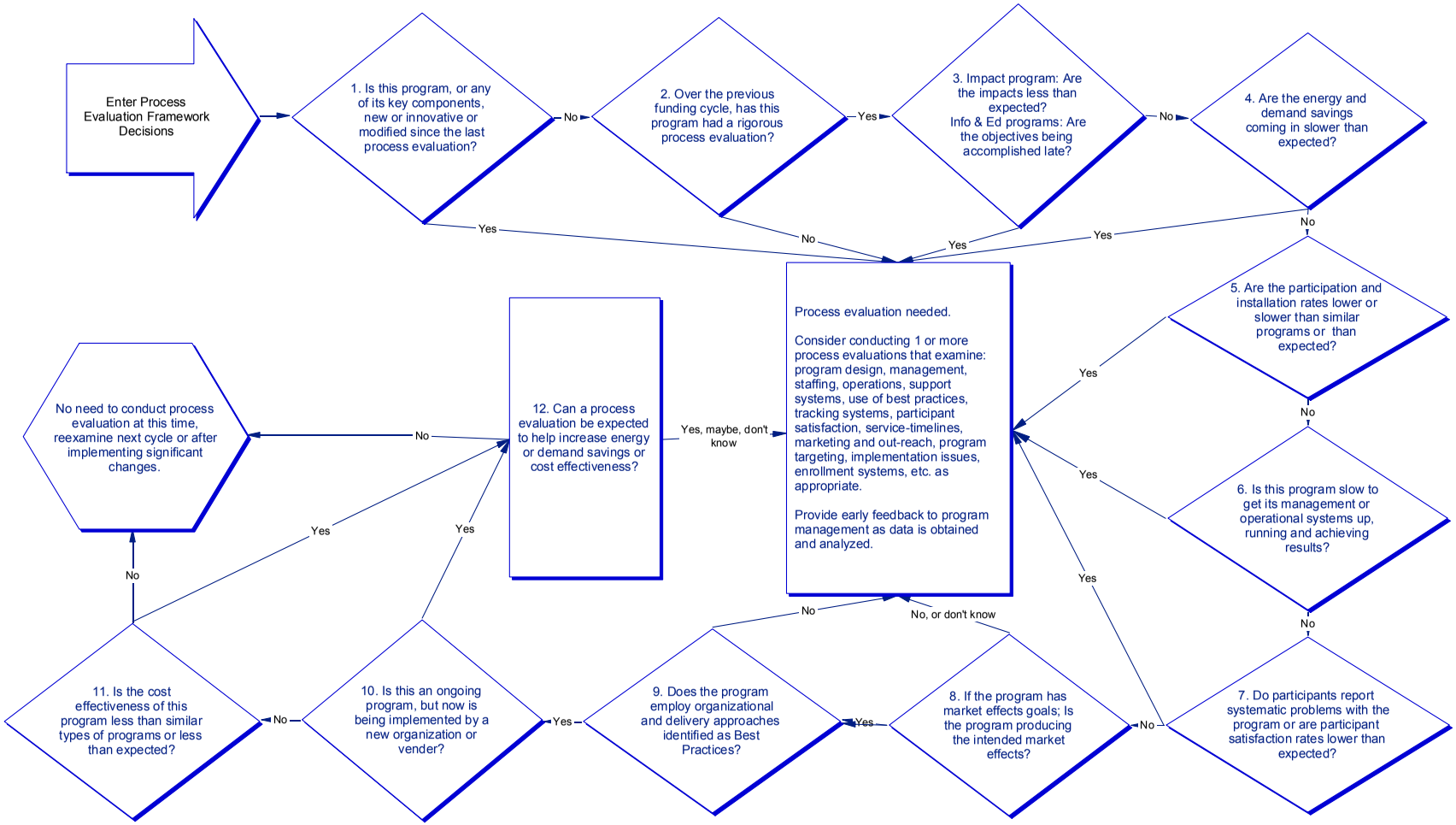

Figure 8.1: Process Evaluation Roadmap Decision Path

The reader will notice that there are a series of decisions along the path. In some cases, the results of a decision lead directly to a suggestion to conduct a process evaluation. In other cases a decision leads to an intermediate decision. This intermediate decision is based on the response to a question concerning whether a process evaluation can be expected to increase energy or demand savings or improve the program's cost-effectiveness.

The following paragraphs present the steps in the process evaluation roadmap.

Step One

Upon entering the process evaluation roadmap, the first step is to determine whether any significant part of the program designs, operational systems, or delivery characteristics are new, are innovative, or represent a change to the program or the program goals since a process evaluation for this program has occurred.

If any key components of the program designs, operational systems or delivery characteristics are new or are considered to be innovative then a process evaluation can be conducted to examine those parts of the program that are considered new or innovative. This evaluation can focus on assessing the effects of those new or innovative characteristics on the operations and expected success of the program and compared to the program theory to determine if the theory and the implementation efforts are consistent. For these reasons the roadmap shows a direct line from the decision concerning the "newness" or "innovativeness" of the program characteristics to the recommendation to conduct a process evaluation. These evaluations help support the informational needs of California policy makers to be able to more fully understand the program characteristics being evaluated and make decisions about their applicability to future programs in California or to make decisions concerning whether those characteristics should be repeated, continued or identified as a best practice.

Step one also focuses on making a decision regarding whether the program has modified or changed any of its key design or delivery approaches pertaining to the management or operations of the program, including management systems, internal or external operations, or service delivery and customer interaction components. If the program has changed any of its design or implementation characteristics since the last process evaluation, then a process evaluation can be conducted to assess those parts of the program that have changed. This evaluation can focus on assessing the effects of those changes on the operations and expected success of the program and compared with the program theory.

If the response to step 1 is "no" there is nothing innovative about this program and "no" key component of the program has been modified since the last process evaluation, then the decision process moves on to step two.

Step Two

In Step two the roadmap looks at how long it has been since the program has had a rigorous process evaluation. If the program has not had a process evaluation of its

designs, operational systems or delivery characteristics during the previous funding cycle, then a process evaluation is suggested for the current cycle. This evaluation would examine the program's key design characteristics, its methods of operation and its program delivery systems, including its management, marketing, product mix and customer interaction processes.

If the response to step two is "yes" the program has had a rigorous process evaluation during the previous funding cycle then the decision process moves on to step 3.

Step Three

For programs with impact goals, this step seeks a decision to determine if the savings (energy or demand) are less than similar types of programs operating in California, per dollar of program costs, or if the energy savings are less than what is expected in comparison to the projected savings presented in the program implementation plans, the program theory, or in the program contract agreements. For programs that have had a previous net impact evaluation conducted on their program (or a very similar type of program) this decision can stem from a review of the net impact evaluation results compared to the program's progress relative to the net impact goals. If the program does not have previous net impact evaluation results, then assessing the projected net impacts resulting from the program's implementation progress can inform this decision. If the expected impacts from installed measures do not match the implementation plans of the program, then the program can implement a process evaluation to determine why impacts are lower or slower than expected. This can be especially important if the reason for low net effects is high free rider levels (for those that have estimated or measured free rider levels). In this case the process evaluation can look at what components of the program are causing the high free rider levels. In making the determination concerning whether the program is achieving less savings than expected, the decision maker should also review evaluation results from similar programs conducted in California and determine if their expected savings are lower, given their differing circumstances. We realize that this will generally be a subjective assessment, especially in the early phases of a program, and decision makers must use their best judgment on using the information and how to interpret it, taking into account the conditions of the programs on which the comparisons are being made (i.e. technology differences, participant differences, weather differences, etc.). In some cases there may not be similar programs with which to draw a comparison, or there may not be impact information available on the programs being compared. This step asks for a best attempt to determine if the program is providing lower savings than what is or should be expected from the program at that point in the cycle. If the assessment indicates that the program is providing savings greater than or equal to that expected and/or to similar types of programs at similar points in their cycles, then the decision process moves to step four. If the assessment indicates that the program is providing savings that are less than expected or less than similar types of program (taking appropriate differences into consideration), then the program can undertake a process evaluation to address the reasons for the lower than expected savings and to develop recommendations for changes to the program that can be expected to improve the program's savings or cost-effectiveness.

For programs without an impact goal, such as information and education programs, this step asks if program objectives are being achieved consistent with the program implementation plans. For programs that have difficulty accomplishing their objectives, a process evaluation can be conducted to identify the reasons why objectives are not being met and what program modification are needed to help bring the program back on schedule.

Step Four

While step three focused on lower than expected savings, this step assesses whether the energy savings (energy or demand) are being achieved at a slower rate than projected. If the project savings levels are coming in too slow and are not meeting expectations for the program then the program should consider a process evaluation to assess the reasons for the slower than expected savings and identify recommendations that can be expected to speed the rate of achieved savings.

Step Five

This step is similar to and in many cases related to the decision associated with step three and four. In this step the decision maker is asked to make an objective decision about the participation and measure installation rates for their program. This assessment should include a review of the program's projected accomplishments with regards to the number of participants served by the program and the number and type of measures installed through the program, and determine if these rates are different than what is needed to accomplish the stated goals in the program implementation plans or in the implementation contracts. If the participation and implementation rates are greater than or equal to what is needed to reach their projected accomplishments with regard to participants and installed measures, then a process evaluation to assess these conditions is not needed and the decision makers move on to step six. However, if the program participation rates or the measure installation rates are lower than projected, then a process evaluation can be conducted to identify the reasons for the slower than expected participation or measure installation rates. This process evaluation should provide recommendations that can be expected to increase the participation or measure installation rates for the program.

Step Six

Step six involves a decision regarding the ability of the program to get its management or operational systems up, running and achieving results in a reasonable period of time. This again is a subjective assessment. However, programs that are slow to move from an approved program plan to an operational program, with management and operational systems in place and functioning in a way that allows the program to produce timely results, can consider conducting a process evaluation. This evaluation can focus on the reasons for the slow operations, and focus on providing recommendations for speeding up the development of the program's management and operational systems. Programs that have their management and operational system up and running and supporting the program in time for the program to function according to the program implementation plans do not need to conduct a process evaluation for this issue.

Step Seven

Step seven is placed in the roadmap to trigger a process evaluation if there are problems from the participant's perspective. These include problems with the program designs, operations or services, which lead to excessive participant complaints or reduced levels of participant satisfaction. While all energy programs can expect minor customer complaints from time to time, and can, for a very small fraction of participants, expect satisfaction scores that might be lower than desired, this step asks if there are systematic problems with the program from the participant's perspective that produce higher than desired complaints or lower than expected satisfaction levels. This again is a subjective assessment. However, if the program is having any significant level of participant interaction problems that stem from the design of the program, the methods of operation or the services received, a process evaluation assessing these problems and the reasons for these problems should be seriously considered. The evaluation should focus on providing program design changes, changes to operational systems, or changes to the customer interaction components that can be expected to reduce complaint levels and increase levels of satisfaction. Complaints or low satisfaction levels from one or two percent of customers may be considered reasonable; however, complaints or low satisfaction levels from five percent or more of the participants should be considered excessive. The goal of the CPUC is to provide energy-resource producing programs that have high customer satisfaction rates, low complaint levels and cost-effective impacts. This step supports that goal.

Step Eight

Step eight is the final step in the roadmap that leads directly from a program associated decision directly to a recommendation to conduct a process evaluation. This step is for programs that have market effects or market transformation goals in their program implementation plans or in their implementation contracts. These programs can conduct market effects (transformation) evaluations via the Framework as indicated in Chapter 10 of this document. If this evaluation indicates that the market effects projected in the implementation plans or the program is achieving program contracts, the program does not need to consider a process evaluation to address market effects issues. However, if the evaluation indicates that the program goals pertaining to market effects are not being achieved, a process evaluation that works with a market transformation or market effects evaluation can be conducted to address the reasons for the low effects and provide recommendations for how the program can be improved, if possible, to increase the effects in the market.

Steps nine, ten, and eleven incorporate an additional decision in the sequence before a determination can be made regarding a recommendation to conduct a process evaluation. The added step in the decision process is a subjective determination regarding whether the program can be expected to increase energy or demand savings or increase the cost-effectiveness of the program as a result of conducting a process evaluation (see step twelve).

Step Nine

In Step nine, a decision is made regarding whether the program employs appropriate "best practices" identified in the 2004 National Best Practices Study[180] pertaining to the design and operations of Public Good Charge programs, or in the services or service delivery approaches used by these programs. The CPUC desires that all energy efficiency programs use best practice approaches that improve program operations or service delivery whenever possible and appropriate (cost-effective). This decision rests on the completion of the 2004 National Best Practices Study (or other best practice studies) and the identification of approaches and designs that can be considered best practices. This study will provide a presentation of best practices as a resource to program designers, managers, administrators and contractors. This step in the roadmap decision process asks programs to assess these best practices as to what might be applicable to their program and whether incorporating additional best practices might help their program increase savings or make it more cost-effective.

If the program is not using the best practices identified for their type of program, then a process evaluation to examine the potential for use of the applicable best practice is suggested. This process evaluation can examine the current program designs, operations and customer interaction systems and make recommendations for the inclusion of the appropriate best practices. Again, this is a subjective decision process that may be overridden or modified by the CPUC if the CPUC determines that a process evaluation to examine the use of best practices is needed.

Step Ten

In step ten, the decision is based upon whether a program is being implemented by an administrator or contractor who has not implemented this type of program in the past 5 years within the state of California. If this is the case, then the process evaluation decision moves to the next level (step eleven) to decide if a process evaluation can be expected to increase savings or cost-effectiveness. Programs that are provided by administrators who may not be familiar with California's special conditions (weather, markets, customer opinions and needs, consumer protection laws, past programs and program operations, etc.) may be at a disadvantage compared to others who are established and experienced with California's energy programs and markets. Organizations and vendors who have not provided the type of program they currently offer need to assess their operations and determine if they think their program can benefit from a process evaluation that examines their structure, operations, services and service delivery efforts to identify ways that the program can be changed to be more effective at achieving energy impacts in the California environment.

Step Eleven

This step involves a decision associated with the program's cost-effectiveness. If the program is operating at lower than expected levels of cost-effectiveness, then a process evaluation to examine the reasons for this performance is suggested. While most programs experience start-up and early rollout conditions that act to lower a program's cost-effectiveness ranking, these conditions should be overcome early in the program

[180] *National Best Practices Study.* (Quantum Consulting Inc. 2004).

cycle. If program design, operational or market condition are not allowing the program to operate at their planned levels of energy impacts per dollar spent, then a process evaluation to address these issues should be considered. This decision involves a determination of whether the current program cost-effectiveness is lower than what is described in the program implementation plan or the implementation contract for the current period of the program. If during the course of the program's implementation efforts the cost-effectiveness falls below what should be expected from the program, then a decision should be made concerning whether a process evaluation can be expected to help increase energy or demand savings or help increase the program's cost-effectiveness.

Step Twelve

Step twelve is the final decision step in the process evaluation roadmap. In this step the administrator makes a determination (based on their expert knowledge, the projected net impacts of similar programs included in the CPUC's program portfolios, the portfolio of progress reports available to program administrators from the CPUC, and the program evaluations conducted in California in the past five years that are available on the CALMAC website), if they think that their program can benefit from a process evaluation. If the administrator does not think that a process evaluation will help increase their program's energy impact performance or their cost-effectiveness, then a process evaluation is not recommended. However, if the administrator determines that a process evaluation may help achieve greater energy impact or improve the cost-effectiveness of their programs, or if they are not sure or don't know if a process evaluation would help, then a process evaluation that focuses on assessing program designs, operational practices, technologies, and market and customer interaction practices, to determine if the program can improve its operational practices or delivery efforts is recommended. Again, the decision process in this case is being made by the administrator of the program, but may be subject to revision by the CPUC.

Conducting the Process Evaluation

If anyone of the above detailed decision steps identify the need for a process evaluation, it is recommended that the program budget and plan for the implementation of the needed evaluation by an independent program evaluator that meets the skills requirements presented in this chapter.

Process Evaluation Budgets

Typically process evaluations range in cost from a low of around $10,000 to $15,000 for small evaluations of limited scope, to an average cost of around $30,000 to $60,000 for evaluations focusing on multiple issues with some level of supportive field data collection and detailed assessment efforts. However, it is not unusual for process evaluations that focus on a wider range of issues, involving customer surveys, interviews or focus groups with on-site examinations to run in excess of $60,000. For these reasons process evaluations should be scheduled into the program evaluation planning and budgeting process along with other evaluation efforts. For programs that are funded over

multiple year cycles, the process evaluation may need to be structured over the course of the program to minimize the financial impact of the evaluation effort in a given year. However, process evaluations should typically be conducted early enough in the program cycle to be able to use the evaluation results to improve the program.

In establishing the evaluation budget many programs will need to prioritize the process evaluation issues on which they should focus their process evaluation dollars. It is up to the program administrator, (perhaps in consultation with their evaluation contractor) to prioritize the process evaluation issues to be addressed within their process evaluation budget. However, the CPUC requires that the evaluation plans and budgets be approved by the CPUC prior to their implementation. This process will include a review of the program evaluation plans and budgets and negotiations concerning the scope and focus of these plans. The CPUC may instruct the administrator to change their evaluation plans to meet the evaluation goals, objectives and priorities of the CPUC.

Chapter 9: Information and Education Program Evaluation

Preface

This chapter of the Framework focuses on energy programs that provide information or education services to their target markets and addresses how these programs can be evaluated. A broad definition of information and education services is used to define the types of programs and efforts for which this chapter is applicable, including advertising, public service announcements, education efforts, training activities, outreach efforts, demonstration projects, and other information or communication-based efforts. These programs may be targeted to either end use customers or to other market actors whose activities influence the energy-related choices of end use customers. Market actors that influence end use customer choices include installation and repair contractors, retailer staffs, architects, design engineers, equipment distributors, and manufacturers and others.

Typically, California information and education programs have one or more of the following general goals:

1. Educate energy consumers regarding ways to increase the energy efficiency of their buildings and activities in order to convince them to take actions that help them manage their consumption or adopt more energy efficient practices.
2. Inform energy consumers and/or other market actors about program participation opportunities in order to increase enrollment in these programs.
3. Inform energy consumers and/or other market actors about energy issues, behaviors or products in an effort to transform the normal operations of the market, so that the targeted market actors make more energy efficient operational, purchase or behavior-oriented decisions *without* direct program interventions or incentives.

This chapter of the Framework focuses on the evaluation needs of California's information and education programs and the evaluation decisions associated with these programs. The information presented in this chapter is especially important to information and education program administrators, and to evaluation managers and their staff who conduct these types of evaluations, but also to individuals responsible for the evaluation planning budgeting and oversight functions. This chapter is also suggested reading for energy regulatory and policy managers and portfolio managers. The chapter is also written to provide a level of subject familiarity to individuals responsible for cost-effectiveness assessments and other evaluation professionals who may need to support or inform an evaluation of information and/or education program effects.

Introduction and Key Issues

It can be argued that almost every energy efficiency program provides some level of educational and/or informational content. However, if the program has been created primarily as a conduit that leads participants into other programs or services, or it provides training and education on energy efficiency options to customers and other market actors, then the program should not be expected to meet the same cost-effectiveness requirements as programs that are offered expressly as a way of acquiring energy resources. That is, these programs may not be expected to immediately produce cost-effective energy resources, if they are not designed with that intent in mind. Rather, information and education programs are typically designed to acquire energy or peak savings indirectly through changes in behavior after exposure to the information, over time (market transformation), or via increased enrollments in other resource acquisition programs. As a result, information and education programs are generally not expected to pursue the Framework's energy impact or market transformation evaluation paths. Instead, a unique information and education effects path is provided in the Framework for these programs. This path is an alternative to conducting impact or market effects evaluations, and is in addition to the appropriate process evaluations suggested for these programs.

As a result, information and education programs are expected to have an energy impact evaluation instead of an information and education effects evaluation only if they have energy acquisition goals on which their cost-effectiveness is based. Likewise, programs do not need an information and education effects evaluation if they have market effects goals and if their cost-effectiveness is based on achieved market effects. This treatment allows the information and education programs to have no more than two types of evaluations per program funding cycle depending on the specific objectives of the program.

The Framework structure is focused on individual programs. While following the information and education evaluation roadmap (decision path) will generally mean that only two types of evaluations will be conducted on each program, it does not preclude the conducting of additional evaluations on these programs or on groups of information or educational programs. There may be cases in which additional evaluation efforts are needed to understand the effects of these programs. For example, while individual information or education programs may not have program-specific market effects, these programs examined in clusters may have substantial market effects or energy impacts. It may be desirable from occasionally to examine the energy or market effects of these programs as a group in order to better understand their combined effects on energy or on their target markets. In these cases, additional evaluations may need to be conducted as overarching studies. A discussion of these studies is presented in Chapter 15 of the Framework (Overarching Evaluation Studies).

Evaluation efforts for information and education programs need to focus their evaluation efforts on the activities that are appropriate for their specific type of program and program goals. For most information and education programs, there are two primary

evaluation paths that are suggested in the Framework. These paths are summarized in Figure 9.1 and are covered in detail within the relevant chapters of the Framework that deal specifically with each of the types of evaluation efforts suggested.

The following diagram indicates that information and education programs can be evaluated as appropriate for the specific characteristics of the program being considered.

The CPUC advises that programs should prioritize their evaluation efforts so that effects evaluations have the highest evaluation priority followed by process evaluations, and that energy impact evaluations and their associated M&V efforts are more important than market effects or information and education effects evaluations for programs with energy impact goals. This is a departure from the previous California evaluation practice for information and education programs and is presented here to bring attention to the need to consider both program effects and process evaluations for these programs. The suggested evaluation efforts are summarized below in Figure 9.1.

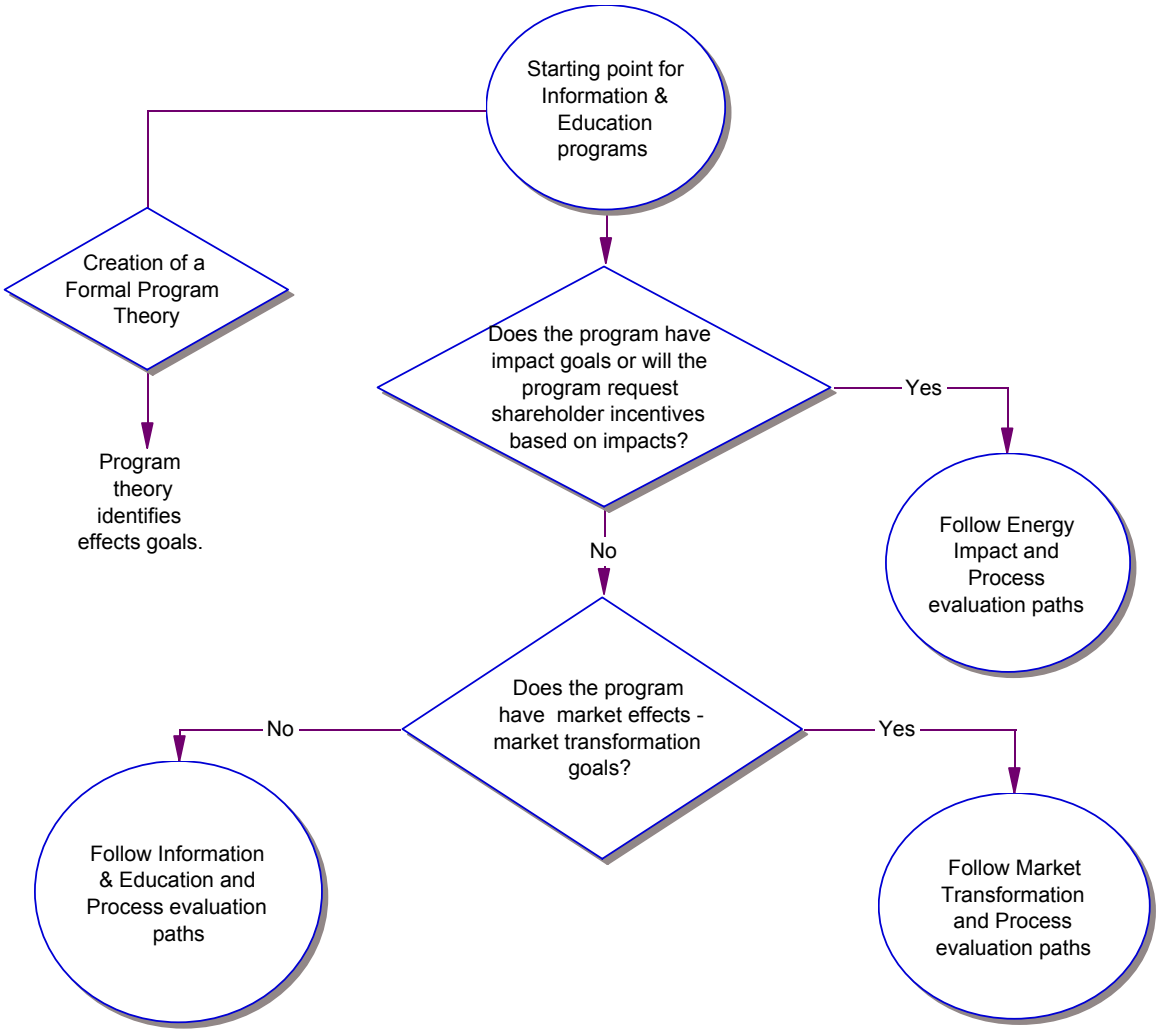

Figure 9.2: Overview of Information and Education Evaluation Efforts

The four possible evaluation paths for information / education program are described below.

a. *Process Evaluation*. All information and education programs should consider following the process evaluation path presented in the Process Evaluation chapter of this Framework (Chapter 8). These evaluations help improve program design and implementation processes in order to improve their effectiveness or operational efficiencies. These evaluations also help document the operational processes that are employed so that they can be considered for use by other programs or as a best practice. All information and education programs should include a periodic process evaluation.

b. *Market Transformation Evaluation*. Programs that have a goal of transforming a market or portion of a market or of changing how a market operates should follow the Market Transformation evaluation path presented in Chapter 10. These evaluations help document the amount of market change that has occurred as a result of the program's efforts and the effects of these changes on California's

energy supplies. Programs with market transformation goals as their primary information and educational goals should consider conducting a market effects evaluation consistent with the market transformation evaluation roadmap.

c. *Impact Evaluation.* California policy makers need to know how effective specific information and education programs are at increasing enrollments into the impact programs. Programs that are designed to refer energy consumers to other programs having resource acquisition or procurement goals need to coordinate their evaluation activities with the evaluations of the programs to which they are referring customers. It is important to assess the degree of success that information / education programs have in increasing customer awareness, enrollment and participation in the targeted resource acquisition programs. This type of evaluation typically takes the structure of including questions in the evaluation surveys and interviews of the resource acquisition program evaluations to determine the different ways participants hear about the program, which information sources helped guide the participant into the program enrollment process and questions that allow the evaluator to determine the impact of the different information sources on the decision to enroll in the program. This activity usually involves adding researchable issues to the impact evaluation plans of the resource acquisition programs that deal with the success of the information or educational program's ability to influence customers to participate in the impact program.

d. *Information/Educational Effects Evaluation*. If the program's goal is to lead customers or targeted market actors to take actions that reduce energy use (in the short-term or in the longer term), and is classified as an information or education program (rather than energy impact, resource procurement, or market transformation) then the administrators should consider following the information and education effects program evaluation path described in the remaining sections of this chapter. The primary focus of this evaluation is to document the degree to which the program is achieving its desired effects within the markets targeted by the program.

Skills Required for Information/Educational Evaluations

Evaluations of information/education programs should be conducted by individuals or organizations that understand the relationships between providing information or training to different types of energy consumers and other market actors and the effects of that information on the behaviors of the individuals exposed to these services. Information and education program evaluators should be trained and experienced in conducting social science research specifically in the field of energy efficiency programs. Evaluators need a strong understanding of the cause and effect relationship between exposure to energy technology and behavior information and possible actions, and of the ways to test the causal processes involved.[181] Evaluators should also have a strong understanding of the

[181] A study that provides an introduction to various methods of testing causal relationships, and provides an analysis of methodological lessons learned after analyzing one of the country's oldest continuously

application of experimental and quasi-experimental research designs on human populations and be skilled at establishing evaluation approaches to test for net program effects.

Identifying the Issues

Information and education programs play a crucial role in the promotion of energy efficiency practices and measures in California. A significant number of programs fall under this classification, and information/education activities are significant parts of several resource acquisition programs. Consequently, documenting the effects of information and education programs and activities is an important component of the Framework. These evaluations help identify which information and educations programs are meeting their goals and the degree to which those goals are met. It will also identify those programs that excel in the accomplishment of their goals. The primary purposes of the information and education program evaluation efforts are to:

a. Provide information on the effects and effectiveness of the programs in motivating customers to either take efficiency actions on their own or to increase the possibility these customers will access or participate in other efficiency programs in order to determine which information and education programs are helping to meet California's current energy supply needs or the energy supply needs of the future;

b. Identify programs that need to be modified or improved to be more effective;

c. Help identify best practices in the energy information and education program practice so that these practices can be documented, shared and replicated;

d. Provide a system for documenting the accomplishments and the benefits received from spending Public Good Charge or energy procurement funds;

e. Help policy makers and resource planners determine which program services to fund to help acquire future energy resources;

f. Help resource supply planners identify a mix of energy resources that can be cost-effectively acquired to meet the energy needs of California's energy consumers.

Together the above reasons express the importance of and reasons for evaluating California's information and education programs.

Making Decisions and Selecting Methods

Once a program has been identified as an information or education program designed to cause people to take actions as a result of exposure to the program, the evaluation team needs to review the program plans, and particularly the detailed program theory

---

operating energy efficiency programs is "How Can We Tell if Free Information is Really Transforming Our Market?" (Conlon et al. 1999).

associated with the program to identify the potential causal relationships between exposure to the program efforts and actions taken as a result. The program theory should describe in detail how the program promotes, changes or influences energy efficiency decisions. (See program theory discussion in Chapter 4: Evaluation Overview and Issues) Typically this relationship is expressed through a formal program theory that may be linked to an implementation theory, a market operations theory or to a program logic model.[182] Through the program theory process, program providers should describe how their program provides customers with information, education and/or is being used to drive customers to other energy efficiency offerings or to take actions. Evaluators should then establish their evaluation approaches to test these relationships and effects to document and quantify actual effects and to confirm theoretical causal relationships. However, evaluators should not rely solely on the program's version of the program theory or the accompanying logic model to plan their evaluation efforts. Typically evaluators can identify other potential alternative causal relationships not reflected in the program theory or logic models for why a specific program effect may or may not be achieved. Evaluators should use the program theories and logic models as guides in the evaluation planning process rather than focus their evaluation efforts only on those relationships presented in the program theories and logic models.

Topics covered in the program theory should include:

- the educational or information subjects on which the program will focus and the efforts and activities to be undertaken;
- the specific education or information transfer methods and mechanisms that will be employed in the implementation process (audits, workshops, training classes, announcements, demonstrations, ads, etc.);
- the target market sector or sectors, including, as appropriate, market segments or sub-segments and the geographical market areas the program is designed to reach;
- awareness, understanding or knowledge of goals for target markets;
- the expected effects in terms of what the recipient is expected to do or accomplish as a result of the information or education efforts;
- the timeframes in which the expected results are to be accomplished;
- the barriers that the information or education must overcome to be successful; and
- the educational goals they are planning to meet within their program's market in terms of end effects.

[182] One of the first uses of a logic model in the energy program field was in an evaluation of a children's educational project, "Detecting Behavioral Change from a Visit to a Children's Museum Energy Conservation Exhibit." This study found the program theory invaluable for structuring the evaluation approach. Evaluators were able to use the model for structuring the evaluation, making it possible to find the small changes made in energy consumption due to a children's museum exhibit. (* Peters, McRae et al. 2000).

A baseline study should also be conducted in order to understand initial energy efficiency issues and to establish the baseline from which program effects can be assessed. Issues to address in the baseline studies include:

- size and composition of target markets;
- pre-program awareness or knowledge levels;
- pre-program information and education sources;
- extent of exposure to and use of pre-program education or information sources;
- pre-program status of the target market relative to the intended results of the program; and
- relative pre-program adoption and behavior patterns.

The baseline study does not need to be program specific, but it should provide evaluation planners with enough detail to assess program-specific attributes by desired segments of the program population (i.e. income level strata, hard-to-reach population characteristics, geographic location, etc.). That is, the baseline should be well-understood so that the incremental change in the actions or behaviors of the target market can be estimated using accepted evaluation approaches for the specific target groups and sectors on which the program's efforts are targeted.

The timing of the evaluation effort is a critical consideration in structuring the information and educational program evaluation effort. To document the effects of the program, in terms of actions taken by customers or other market actors exposed to the program's efforts, the evaluation will need to wait until the program has had the time to accomplish their goals to a level that can be measured. This may mean that the evaluation effort may need to be delayed until the program theory indicates that the targets of the program's efforts have had enough time to accomplish the intended results.[183] It is not unusual for an evaluation to be delayed for six months to a year or more. However, in some cases it is advisable to establish a time series evaluation approach in which the program's effects are examined periodically to identify the rate of change of the intended effect in the target population. A time series approach can be useful for documenting the rate of change, and if necessary help identify improvement opportunities that can be addressed in order to increase the rate of change.

Programs that inform market actors in a way that is expected to cause them to take specific actions will need to identify the target markets for their efforts in a way that permits evaluators to contact members of that market and assess the program's effects.[184]

---

[183] This concept was applied in "Cinergy's Home Energy House Call (HEHC) Program: An Information Program That Changes People's Lives. The HEHC Program was evaluated from six months to two years after participation to determine how many households were implementing energy efficiency measures after an expanded audit of their home. (* Riggert et al. 1999).

[184] An evaluation of a targeted information and education campaign is discussed in "Transforming Markets Through Education and Information: A Study of the Pacific Energy Center.," This study examines the effects of the Pacific Energy Center's campaign to transform the commercial building market to make it more energy efficient. (* Reed et al. 1999). Full PG&E report: (Reed et al. 1998).

Broader consumer-based message efforts can in some cases be difficult or costly to evaluate due to the need to find those with recall. However, energy information messages broadcast over television or radio stations as part of a relatively large campaign can expect to be successful at finding and testing consumers for recall. These customers can be assessed in order to evaluate their reactions and behaviors to the message by using random-digit dialing and a sample-quota system for those with recall.[185]

Smaller general advertising/information efforts, on the other hand, may find the cost to document the effects of the program to be prohibitive. However, policy makers should also understand that if the cost to identify and contact an individual exposed to the program's message is prohibitive, the message dilution factor (percent of a target market exposed to a message) associated with that program may be so low as to cause the program to have no significant measurable effects on the target population. For a broadcast message, an evaluator may need to contact hundreds of potential readers or listeners to find a few who were exposed to the message and who remember it to the extent that they can describe the influence of the message on their behaviors. It may not be practical to spend thousands of dollars in data collection costs to acquire program effects information from such a small group of individuals; however, it may be very practicable to contact potential viewers or readers in a specific market to identify the percent of the targeted population who recall the program's efforts. If recall is low, then policy makers need to determine if the recall rate is worth the program expenses.[186]

To help identify the population reached by an information or education program designed to influence behavior, the program implementers should collect exposure contact information on the people they reach whenever possible.[187] For example, if residential or commercial energy audits are provided, then contact information on the recipient of that service should be maintained and made available to the evaluation staff. Likewise if the program conducts door-to-door efforts within targeted neighborhoods or communities, the program should maintain detailed information on the specific program areas targeted so that individuals residing in the target areas can be identified via city directories, geo-coding capabilities, or field efforts that sample in the target area.[188] In many cases it will

---

[185] An evaluation of a large advertising effort was able to conduct this type of evaluation. This evaluation was particularly unusual in that the advertising effort was a concentrated three-month campaign to induce load-shifting behavior in the summer (among other energy saving tips provided). The 2002 evaluation included random pre and post-surveys immediately surrounding the campaign. The surveys were designed to gather input to actually calculate peak load reduction. The results were used within a method to estimate peak load reduction and large load-shifting impacts from the advertising effort were found. "Quantifying Load-Shifting Benefits from an Advertising Campaign." (* Engle et al. 2003).

[186] "Evaluating the Impacts of Education/ Outreach Programs: Lessons on Impacts, Methods, and Optimal Education," provides a comprehensive overview of various types of advertising campaigns and the evaluation approach to identify the exposure and recall rates. (* Green et al. 2000).

[187] An evaluation of a low-cost residential DSM program found that on-site visits were more likely to encourage the replacement of existing appliances with more energy efficient models. A complete contact database made this evaluation possible. See "Evaluating Educational Effects in Pacific Gas and Electric's Energy Savings Plan." (* O'Meara et al. 1994).

[188] One study of an educational campaign done by Puget Sound Power & Light and an advertising firm utilized mass media and direct marketing in a campaign to push energy conservation. The evaluation

be possible to collect recipient contact data if the program is initially designed to do so. For example, if the program distributes energy efficiency information at a public event via a hand-out information packet, participants who receive the handouts can be offered an incentive to provide their contact information in order to participate in a free drawing for a desired incentive (for example a drawing of $100 per hour over a 10 hour distribution period) in exchange for agreeing to be contacted by the evaluation staff. However, if an incentive to collect the information is used, designers need to minimize the influence of the incentive as a filter for the types of people who may provide their contact information. The ability to identify and contact individuals exposed to the program's efforts is often a critical component in the ability to evaluate an information or education program.

In some cases it may not be possible to collect contact information at the time the program services are provided. In other cases it may not be affordable to conduct the appropriate screening efforts within the evaluation process. In other cases, the cost of a formal evaluation may simply be too high for the size of the program being evaluated. In these cases, whenever possible, program evaluation efforts should be incorporated into the program delivery services themselves. One example of this approach is an evaluation questionnaire administered at the time of the exposure to the program for all or a sample of the market receiving the program service. In these cases, the evaluation effort should obtain program operations and delivery feedback information, but also, to the extent possible, obtain program effects or anticipated program effects information. In addition, there should be an attempt, as part of the real-time evaluation efforts, to collect participant contact information for follow-up evaluation contacts.

Primary consideration in determining whether an information and education evaluation needs to be conducted include:

- the ability to conduct the evaluation,
- the value of the information to feed long-term energy supply decisions,
- the cost of the evaluation for that specific program cycle, and
- the anticipated value of the evaluation results toward longer-term program selection choices.

It may not be desirable to allow continued funding for program efforts that cannot document energy or demand savings nor information or educational effects through their evaluations.

Program Evaluation Design, Tools and Examples

The purpose of this section is to aid in the selection and design of the evaluation strategy for information and education programs. Evaluators should examine the following list of evaluation topics for potential inclusion in their evaluation plan and identify evaluation strategies that allow for collecting and analyzing the data required for these topics.

---

shows unintended benefits, and successful results, partially due to the known geographical area of the campaign. See "Conservation Advertising Campaigns and Advertising Effectiveness Research: The Right Combination to Solidify the Conservation Ethic." (* Auch et al. 1994).

Examples of information and education Program evaluation topics:

- Number and percent of customers reached, or made aware;
- Number and percent of customers reached that take recommended actions;
- Number and type of actions taken as a result of the program;
- Changes in awareness or knowledge by topic or subject area, by type of customer targeted;
- Customer perception of the value of the information and/or education received;
- Elapsed time between information exposure and action(s) taken by type of customer targeted;
- Attribution of cause for actions taken when multiple causes may be associated with the actions taken;
- Demographic, firmographics or psychographic information as appropriate;
- Influence of program on dealers, contractors and trade allies;
- Effects of program on manufacturers and distributors; and
- Identification of barriers experienced by program, and the development of recommendations for addressing those barriers.

The program administrator will want to incorporate as many of these research objectives in their evaluation plans as practical for the specific program(s) being evaluated and the available evaluation resources.

For each program, a detailed evaluation plan should be developed that incorporates the appropriate elements from the research topics listed above. The following types of data collection methods and tools may be appropriate for these evaluations:

- Surveys: telephone, mail, Internet, e-mail;
- Interviews: in-depth, small group, individual;
- Focus groups;
- Site visits, verification visits;
- Pre & post testing to gauge effectiveness of training initiatives;
- Follow-up efforts (e.g. six months after participation) to assess how training or information has influenced energy efficiency choices, actions or behaviors; and
- The use of comparison groups so that net effects of the program can be estimated.

Table 9.1 lists several types of information and/or education programs and an example of an evaluation issue that can be addressed when conducting an evaluation of these programs. An example of the various tools and strategies that can be used to address each of these issues are also presented in this table. This table presents a single

evaluation issue for each type of program, not to suggest that the evaluation effort should focus on only one issue, but to illustrate an evaluation approach for addressing that issue. The evaluation approach that can be applied to a specific evaluation goal should be coordinated with an information and education program evaluation expert to design an evaluation approach appropriate for the each program.[189]

[189] An example of an education program evaluation specifically designed to work with evaluating "educational unit" programs can be found in "The Enduring Effects of an Elementary School Energy Education Program." This evaluation of an elementary school program provided evidence of long-term and cumulative impacts. The program used participant/non-participant comparison surveys and other information to provide an estimated value for increased conservation behavior by students. (* Hanson et al. 1995).

Table 9.1: Information and Educational Evaluation Methods and Tools

Type of Program	Issue to be Evaluated	Evaluation Tools/Methods
Residential Energy Audit Program	Actions taken as a result of the program audit	Survey of program participants compared to a comparison group of non-participants.
Residential Energy Audit Program	Level of increased education or knowledge	Post-program knowledge survey battery compared to a matched comparison group of non-participants, or pre and post program questionnaires to measure differences compared to a comparison group.
School-based Student Energy Education Program	Level of increased education or knowledge	Teacher-administered tests/surveys of participants and non-participants, or pre and post testing of participants and non-participants.
School-based Student Energy Education Program	Actions taken as a result of the training	Teacher-administered tests/surveys of participants and non-participants
Energy Star Promotional Program	Increase in ENERGY STAR® appliance sales due to advertising	Compare ENERGY STAR® appliance knowledge and sales to similar cities and/or stores in which the program is not provided.
Energy Efficiency Program / Services General Advertising Campaign	Increase in awareness levels and participation in energy efficiency programs and service offerings	Surveys of matched populations exposed and unexposed to the campaign or the same population before and after the campaign. Comparisons of information sources and program enrollment causes between areas exposed and not exposed to the campaign.

Training Programs

Programs that provide training to their participants should consider incorporating the following types of research objectives into their evaluation plans:

- Pre-program level of knowledge to compare with post-program levels,
- The specific knowledge gained through the program,
- Relevance and usefulness of the training as it relates to their specific needs and opportunities to use the information,
- Future opportunities and plans for incorporating the knowledge gained into actions or behaviors that provide energy impacts,
- Whether participants would recommend the training to a friend or colleague, and

- Participant recommendations for improving the program.

Tools used to evaluate training programs in general may include on-site pre/post testing of participants and non-participants, typically involving surveys or questionnaires that may be coordinated with participant satisfaction questions to understand the value of the training and how the knowledge gained during the training will be used. Follow-up post-program surveys of participants are also useful in order to determine knowledge retention issues and applicability or use of training.[190]

Programs with large training efforts or ones that are designed solely for training should have evaluation designs that are mindful of the rich literature and methods on evaluating training programs that are available from the larger evaluation community.[191] An in-depth training evaluation might attempt to measure, report, and make recommendations on how the training meets its goals and objectives and how it fits into the overall goals of the program.

Kirkpatrick developed one of the best-known evaluation frameworks for classifying impacts of training programs. He suggested four levels of evaluation for training programs.[192] Kirkpatrick suggests that the impact/influence of the training can be measured, assessed, and examined for potential improvements across four levels of scope, including:

1. Reaction level - The satisfaction of the trainees (see additional information on assessing levels of satisfaction in Chapter 8: Process Evaluation);
2. Learning level - The level of knowledge gained (measured by comparing pre and post training test results);
3. Work Behavior level - How the knowledge has been transferred into workplace behavior changes; and
4. Organizational level - Where knowledge gained has been used by the trainee to influence organizational behavior changes.

These principles were successfully incorporated into an internal preliminary energy efficiency evaluation planning effort for the EPA's ENERGY STAR® Building benchmark training effort.

---

190 "The Enduring Effects of an Elementary School Energy Education Program" is an example of how evaluation tools and methods are used in an evaluation to determine the measurable effects of education on students' energy-related knowledge, attitudes and conservation practices. (* Hanson and Siegal 1995).

191 See: *Handbook of Training Evaluation and Measurement Methods*. (Phillips 1997); *How to Measure Training Results: A Practical Guide to Tracking the Six Key Indicators*. (Phillips et al. 2002); and *The Bottomline on ROI: Basics, Benefits, & Barriers to Measuring Training & Performance Improvement*. (Phillips 2002). Additional papers and information is available from the American Society for Training and Development.

192 Originally published as four articles in the *Journal for the American Society for Training Directors*, then revised. (Kirkpatrick 1996).

There have also been several alternatives and revision recommended to Kirkpatrick's initial four-level approach: a five-level approach by Kaufman; a four-level approach called CIRO by Warr, Bird, and Rackham, and a five-level approach developed by Phillips and Phillips.[193] The five-level approach by Phillips and Phillips incorporates evaluating the return on the investment of training dollars and may be particularly useful for energy efficiency training programs.

Steps in Deciding if an Information/Education Effects Evaluation is Required

The information and educational program evaluation diagram presented in Figure 9.2 below is a roadmap to help guide planners through the steps required to decide if an effects evaluation is needed for the current program cycle.

Step One
The first step is to determine if the program should be or is classified as an information or education program. In California this decision is made in the early program planning and design efforts and is usually coordinated with the CPUC. Typically an information or education determination is made if the program provides only information or instructional services and if the program does not, in itself, have energy impact or market transformation goals that must be obtained. The program is also usually classified as an information program if its primary goal is to refer participants to other energy programs that do have energy impact goals.

Step Two
The next step is to determine if an information or education program effects evaluation has already been conducted for the program as currently designed and implemented. If not, then a program effects evaluation is suggested using the information and education program evaluation processes described above.

Step Three
In general, if a program effects evaluation was conducted for this program as it is currently designed and implemented, but was conducted prior to the previous program cycle, then a program effects evaluation is suggested for the current program cycle. This requirement allows the tracking of the effects of the program at least every two program cycles so that the effects of the program can be tracked over time to determine if the program is performing as intended and is accomplishing the program goals.

Step Four
Step four includes a group of conditional decisions that allow for an effects evaluation to be conducted more often or less often, depending on past evaluation efforts, the comparison of the results from these efforts, and if the program has significantly changed over that period of time. There are a number of possible outcomes to this decision

[193] All these approaches are discussed in (Phillips 1997).

process. These are presented below in bulleted formats to allow the reader to see the different possible decision results. Following the diagram (Figure 9.3) will help the reader move through the decision process.

Potential results from following the roadmap are:

- Even if an information or education effects evaluation has been conducted during the past program cycle, and if multiple evaluations have been conducted in previous cycles, a new effects evaluation is suggested if the program's goals or methods have significantly changed since these evaluations were performed.
- Programs with multiple evaluations over the past two program cycles that have not had significant program changes in their methods or goals do not need to consider an effects evaluation if the results of these evaluations are comparable and the findings are within 5% of each other. However, if the results of the previous evaluations show differences greater than 5% across comparable metrics, then a new effects evaluation is recommended.
- If a program effects evaluation was conducted on this program in its current design (program has not changed), and this evaluation was conducted *within* the last two program cycles, then an effects evaluation for the current program cycle is not recommended.

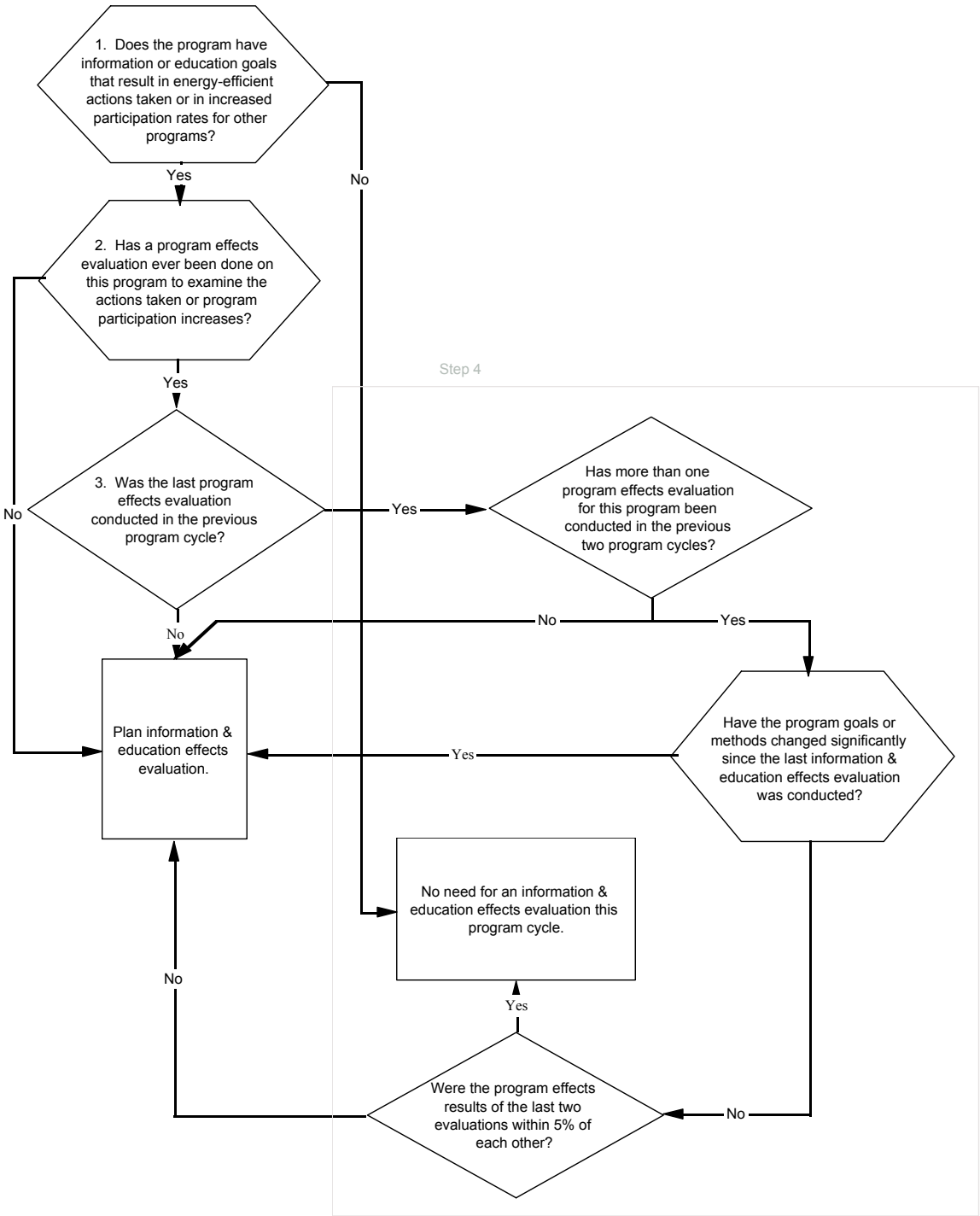

Figure 9.3: The Information and Education Evaluation Decision Roadmap

Chapter 10: Market Transformation Program Evaluation

Preface

This chapter of the Framework provides guidance on when a market transformation evaluation should be conducted and discusses the focus and activities associated with the evaluation effort. This chapter also discusses the general types of market-focused evaluations that can be conducted to support program planning and operational efforts, and at the same time help document market transformation (MT) program effects. This chapter also includes the market transformation evaluation roadmap. This roadmap (Figures 10.1 and 10.2) provides the step-by-step decision path to help program administrators and others determine when a market transformation evaluation should be considered. The information presented in this chapter can be useful for policy-makers, as well as program administrators, and implementers needing an overview of the types of market-focused studies that are available for their consideration.

Four important principles are central to the recommended approach for conducting evaluations of market transformation programs. These are:

1. Market transformation program evaluations need to be conducted at the market, sub-market or niche market level rather than at the program level. This can mean conducting a market evaluation on a group of programs operating in the same market or conducting multiple market studies for a program operating in a number of markets.
2. There are a number of important conditions and activities needed to be able to evaluate market transformation programs, including, an assessment of the program theory/logic model (PT/LM), a characterization of the market(s) in which the program(s) operate, the availability of baseline studies that provide a "starting point" for assessing market effects, market progress studies, and causality assessment that examines the linkages between the program and the observed market changes (where the program-induced changes are defined as the market effects).
3. If a MT program evaluation is conducted to document program-created market change as part of an effort to estimate the energy impacts from a MT program, an energy impacts evaluation may still, in some cases, need to be conducted to verify the impacts achieved through the MT program.
4. A MT evaluation critically evaluates causality and sustainability. It is recommended that long-term market effects only be claimed under three conditions:
    a. If the program theory specifically identifies the market effect(s) to be measured in the evaluation and provides theories supporting the causal relationship between the program's efforts and the expected market effect(s),
    b. If these efforts are supported in the program theory as being sustainable (i.e. last beyond the program period), or

c. If sufficient evidence is provided through a MT evaluation indicating that the market effect(s) have a high probability of being the result of the program's efforts.

Skills Required for Market Effects Evaluation

This chapter provides the market transformation program evaluation roadmap. In doing so, it makes references to a variety of steps and types of studies for which it only provides a very brief description. It does not provide the in-depth discussions needed to learn how to perform these studies. Instead, the chapter focuses upon the basic outline of the studies needed and refers to other documents and articles as reference material. A working knowledge of this body of literature is recommended for policy makers, administrators, program designers, and evaluation managers involved in decisions regarding market transformation program's evaluation efforts.

The basic foundation for market transformation evaluation research is based upon research approaches developed within the larger field of program evaluation and market research. This body of research is found outside of the field of energy efficiency program evaluation. Most energy program evaluators will need to develop some familiarity with the adaptation of evaluation approaches in order to conduct market transformation evaluations of energy efficiency programs. Evaluators will also need to be aware of the current practice and history of the evolution of market transformation evaluation within the energy efficiency program literature. Training or experience with research dealing with microeconomics, diffusion of innovation[194] and adoption theory is critical to market transformation evaluation concepts and practices. At the same time, knowledge of program evaluation methods and issues, including the use of program theory and logic models and alternative hypothesis development and testing within the field of evaluation, are more important than academic economics for these types of evaluations. Depending on the nature of the market barriers or the program interventions, an understanding of social psychology and the associated social science research approaches can also be helpful. Market transformation evaluation researchers also need a solid understanding of how markets operate, what kinds of conditions and events move markets toward product adoption behaviors and what kinds of program-induced and non-program-induced conditions influence changes in market conditions that can be examined in a MT program evaluation.

[194] The seminal work in this area is *Diffusion of Innovations, 4th Edition*. (Rogers 1995).

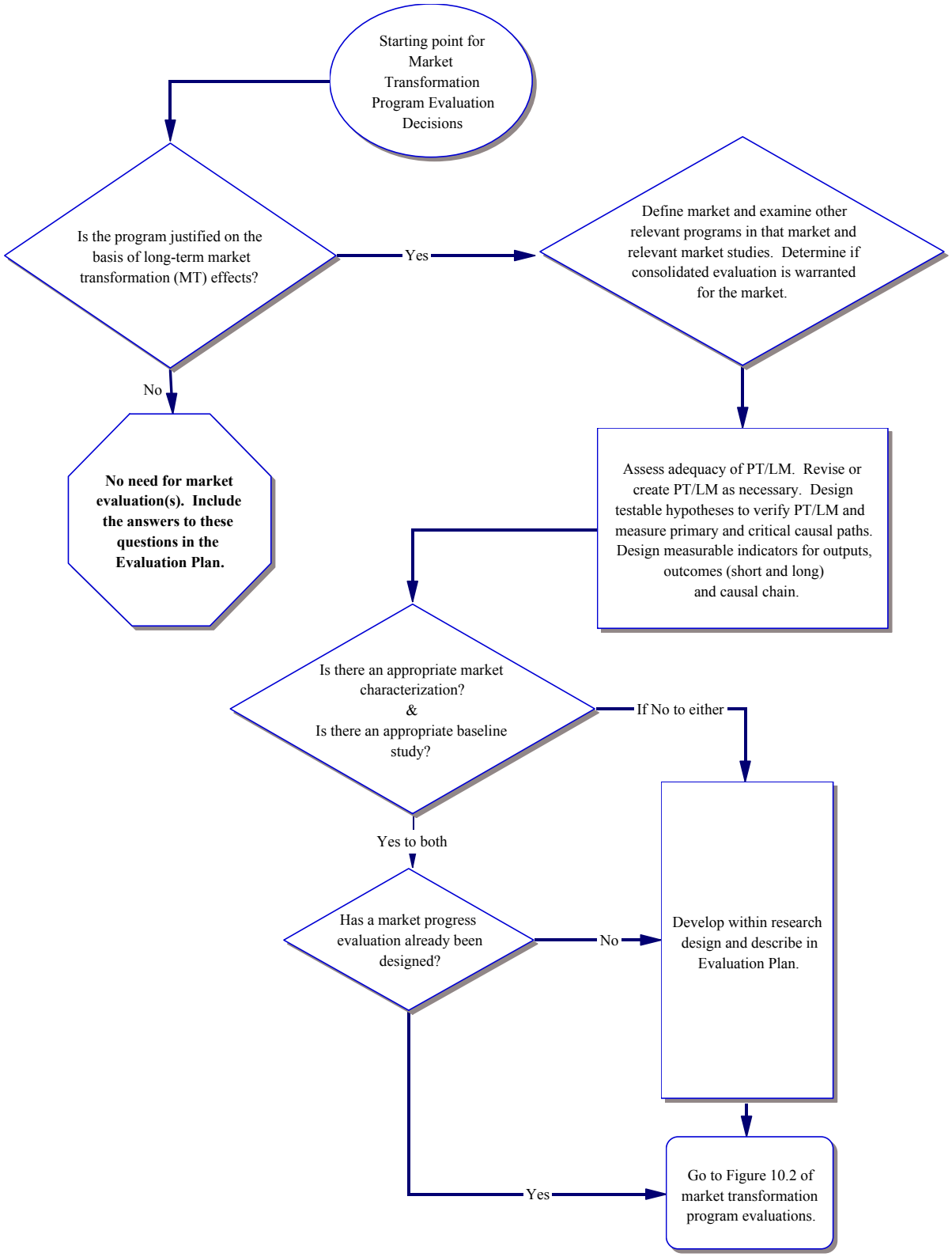

Figure 10.1: Market Transformation Program Evaluation Roadmap, First Half

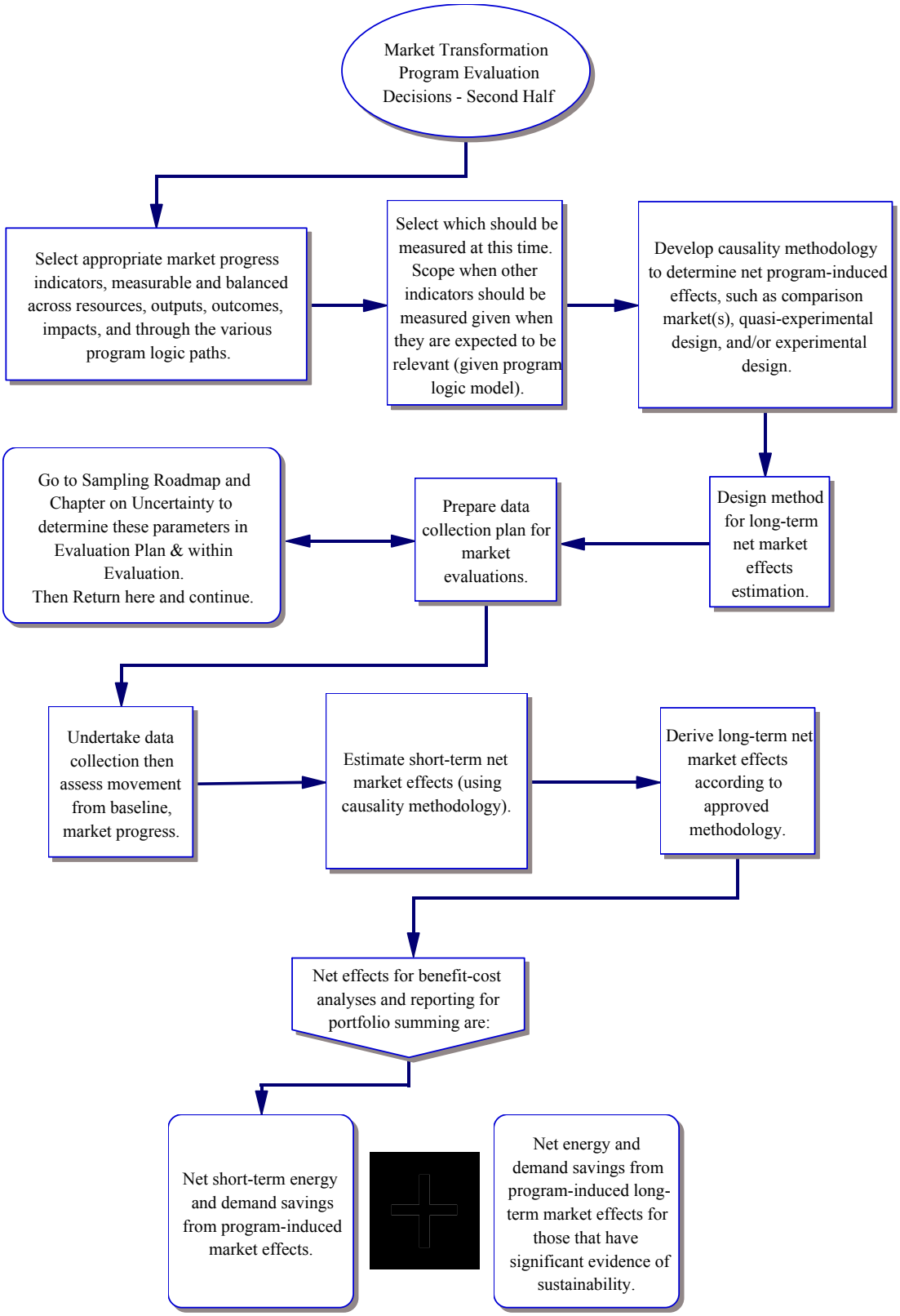

Figure 10.2: Market Transformation Program Evaluation Roadmap, Second Half

Introduction and Key Issues

As discussed in Chapter 5, Umbrella Roadmap, any program can claim to have an effect on the market in which it operates. Many programs have short-term market effects due to the interest they produce among trade allies, the learning that takes place during program influenced market transactions, or the contagious information exchange generated among participants and their peers. Short-term effects are best classified as "spillover" from the program intervention. As of 1998, few energy efficiency programs could claim long-term market effects, not because they are hard to measure, but because they are hard to accomplish. The main reason that (*most*) programs do not accomplish lasting market effects is because they are not designed specifically to address this goal (*often because of regulatory policy directions given to program designers*).[195]

Without going into a full discussion of how to design market transformation (MT) programs to achieve long-term market effects, it is important to reiterate the elements that are likely to be found in a logic model for a MT program. These include a sustained effort, substantial non-energy benefits, allies in the market who see opportunities to capture benefits, and knowledge of the key communications relationships in the market.[196] Of these, a planned sustained effort is a key indicator emphasized throughout Chapter 7 of the *2001 Framework Study*. There are multiple examples in the *2001 Framework Study* that illustrate how every program will have forward-moving "spillover" into future times, whether it is caused simply by education, or rebates, or through a number of program efforts. The longer an effective intervention is placed in the market, the greater the penetration will likely be, and subsequently, the longer the carryover effect can be expected to influence the market. If long-term effects are forecast, then it is important to plan for long-term measurement and evaluation. Sometimes paradigms change and programs are dropped, without a consideration for continuing to document the long-term market impacts.

Estimating the magnitude of the program's effect on the market (and the ultimate outcome in terms of energy and demand savings) to include evidence of causality, and the sustainability of the market effect (the level of future savings after program discontinuation) is the toughest issue surrounding evaluation of MT programs. In order to determine the cost-effectiveness of a market transformation program, the evaluation(s) needs to address the period over which the market effect will remain, the level of effect experienced in the market over time, the degree to which the program's efforts caused the market effect, and the amount of energy savings provided by the effect.

---

[195] *Market Effects Summary Study, Volume 1* (of 3). (Peters et al. 1998). A summary can be seen in "Measuring Market Transformation: The 1997/1998 California Market Effects Studies." (* Peters et al. 1998).

[196] "Wheat, Chaff and Conflicting Definitions in Market Transformation." (Keating et al. 1998).

Market Identification and Evaluation Consolidation

As discussed earlier in this chapter, given the nature of their objectives, MT program evaluations need to be structured around markets rather than around a specific program in order to identify program-induced activities and the relevant market changes. Where there is a direct relationship between a program's activities and a market effect, the design of the evaluation may not be a difficult issue. However, it is not uncommon to find multiple programs targeting the same customers within a market or targeting different customers across a set of markets. Likewise, even a single program may target multiple markets.[197, 198] In these cases, it may be appropriate to assess whether a consolidated evaluation covering multiple programs, or multiple markets, is desired.

An impediment to developing clear criteria for when to conduct consolidated evaluations is that there are no universally accepted definitions within the energy efficiency industry pertaining to what constitutes a program's market. For example, most observers would agree that a homeowner buying a CFL at a discount store and a manager of a manufacturing plant negotiating the purchase of new production equipment are participating in different markets. However, fewer program administrators or evaluators agree on whether a homeowner buying a CFL at a discount store, a small business owner buying CFLs by mail order, or a different homeowner buying a room air conditioner off the shelf at the same discount store, are participating in the same market.

Clearly, the question of whether two different programs are targeting the same market will often be one of degree, rather than one that lends itself to a simple yes or no answer. However, while there is no single universally accepted taxonomy of energy efficiency markets, there is more consensus as to the characteristics that tend to delineate and define markets. For example, most energy efficiency program professionals agree that the more the two energy efficiency related transactions share the following list of characteristics, the more likely it is that they are taking place in the same market:

- The same or similar product functions or categories of functions.
- Overlapping distribution chains.
- Nearby geographic locations.
- The same or competing manufacturers.
- Demographically or firmographically similar buyers.

---

[197] This is particularly likely to be an issue in a state in which market transformation is a secondary rather than a primary policy objective – a category that would seem at this time to include California. In a state in which market transformation is the primary policy objective, it is likely that programs will have been structured to correspond as closely as possible to individual markets, and thus there will tend to be a one-to-one correspondence between markets and programs. In a state in which market transformation is a secondary objective, programs are likely to have been structured based on other criteria.

[198] The converse issue, not discussed here due to space limitations, occurs when a single program targets multiple markets.

The decision to consolidate different programs for a MT program evaluation is often a subjective decision based on the relative extent to which the programs share market characteristics (such as those listed above). Likewise, the decision to address multiple markets in a single program evaluation should be based on the extent to which the transactions covered by the program share such characteristics.

Institutionally, it seems unlikely that any one evaluator or entity will have enough familiarity with the full range of programs being offered across the state to reliably assess whether there are other programs sharing enough of the market characteristics listed above that they should be consolidated for evaluation purposes. Thus it would seem that subjective decisions about consolidation would best involve a decision process considering multiple and perhaps conflicting issues. To address these issues it may be best to establish a process in which multiple issues can be addressed, considered and decided in a reasonable and timely manner. Such a process should be designed to operate within a more comprehensive evaluation planning process.

Assessing and Using Program Theory/Logic Models for Market Evaluation

This Framework emphasizes the importance of identifying, understanding and assessing the program's program theory and the associated program logic model (PT/LM). The Market Transformation Program Evaluation roadmap emphasizes the importance of understanding the goals of the program early in the evaluation planning process.[199] If the program theory and the program goals explicitly identify market changes as a program effect or if it is clear that the cost-effectiveness of the program is predicated on counting long-term market effects, then the logic model should explain the key linkages between program activity and predicted effect. These explanations should include a clear description of the targeted market(s) and a detailed description of their operations, the market hypothesis on which the program activity and the expected effects rely, a description of the baseline condition that is expected to occur without program intervention, and the causal linkages that lead from program activities to the accomplishment of the program's goals. A benefit of this type of description is that it forces the program planners to be explicit about what they are doing, why they are doing it, and what the results will be. Without this information it is difficult to understand how the program expects to accomplish its goals. For example, a general goal of "taking X actions to produce Y effect" provides little information that explains the path between program actions and desired results.

Because of the importance of the PT/LM for MT evaluation, a critical first step in the MT program evaluation roadmap is assessing the PT/LM and revising it. If a PT/LM does not exist, it may be necessary for the evaluator to create one that is supported by the program administrator. The ability of the evaluator to design an effective evaluation rests on the

[199] An introduction and references to the program theory/logic modeling literature can be found in Chapter 4 of this California Evaluation Framework. Examples of logic models are also provided in the Appendix to that chapter.

accuracy and comprehensiveness of the PT/LM. A poorly developed PT/LM can lead to a poorly designed evaluation plan unless the evaluator fully understands the program goals and activities and the operations of the market without the benefit of the PT/LM to guide the evaluation planning process. This means that the evaluator should not just grab an available PT/LM for the program or a similar program and use it to plan the evaluation. A critical assessment of the PT/LM or the creation of a PT/LM needs to be taken seriously by both the program team and the evaluation team. In this effort it is important to assess the PT/LM in terms of its thoroughness and accuracy in describing how a particular market operates and the various causal mechanisms lying within its patterns of operation. The PT/LM should be informed by a market operations theory.[200] The market operations theory is typically developed from a market operations analysis or a market baseline study. The market operations theory explicitly details the operations and operational linkages in the market that the program will need to influence.

With a PT/LM that accurately describes both the program logic and the operations of the market; an evaluator can move quickly to design a market evaluation.[201] The fundamental concept of evaluation is that it tests the program planning assumptions in the field.[202] (See the *2001 Framework Study*, Chapters 3 and 6.) The logic model provides the assumptions behind the intervention and the program design. However, this roadmap suggests that if the goal of the program includes market effects, and if the market characterization, the baseline condition, and the market hypotheses and causal paths are not clearly presented in the PT/LM, then the evaluation should first clarify these issues with the program designers/implementers.[203] Seldom can an evaluation provide enough evidence that the program induced the market effects without the causal relationships being clearly defined and the program testing these relationships in conjunction with the documentation of market changes that occur as a result of the changes caused by the program.

The logic model should provide a set of testable assumptions, some of which will provide market progress indicators and some of which may stretch the evaluation efforts out over a number of years.[204] One of the roles of the evaluators will be to sift through the hypotheses that can be tested to identify those that are most crucial for assessing whether the logic model is accurate, showing progress of the program toward changing the market, measuring or estimate energy savings, and testing the long-term effects. The

---

200 "Merging Program-theory and Market-Theory in the Evaluation Planning Process." (* Hall and Reed 2001).

201 The PT/LM and market models will inevitably be refined as more is learned about the market through market transformation evaluations and program implementation experience.

202 In a pure acquisition program, the evaluation tests the assumptions behind the planned savings – the hours of use, the change in watts, the loading factors, all the costs, and the participation rate, as well as the freeridership. The assumptions being tested for a market transformation initiative more likely lie in whether the program intervention creates the expected outputs, short-term outcomes and whether these lead to the desired long-term outcomes.

203 We note, however, that this does not relieve program designers of their obligation to attempt to ensure that the market hypotheses and causal paths in the program theory are clear to begin with.

204 The evaluation plan (as described in Chapter 5) should take this into account and indicate where a particular evaluation plan fits into the long-term evaluation plan for this particular market.

evaluation needs to test the assumptions upon which the logic model is built and whether the critical linkages actually occur (output to outcome to longer-term outcome to ultimate outcomes).[205] Not all assumptions are equally important, and it is possible for an evaluation to get bogged down measuring activities instead of outputs or outcomes of a program. (See Chapter 5 in the *2001 Framework Study.*) Both Weiss and Rogers[206] have recommended that evaluators consider testing the underlying mechanisms on which the program rests, which may include analyzing the psychosocial assumptions that can answer the question of why as well as how. Prioritizing the assumptions to be tested is important in order to control evaluation costs and focus evaluation efforts on those activities that provide the best assessments results.

There is an extensive literature on the use of program theory and logic models in the program evaluation field outside of the smaller field of energy efficiency program evaluation. The authors recommend that this literature be used within the energy program evaluation field to support and evolve the use of program theory in the evaluation planning process. It is important that the evaluation team understand both the strengths and the weaknesses of using program theories as a data source for developing the evaluation design. Program theory's relevance to the evaluation of market transformation can clearly be seen by a classic evaluation textbook's treatment of the subject.

> "*Early data on measures relevant to the program's long-term success are especially useful when the program has an extended time line. ... Theory-based evaluation calls for the collection and reporting of data on early phases that are conceptually linked to the program's ability to produce long-term success. That is, the data are not just desired interim outcomes; they are the interim outcomes that are considered to be paths to long-term effects. They are part of the assumed casual chain*."[207, 208]

Once the desired testable hypotheses are isolated, the evaluator needs to deal with three key elements of the evaluation design. These are baseline, measurability, and causal arguments.[209] Each of these is further discussed later in this chapter.

There have been market transformation programs that intervene in the market by influencing key market actors to make them proponents for energy efficiency. Often, in these cases, the program is trying to create "agents of change." There is significant

---

[205] Chapter 4 provides more information and references concerning program theory and logic models.

[206] "Which Links in Which Theories Shall We Evaluate?" (Weiss 2000); *Program Theory in Evaluation: Challenges and Opportunities*. (* Rogers, Hacsi et al. 2000).

[207] *Evaluation: Methods for Studying Programs and Policies*, 2nd edition. (* Weiss 1998).

[208] See also the Chapter on program theory in the largest-selling evaluation textbook: *Program Evaluation: A Systematic Approach, 6th Edition*. (Rossi et al. 1999); *Program Theory Evaluation: Practice, Promise, and Problems.* (* Rogers et al. 2000); *Program Theory in Evaluation: Challenges and Opportunities* (* Rogers, Hacsi et al. 2000).

[209] There are plenty of other issues, such as timing, cost, persistence, and relationship among the various programs and their evaluations, but these are central to the design.

theory from the diffusion of innovations literature[210] that can be referenced when examining the program theory in this situation. Prior research across many different fields has identified the key characteristics of effective change agents and the steps they use to help create the desired change. This prior work can be used to establish evaluation criteria for assessing the MT program's selection of certain entities or individuals as agents of change. These criteria can also be used to evaluate whether the changes for which the selected agents are supposed to be the causal drivers are actually occurring within the marketplace and whether the changes caused by the agents appear to be due to the program. These types of evaluation efforts can serve as an example of how theory-based evaluation can enhance MT program evaluations.[211]

Market Characterization and Baseline Measurement

Conceptually, market characterization and baseline measurement are two distinct activities, each of which plays a unique role in market evaluation. In practice, however, the boundaries between these two activities tend to be somewhat porous. For this reason market characterization and baseline measurement is discussed jointly in this section while exploring how the two activities are related.

Market Characterization can generally be defined as a qualitative assessment of the structure and functioning of a market, the primary purpose of which is to understand how the market operates in order to be able to effectively change the way in which the market functions.

Baseline Measurement can be defined as the quantification of key market indicators that have been or can be influenced by a program intervention. The primary purpose of the baseline measurement is to provide a basis for later comparisons of the status of the market after program intervention, in order to help assess the impact of the program.

Logically, while both market characterization and baseline measurement rely on the same kinds of market research data, there are at least three important distinctions between the two activities:

- Market Characterization efforts should generally precede baseline measurement. It is better to develop a qualitative understanding of a market before attempting to rigorously measure various elements and indicators of that market.
- From an evaluation perspective, the key purposes of a market characterization are to provide a starting point for assessing (or developing) the program theory, to help shape relevant baseline measurements and later help plan the market evaluation activities. The key function of the baseline measurement, on the other hand, is to help support an assessment of the impact of the program on the market.

[210] See Chapter 9 in *Diffusion of Innovations, 4th Edition*. (Rogers 1995).
[211] "Local Government Associations as Agents of Change." (* Megdal et al. 2000).

- Baseline measurement, due to its critical role in evaluating market effects, should generally be performed by evaluators. Market characterization, in contrast, may be performed either by program evaluators, market research professionals or by program staff.

In practice, however, market characterization and baseline measurement can often overlap. Budget constraints or small population sizes may force limitations on the size of baseline samples (such as the need to do interviews with the same market actors over a short period of time), thereby blurring the distinctions between the two activities. Likewise, time constraints may force evaluators to use the same data collection efforts for purposes of both market characterization and baseline measurement, thereby further blurring the boundaries. Nevertheless, it is still important to keep in mind both the differences and the similarities between these two MT program evaluation activities.

A complete taxonomy of all the elements that might be needed in a Market Characterization is beyond the scope of this project. What is known about a market should first be explored via secondary research. Next, the design of the market characterization study should then be based on the results of the usefulness of the findings from the secondary research. The research should be constructed so that together, the secondary and primary research should ensure an understanding of the current market structure, and operation. Previous work in California provides some guidance on this issue. For example, a white paper prepared in 1998 under the direction of the California Board for Energy Efficiency[212] listed the following types of information that should be considered in designing a market characterization or baseline study:

a. A summary of the specific technologies, services, or products being exchanged.

b. A summary of the major market participants and the nature of the transactions and other interactions between them, including buyers, sellers and intermediaries.

c. A description of how information concerning products and services flows through the market actors and the identification of key information hubs.

d. A description of the distribution chain – i.e., the variety of paths that a product follows on its way from a manufacturer to an end user. A helpful tool here might be a product flow diagram.

e. A description of the geographic boundaries of the market.

f. A description of the circumstances and settings under which transactions tend to occur, including the sales practices and the market events that tend to result in transactions within the market (for example, a decision to remodel precipitating the purchase of a new C&I lighting system.)

g. Approximate estimates of the number of buyers, sellers and intermediaries in the market, as well as an order of magnitude estimate of the total annual sales of relevant measures and services.

[212] *Proposed Guidelines for Conducting Market Assessment and Evaluation.* (CBEE Technical Services Consultants 1998).

Types of Market Indicators to be Studied

For both market characterization and baseline measurement, the types of market indicators to be studied should be driven in a large part by the (revised/assessed) program theory. In general, any key market indicator that the program theory predicts will be changed by the program should be considered for inclusion in either a market characterization or a baseline study. However, markets are constantly in a state of change. Likewise, program theories can and do change over time to reflect added knowledge of how a market works, to reflect changes in how the market operates and to reflect changes in a program's thinking about how to influence a market. If these changes are significant, a market characterization or baseline study that considered only those market indicators addressed under the initial program theory may be outdated before it is even completed. Thus it is important that market characterization and baseline studies seek to include not only those market indicators that are important under the initial program theory, but also those that could become important later. It is often the evaluator's job to predict what indicators may be important in the future and include those metrics in the study.

Often one or more market indicators can be created for each hypothesized outcome in the logic model. Similarly, there can be a testable hypothesis and an associated metric(s) embedded within the relationship between different program results. That is, multiple program results may be influenced by the same activities resulting from a single relationship or event hypothesized in the program theory. The ability to chart the flow from intervention to outcomes to further outcomes and the interactions of outcomes compared to the program theory is an important part of the evaluation effort. Assessing each link can provide information that can separate problems with the theory associated with the causal effects (the basis of program design) from failure of the program to set events in motion to achieve those effects. Disaggregating "theory failure" from "program failure" is a necessary assessment in order to improve the program should the causal expectations not go as planned. (A more detailed discussion of theory failure versus program failure can be found in Weiss.[213]) Without this assessment, it is difficult to know what needs to be changed and whether a change has a likelihood of succeeding.

Market indicators will likely be selected based upon the theoretical foundations within the evaluation team's program theory. Programs designed to speed the adoption and use of energy efficiency products may use research methods that focus on study approaches that measure the rate of customers moving down the product diffusion path.[214] These measurement approaches focus on the extent to which customers move along the diffusion path steps of technology awareness, persuasion, decision, implementation and

---

[213] *Evaluation: Methods for Studying Programs and Policies*. (* Weiss 1998).

[214] Rogers describes the stages of the innovation-decision process as knowledge, persuasion, decision, implementation, and confirmation. See Chapter 5 in *Diffusion of Innovations, 4th Edition*. (Rogers 1995). Others, such as Hassinger in "Stages in the Adoption Process," argue that need is required before someone will listen to messages that will provide knowledge about the alternative. (Hassinger 1959). Other authors have looked at the continuum as awareness, knowledge, adoption, and confirmation. There have also been other variations and suggestions that the process may be somewhat different depending on the product or service and entities involved.

confirmation.[215] In addition, evaluators can also measure the elements associated with purchase behavior, including the number of each of the different supply chain actors involved with the technology, store stocking behavior, and technology advertising/promotion behavior. A program focusing on eliminating specific market barriers will need to have measurement of these barriers and how they might be changing over time. Generally, the ultimate outcomes would include a market indicator measuring an analysis of efficient market share, i.e., the percentage of the measures or services sold that meet the energy efficiency criteria.

Orienting Theoretical Perspectives

There is a maxim that is popular in certain academic fields suggesting that there is nothing quite as practical as a good theory. The same is true for MT program evaluation: rooting market effects studies within a broad theoretical perspective regarding how markets work can help lend support to the program theory, improve the credibility of results, and provide important insights that might otherwise have been missed.

A wide range of such theoretical perspectives on markets is possible. However, one framework that has been used to help shape many past market evaluations in California and elsewhere is the market barriers paradigm first developed in the 1996 *Scoping Study*.[216] Under this paradigm, markets are viewed as being pervaded by certain recurring structural characteristics (market barriers) that tend to lead to gaps between the actual level of investment in energy efficiency and the level of investment that would be societally optimal. Thus, a key objective of market characterization, baseline, and market progress studies is to document these barriers and the extent to which programs reduce them.

Another theoretical perspective that has helped shape past market evaluations is diffusion of innovations theory. Briefly, the literature on diffusion of innovation provides a general model describing how products and ideas penetrate markets.[217] The theory includes elements that account for the structural, social and cultural environment in which decision-making occurs, the characteristics of firms and individuals that influence decision-making, the criteria that potential adopters use in evaluating the characteristics of products, and the characteristics of the product itself. MT evaluation studies that draw on diffusion of innovations theory tend to focus on these issues, and on the manner in which energy efficiency programs advance the diffusion of energy efficiency technologies.

---

[215] Hall and Jordan evaluated pre and post-program movement through the diffusion of innovation steps for several USDOE programs in *An Evaluation of Selected Technical Assistance Services Provided by the Federal Energy Management Program: Results of a 1999 Customer Survey*. (Jordan et al. 2001); and *The 2001 FEMP Customer Survey: Study Report*. (Hall et al. 2001).

[216] *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs* (* Eto, Prahl et al. 1996).

[217] Diffusion of Innovations, 4th edition. (Rogers 1995).

Other theoretical perspectives to help shape market effects studies have been proposed, including transaction flow analysis, communications theory, and network analysis.[218] [219] In addition, there is a vibrant and growing literature on the demonstrated utility of the various theoretical perspectives that have been applied and on the potential compatibility of competing perspectives.[220] In fact, Kunkle and Lutzenhiser argue that both are needed to create better energy efficiency policies and programs and to measure and evaluate them, a broadening of the perspectives and the paradigms examined should occur.[221]

At the end of the day, the authors believe that what is most important is not to choose from among the available theoretical perspectives on markets, but simply to recognize the critical role that any broad theoretical perspective on markets can play in underpinning market effects studies.

Baseline Measurement: Some Practical Considerations

In examining the program goals evaluators frequently learn that the program will "increase," "change," or "improve" something, but the program planners may not know where they are starting from – the baseline. This problem needs to be anticipated.

- If a baseline is not available, it should be determined.
- It is important to recognize that baselines are not static points, but slopes of expected change.
- In measuring baselines researchers need to keep in mind that many changes can occur without a program.
- Sometimes it takes an evaluation method or process to obtain a good estimate of where the baseline will move in the absence of the program.[222]

---

[218] See, for example, "Using Diffusion and Communications Theory to Expand Market Barrier Examination in MT Measurement." (Megdal, Lee et al. 1999).

[219] "Integrating Perspectives from Alternative Disciplines to Understand Market Transformation Policy in Energy Markets." (Megdal 1998). Though they do not use the term, most of the market barriers discussed in the *Scoping Study* (* Eto, Prahl et al. 1996) can be found in earlier literature discussions within Transaction Cost economics. One example of literature that looks at market operation and change from a network perspective includes Rosen's work on word-of-mouth marketing, "buzz": *The Anatomy of Buzz: How to Create Word-of-Mouth Marketing*. (Rosen 2000).

[220] See, for example "Why Can't We All Just Get Along? A Reconciliation of Economic and Innovation Diffusion Perspectives of Market Transformation." (Mast 1999); and "Does Talking About Barriers Just Get in the Way?" (Goldberg 2003).

[221] "Beyond Market Transformation: Some Perspectives on Energy Evaluation and Research and the Energy Efficiency Movement." (* Kunkle et al. 1999).

[222] One such process, though not the only potential one, is a Delphi process that uses market and program experts in a way to gather their opinions while not allowing any one or small group of individuals to overly influence the group outcome. See "Dressing the Priestess: Preparation for and Results of a Delphi Study for a Residential New Construction Program." (Blake et al. 2003). Recognize, however, that a Delphi process is not so much a forecast as it is a method to put together opinions without undue influence from forceful individuals. As such, it is only as good as the process undertaken to obtain a balanced and unbiased panel (e.g., not an over-representation of personnel involved with programs versus those knowledgeable about the markets).

- Even experts need to know where we are today in order to forecast where we might be over time.

Further, evaluators sometimes hear: "But the baseline is too hard (or costly, or vague, or changeable) to measure." It is possible that something is too costly to measure, and that other surrogate variables may need to be identified,[223] but clearly if something is too hard to measure, too changeable to get a handle on, or too vague to define, the evaluation is in trouble and the program may be meaningless. If we can't measure it in the baseline, how will anyone measure it as a program progress indicator or outcome variable?

Market Indicators and Measuring Market Progress

Chapter 6 on Impact Evaluation was concerned about estimating the program's "impact" on energy and demand savings. Impact evaluations are generally thought of as measuring the primary effects of resource acquisition efforts. MT programs also generally have ultimate goals of energy and demand savings. However, as discussed above these are obtained more indirectly through interventions designed to change the market as described within the program theory. (See the *2001 Framework Study* for a much more thorough description and the theoretical foundations for these various types of program interventions.) The effects being measured from MT programs are, therefore, those program induced market changes and ultimate changes. There are several differences between measurement of market transformation and resource acquisition efforts. These involve the questions being posed, the groups examined in the research, and how effects are measured.[224] One of the key differences is the need to identify and measure market indicators in order to measure market change, and then market effects (as the degree of market change(s) that are program induced).

The selection and timing of market indicators to be measured in a MT program evaluation is as much an art as a science. However, discussed below are a number of general principles suggested either by the literature on market effects evaluation[225] or the authors' experience.

Another method is the dynamic forecasting method provided in Chapter 7 of the *2001 Framework Study*. This method was then used in Wisconsin. A summary is available in "Dynamic Modeling of Market Effects & Spillover with Limited Information." (Goldberg et al. 2003).

223 Avoid getting trapped into looking at variables that are easy to measure, but aren't really related to the target variable – "looking for the keys where the light is good, instead of in the dark place where you dropped them."

224 A comparison of the measurement issues in impact evaluation versus MT evaluation is presented in "Methods & Measurement Issues for a DSM Evaluation Versus a MT Market Assessment and Baseline Study." (* Megdal et al. 1999).

225 The term "MT program evaluation" is used to cover all the types of studies needed: PT/LM assessment/refinement, Market Characterization, Baseline Studies, Market Progress, and Causality Assessment. Market effects measurement or evaluation generally is used to refer to the subset of Market Progress (with comparisons to Baseline Studies) and Causality Assessment.

The Need to Tell a Story

Perhaps the single most important tenet regarding the selection and timing of market indicators to be measured is the need for the evaluation to tell a coherent story about the chain of events that the program is intended to cause to occur. This point was made first by Feldman, et al.,[226] was elaborated on in the *2001 Framework Study*, and has been discussed in other papers as well. In general, "a coherent story" means that the evaluation should systematically explore the extent to which the chain of causally linked events hypothesized by the program theory is actually occurring. Toward this end, it is important both to include in the evaluation those market indicators that play the most pivotal role in the program theory, and to ensure that the evaluation is looking for changes in these indicators roughly when predicted by the program theory, rather than sooner or later.[227]

The Need for Targeted Studies in the Early Years

There is a tension in market evaluation between the need for timely feedback on program performance and the relative slowness with which most market transformation initiatives change energy efficiency markets. It may be years before meaningful change can be observed in overall market indicators.[228] Feedback may be needed much sooner than that, either to refine the program, or to decide whether it is working well enough to be continued. This tension can be partially resolved through the judicious combination of well-linked proximate and ultimate market indicators. However, it may still be difficult to obtain feedback on market effects early enough in a program's life cycle.

One approach that is useful for dealing with this challenge is to attempt to isolate subsets of market actors who can be expected to show attitudinal and/or behavioral changes particularly early if the program is working as expected – for example, market actors who come into direct contact with the program, including, but not necessarily limited to, program participants.[229]

Another primary use of early evaluation results is for testing the program theory and providing feedback to program planners and implementers to allow them to revise the PT/LM and refining the program design. Blumstein et al. describe the need for MT pilot efforts and using a continuous program improvement process as part of an approach based upon theory-based market transformation program development. These authors

---

226 "Measuring Market Transformation: First You Need a Story …" (Herman et al. 1997).

227 Of course, it is often useful to measure market indicators well before they are hypothesized as changing, in order to provide a baseline. However, our focus here is on when the evaluation anticipates and tests for a *change* in these indicators.

228 "Market Transformation: Expectations vs. Reality." (Hall et al. 1999).

229 Looking for behavioral effects on the part of program participants in the context of a market evaluation should not be confused with the measurement of participant energy savings in the context of a traditional impact evaluation. The key difference is that in the former case we are typically looking for relatively lasting behavioral changes that are tied to the longer-term market effects predicted by the program theory, while in the latter case we are interested mainly in near-term energy savings, and the participant behaviors that drive these savings.

suggests that this system of program planning, pilots, and evaluation is needed due to the lack of experience and understanding by some of the program designers and operators concerning how to transform markets (rather than just make resource acquisitions). Evaluation and market testing can be used to maximize program-associated learning and minimize program design and operational risk. These authors argue for the critical role that early MT research provides in helping program administrators and others understand market dynamics and the role of program intervention within a market.[230] Another example of the value of early MT evaluation is the program theory and evaluation work conducted on the California Nonresidential Standard Performance Contracting (SPC) program in 1999. This study demonstrates the importance of using a PT/LM process and early evaluation findings as input for significant program improvement recommendations.[231]

The Usefulness of a Consistent Framework and Institutional Process

Hastie et al. describe a framework and institutional process used to drive the evaluation of market transformation initiatives at the Long Island Power Authority (LIPA).[232] Under this framework, program designers complete a table, using a standardized format, showing the chain of specific behavioral effects the program is expected to have on market actors. Columns of the table show the categories of market actors influenced; rows show different time periods, ranging from less than one year after program implementation to more than five years; and individual cells show the specific behavioral change anticipated for a given market actor in a given period. The table is accompanied by a narrative summarizing the causal links between all of the anticipated behavioral changes. Program evaluators critique the table and narrative developed for each program as part of the development process. Following the review by the evaluators, the program designers revise it as appropriate. The resulting "standardized" program theory is then used to drive the selection and timing of market indicators to be measured in evaluating the program.

Hastie et al. conclude that this framework had a number of advantages, including: (1) illustrating deficiencies in the logic or understanding of specific market segments; (2) illuminating linkages between stimuli and effects and among market effects; (3) making the market transformation approach more tangible to the utility's Board of Trustees; (4) making obvious the market indicators that needed to be tracked; and (5) serving to ensure that all key market effects were being targeted by specific implementation activities. LIPA's experience would seem to suggest that it might be beneficial for California to develop a consistent framework and institutional process for the selection of market indicators to be studied in market transformation evaluations.

---

230 "A theory-based approach to market transformation." (Blumstein et al. 2000).

231 "Applying a Theory-Based Approach to California's Non-Residential Standard Performance Contract Program: Lessons Learned." (Goldstone et al. 2000).

232 "A Systematic Application of Theory-Based Implementation and Evaluation of Market Transformation Programs." (* Hastie, Prahl et al. 2000).

Dealing with Large Numbers of Market Indicators

A number of market indicators may be desired to be measured either as part of assessing the various program logic model elements and assumptions or as part of tracking market changes. One method of assessing a large number of market indicators is through the use of a binomial test, as proposed by Shel Feldman and used in a market evaluation by a GDS Associates team for the Boston Gas Company.[233]

Don't Forget about Gross Unit Savings

One of the distinguishing features of market transformation program evaluation is that it attempts to measure things occurring in the market in order to test the program assumptions. Social scientists have developed methods to measure many program or program-related effects, but in energy efficiency program evaluation, the ultimate outcome metrics about which policy makers care the most are energy and demand savings. While it is important to measure intermediate steps to claiming market effects, the ultimate value will be based on the energy effects. Identifying a program goal as a "market effect" doesn't mean that the laws of physics are suspended. Evaluators and planners still need the same basic foundation of savings per widget that is required for resource acquisition programs – the only difference is that for market transformation programs, estimates are derived by applying savings metrics to an estimated market change, rather than a building, a set of buildings, or some other easily identified physical unit (Principle 3).[234]

[233] For a summary of this work see "Measuring Market Transformation Progress & the Binomial Test: Recent Experience at Boston Gas Company." (* Spellman et al. 2000).

[234] It is worth noting, however, that there may be differences between the ideal methods for estimating gross unit savings for a market transformation initiative and a resource acquisition program. In the former case, the uncertainties surrounding gross unit savings may be swamped by those surrounding the number of measures the program caused to be adopted. Thus it may make sense to focus the bulk of available evaluation dollars on establishing attribution.

Causality Assessment [235]

It is important to recognize that changes observed in the market are not necessarily market effects induced by the program.[236] Market effects were defined in the *Scoping Study* to be explicitly different from market changes by this fact.

> *A **market effect** is "a change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy efficient products, services, or practices and is causally related to market intervention(s)."*[237]

The fourth principle for a high quality MT evaluation is to critically assess causality and sustainability. The most difficult of the key evaluation planning elements is setting up the evaluation to approximate causal arguments. When examining markets, it is seldom possible to prove causality. Markets are complex and dynamic. Designing programs that contain experimental design through implementation might be ideal. Yet, this is seldom possible and market operations can interrupt the intended experimental design. Evaluators often then can provide evidence of causality, but not proof.[238] Some of the better ways of building evidence include:

- An *a priori* description of the process and steps to a changed market condition that can be traced in the evaluation.[239]

---

235 An early cautionary note regarding this term seems appropriate. There are at least two terms in relatively widespread use to describe the process of assessing the extent to which an energy efficiency program has played a causal role in observed improvements in market functioning: "causality assessment," and "attribution." Both terms have their merits, but both come with pitfalls. On the plus side, the term "causality assessment" has a venerable history in the social sciences, as well as in the evaluation of other kinds of social programs, suggesting that this term offers the advantage of recognizing that energy efficiency program evaluation does not exist in an intellectual vacuum. On the minus side, the term "causality assessment" carries with it a certain hubris, given the profound difficulties of reliably assigning causes for social outcomes, some of which are discussed in this section. On the plus side, the term "attribution" seems more appropriately humble in its connotations. However, on the minus side, this term often seems to end up getting grossly misused in practice – as, for example, when it is used to refer to purely administrative negotiations over which of two overlapping programs should get to take credit for an observed market effect. On balance, the term "causality assessment" has been selected, but the provisional nature of all conclusions regarding the causes of market changes needs to be emphasized.

236 See *Market Effects Summary Study, Volume 1* (of 3). (Peters, Mast et al. 1998), page 33.

237 *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs* (* Eto, Prahl et al. 1996), page 9.

238 To understand the difficulty of the task, one must consider that even true experimental designs with large samples, and ones which are replicated several times, may not show a causal effect, but merely a consistent correlation. An experimental design that adds and removes a stimulus under laboratory conditions can provide both necessary and sufficient evidence that the cause and effect are identified – but this is often not an option available to public purposes programs.

239 There is significant literature within program evaluation on this method. For example, see two proponent articles and one opposed position article in *Program Theory in Evaluation: Challenges and Opportunities*. (* Rogers, Hacsi et al. 2000); "Ascertaining Causality in Theory-Based Evaluation" (Davidson 2000), "Causal Models in Program Theory Evaluation" (* Rogers et al. 2000); and the

- A quasi-experimental design where some, but not all extraneous, confounding factors can be controlled.
- A "comparison group" design (as opposed to a pure, randomly assigned control group) is one quasi-experimental approach that allows the researchers to find a similar market without the program impetus, but hopefully with other important factors at work.
- An explicit exercise in which reasonable alternative hypotheses are considered and shown to be less supported by the facts.
- Because the market actors in a program aimed upstream of consumers may have a clear picture of why they took action, there may be convergent evidence from interviews with market actors.

An early MT evaluation looked at the effects of lighting efficiency programs in New England and assessed causality by using the reduction in market barriers as part of the evidence for attributing the changes seen in the market as being caused by the program. This work is summarized in Meber et al.[240] This study is cited in the *2001 Framework Study* with the comment that this work notes that "the real problem is causality and whether any other factors could account for the observed changes" (pages 5-22).

An example of the use of quasi-experimental design, pre-post analyses, and comparison group analyses for a causality assessment for a market transformation program is provided by analysis done for one of the New York State Energy Research and Development Authority's (NYSERDA) programs. This causality work is summarized in a paper by Megdal et al.[241] This work is reported in NYSERDA's 2003 Evaluation and Status Report and was continued in its 2003 Evaluation and Status Report. (Both are available at http://www.nyserda.org/publications.html.)

Causal linkages between training of sales personnel in retail lighting and appliance stores with changes in ENERGY STAR® equipment was tested in California and reported in Ridge et al.[242] The study found evidence to support some but not all the causal linkages tested. Alternative explanations and external influences were assessed and discussed.

After a design is chosen that will put together an argument for causality (attribution) – if the research supports it – the question of measurement of savings still needs to be answered. Market effects can usually be measured with confidence (assuming the baseline, measurability, and causality elements have been met), but frequently without much precision. Measuring changes for direct participants and tight definitions of participant spillover are easier to make precise. It is much more difficult to measure

---

opponent piece: "The False Choice Between Theory-Based Evaluation and Experimentation." (Cook 2000).

240 "Converging On the Effects of Utility Lighting Efficiency Programs." (Meber et al. 1997).

241 "He Did It! He Did It! – Providing Evidence for Causality." (* Megdal, Coleman et al. 2001).

242 "Testing the Causal Linkage Between Training of Sales Personnel in Retail Lighting and Appliance Stores and Changes in Market Share of ENERGY STAR®–Qualifying Equipment." (* Ridge et al. 2001).

exact program induced energy savings when market effects involve a several step chain of causal logic.

For example, Feldman and colleagues provided significant evidence supporting the Consortium for Energy Efficiency's collaborative effort as a major contributor to the transformation of the residential clothes washer market.[243] In the Pacific Northwest, ENERGY STAR windows started with a baseline penetration of about fifteen percent (15%) of windows sold. Three years later, over sixty percent of the windows sold were ENERGY STAR. This did not happen anywhere else in the country (comparison groups), almost no utilities paid incentives (alternative hypothesis), and the window manufacturers unanimously credited the program in making this happen (knowledgeable market actors). The evidence and the confidence that these programs significantly contributed to these changes is overwhelming, but precisely how many gigawatt-hours were saved is not easy to know. In this case, although the effects are large and clear, policy makers may need to accept somewhat imprecise estimates.[244]

Even with a clutter of multiple programs and historical changes confounding the market baseline, it is still possible to ferret out clear evidence of market effects attributable to a planned intervention. An evaluation of the CFL program in California during the energy crisis by Rasmussen and her colleagues conducted such an exam.[245]

An alternative finding is seen in an evaluation for the "Just Enough Air" program. This program attempted to get wood shops to reduce the vacuuming force of their air-sucking equipment. The results showed that the baseline conditions were higher than expected; the market was smaller than thought; most of the participants said they wouldn't pay for it without utility incentives; there were very few additional market players after the program; and the only spillover was what might happen as a result of disseminating the "how-to" manual that the program produced. There did not appear to be near-term or long-term effects attributable to the program.[246]

Market effects are complex and often difficult to achieve, but when they are achievable, they should also be measurable. The challenges to evaluating market effects are real, but they are not overwhelming. Evaluators have more to track when looking at market effects than when looking for direct effects of acquisition programs. The program logic model may be complex, and many market activities and relationships may have to be

---

[243] *The Residential Clothes Washer Initiative: A Case Study of the Contributions of a Collaborative Effort to Transform a Market*. (Shel Feldman Management Consulting et al. 2001).

[244] In the northwestern windows example, the baseline could have grown to 18% or 24% in the absence of the program. The forward strength of the momentum after the utility support ended may add 15 – 25 % more penetration without further intervention. Frankly, the actual savings from changing the rated U-value of the square foot of window area will probably never match the engineering estimates. An approximate estimate of savings can be made but it will not be precise. (Quantec 2002).

[245] "Addressing Program Attribution in the Wake of the California Energy Crisis." (Rasmussin et al. 2003).

[246] *Just Enough Air: Efficient Pneumatic Conveying. Concluding Memorandum*, (Stout et al. 2003); and *Fan Speed Reduction in Pneumatic Conveying Systems in Secondary Wood Products Industry.* (Quantec 1999).

measured and compared in order to meet the challenges of causality. This may obscure the need to continue to obtain the measurements required to estimate demand and energy savings. It is not useful when accounting for public money to say that the market effects are large, but that savings can't be reasonably estimated. Demand savings will follow from the end use load shape of the relevant measures. The assumptions behind the savings estimates for each widget should be tested when the market effects are counted on to avoid power purchases.[247]

The challenge of identifying "net" savings of market effects is not much greater than for direct acquisition programs. If the baseline is known with confidence, calculating the net savings should be very easy. As it is the baseline of naturally occurring or non-program induced savings is an *estimate* that is made less certain by the fact that it must be *forecast* into future years in a hypothetical situation. However, every method for getting at net savings for direct acquisition programs is confounded by uncertainty (see Chapter 6 on Impact Evaluation). In addition to self-report problems, statistical estimation issues, and incomplete information, none of the field-tested methods ever assesses the full impact of the non-participant spillover. When information spillover is large, many of the actions taken by non-participants may be a result of the program or the program portfolio. Likewise, many of the actions taken by freeriders, may, in fact, be actions that are taken as a result of information spillover from the program or the program portfolio when that information is spread through the market.

Programs need not be justified by their intended market effects, but if these effects are important, they can and should be evaluated and documented.

Assess Sustainability and Plan to Measure Long-Term Market Effects and Post-Program Sustainability

There are two fundamental challenges surrounding assessment of the sustainability of market effects. The first is that sustainability is, by definition, a long-term concept. With energy efficiency markets, as with any human institution, the more time that passes, the greater and more complex the variety of social forces that come to bear, and thus the harder it is to reliably and coherently assess the reasons underlying observed changes. The second fundamental challenge is that, again by definition, sustainability focuses on what happens, or would happen, to the market after a program is withdrawn - and more often than not an assessment of this issue needs to be made while the program is still in operation.[248, 249]

---

[247] In a forthcoming report on the evaluation of the effectiveness of the Northwest Energy Efficiency Alliance, the evaluators suggest that the Alliance continuously test the underlying energy savings estimates for their programs. When these savings are multiplied, in order to estimate the long-term savings from the market effects, the resulting calculations need to be based upon the most accurate impact estimates.

[248] Indeed, to withdraw a market transformation initiative before there is good evaluation evidence regarding the sustainability of observed market changes would be imprudent, as it would unnecessarily risk premature loss of the observed program benefits.

The combination of these two challenges suggests that assessment of sustainability relies as much on reasoning as on empirical evidence. For example, the following are some bases for asserting sustainability that have been suggested by industry observers:[250]

- *Relative Irreversibility* or the existence of demonstrated market effects that are sufficiently fundamental that they would be difficult to reverse once the program is removed. An example of relative irreversibility that is often used is the retooling of manufacturer production lines to favor more efficient technologies in a manner that would be expensive to change back.
- *Fundamental Changes in the Incentives Facing Market Actors,* which may occur when a market transformation initiative causes beneficial changes in the basic market signals facing one or more categories of market actors. This may involve market structure changes that eliminate the market barriers or the development of profitable market entities that continue to drive market transformation. For example, many residential new construction programs over the past several years have sought to increase homebuyer awareness of energy efficiency, reasoning that significant increases in awareness can be expected to lead to more builders positioning themselves as energy efficient, and once a builder has successfully positioned itself this way, it is in its interests to continue to capitalize on the reputation it has established.
- *New Codes and Standards,* which use the force of law to institutionalize program-induced gains in energy efficiency.
- *Disappearance of Inefficient Technologies from the Market*. An example would be magnetic lighting ballasts for commercial facilities, which, following significant program-induced acceleration of the diffusion of electronic ballasts, have come to be relatively rare, at least in larger buildings.

What all these potential foundations for asserting sustainability have in common is that they require both a detailed understanding of the specific manner in which a program has changed the structure or functioning of a market (an empirical component) and a systematic argument as to why either human nature, economic forces, or laws can be expected to preclude these changes from reversing themselves (a reasoning component).

Despite the fact that an initial assessment of sustainability would generally be made while a program is still operating, it is difficult to overstate the desirability of continuing to systematically monitor energy efficiency markets once they have been deemed transformed, and energy efficiency programs withdrawn. All too often, such post-

---

[249] See "Saturation, Penetration, Transformation: How Do You Know When a Market Has Changed?" (* Grover et al. 2002) for a study that begins to assess large changes in penetration versus what may be sustainable as market transformation.

[250] See, for examples: "The Elements of Sustainability." (Hewitt 2000); *Market Effects Summary Study, Volume 1* (of 3). (Peters, Mast et al. 1998); a summary can be seen "Measuring Market Transformation: The 1997/1998 California Market Effects Studies." (* Peters, Mast et al. 1998); "Measuring the Market Effects of Utility Programs: Lessons from California." (Mast et al. 1998); and *CTAC Market Effects Study* (Hagler Bailly Consulting 1998).

program tracking falls by the wayside, as markets that are still being actively targeted for intervention absorb the bulk of evaluation resources. In those relatively few cases in which such post-program monitoring has been done for an extended period of time, the body of evidence regarding sustainability has been mixed. For example, in reviewing evidence from Wisconsin, Prahl and Pigg[251] found that in some cases market effects appeared to persist, while in other cases markets tended to partially regress toward their pre-program state.

Finally, once a reasonable case has been made for sustainability, a forecast of the long-term effects of the program on measure adoption is in order to support cost-benefit analysis. Methods for doing so are a frontier area in MT program evaluation. To date, most efforts to systematically forecast long-term program-induced acceleration of measure adoption have relied on Delphi analysis, under which a panel of experts provides individual forecasts, information on these forecasts is returned to the panel without being attributed to individual panel members, and panel members use this information to revise their forecasts.[252] The *2001 Framework Study* provides an extensive discussion of alternative approaches, including a dynamic modeling approach that relies on modeling the adoption process for repeated time periods. To our knowledge, most of this proposed approach has yet to be implemented anywhere. Recently, however, a pioneering effort in Wisconsin has used a type of dynamic modeling approach to forecast the long-term market effects of a set of programs targeting business customers.[253]

Needless to say, given the earlier discussion of the challenges inherent in assigning causality for even near-term market effects, it is important to remember that even the most rigorous methods for modeling long-term market effects can be regarded as producing highly uncertain results.[254] Given this, when the time comes to use the results of such modeling in a cost-benefit analysis, it may be appropriate to adopt non-traditional approaches that explicitly acknowledge the presence of extensive uncertainty -- for example, analyzing program benefits under multiple scenarios, or focusing on assessing the *probability* that benefits will exceed costs, rather than the more common focus on developing a single point estimate for the ratio of benefits to costs.

---

251 "Do the Market Effects of Utility Energy Efficiency Programs Last? Evidence From Wisconsin." (Prahl et al. 1997).

252 For an example of Delphi analysis in practice see "Dressing the Priestess: Preparation for and Results of a Delphi Study for a Residential New Construction Program." (Blake, Feldman et al. 2003). A word of caution here. The "forecasts" are the informed opinions of individuals knowledgable about the industry, but there are not formal forecasts that statistically are formed from extensive data analysis and modeling. Given this, it is extremely important that the Delphi panel be selected to be balanced between program supporters and those only knowledgeable about the industry. Otherwise, a biased forecast could easily be expected. Similarly, the information provided to the panel and in what form could also create bias. So great care and expertise is needed to properly use this technique and minimize potential bias.

253 *Business Programs Evaluation: Market Effects Pro Forma Estimates.* (Goldberg et al. 2003); and "Dynamic Modeling of Market Effects & Spillover with Limited Information." (Goldberg and Agnew 2003).

254 This does not mean that market transformation should be viewed as a second-class resource. All efforts at long-term forecasting must deal with considerable uncertainty – including forecasting for both energy efficiency resource acquisition, and supply-side construction planning, two of the main alternatives to market transformation.

Chapter 11: Non-Energy Effects Evaluation

Preface

This chapter of the Framework presents the non-energy effects roadmap and issues surrounding non-energy effects evaluations of California's energy efficiency programs. Non-energy effects (NEEs) are those net effects of a program other than energy related (energy or demand) impacts. These might include reduced water use, improved environmental conditions, higher productivity, or other similar benefits. However, these effects may also include increased water use, lower productivity or other changes that have a negative impact on the participant.

This chapter presents a definition for NEEs evaluations, a discussion of the reasons why NEEs evaluations are included in the Framework, and a discussion of the skills needed to conduct NEEs evaluations. The chapter also provides a discussion of the typical evaluation tools and evaluation approaches for researching NEEs, and examples of the various types of NEEs evaluations. The chapter ends with a presentation of the decision steps associated with planning and conducting NEEs evaluations within the Framework.

Introduction and Key Issues

This chapter focuses on the evaluation of non-energy effects within the Framework and specifically identifies the conditions under which Public Goods Charge funded programs can spend evaluation resources to conduct NEEs evaluations. These conditions are specific to the State of California and are grounded within a set of policy decisions that limit the conducting of NEEs research to effects that are approved for research by the CPUC, or on effects in which a specific program's success at obtaining energy resources is directly influenced. This policy direction was provided to the authors by the CPUC staff during the development of the Framework.

An example of a program that could justify NEE evaluation research might be one that promotes energy efficient clothes washers. This type of program not only emphasizes energy savings, but also focuses on reducing the amount of water used. The amount of water saved is one of the key customer effects that cause or influence participation levels. In this case, the level of water savings is a direct determining factor relating to the amount of energy saved. For this evaluation, resources could be spent documenting the amount of water saved in order to confirm the presence of the benefit, and to use those results in promotional efforts to increase participation. Another example would be a case in which a CPUC policy decision establishes a need to evaluate the level and type of generation-associated emissions reduced as a result of California's energy programs. In this case, evaluation resources could be spent on quantifying the level of emission reductions and the value of those reductions. The Framework's NEEs research decision steps presented at the end of this chapter provide additional guidance on this topic. These

decisions significantly limit the NEEs research that might be conducted in California for the types of programs covered by this Framework.[255]

While in some circumstances NEEs evaluations can be conducted under the Framework, the results from those evaluations are not to be included in cost-effectiveness tests used to evaluate the program's ability to save energy (kWh or therms) or to influence demand. However, the results of the NEEs evaluations can be used to help define the total public value of the programs offered (such as in a public purpose test).

Definition

For the purposes of this Framework, non-energy effects are defined as *any program implementation or participation effect that is other than the direct energy (kW, kWh, therms) effect(s) associated with an energy efficiency, resource acquisition, or resource procurement program funded in California through the use of Public Goods Charge or procurement funds that falls within the types of programs covered under this Framework.*

This is, by design, an encompassing definition that includes a wide range of effects. These NEEs can include effects on the program participant, such as increased or decreased productivity, increased or decreased sales levels, increased or decreased levels of water usage, or increased or decreased levels of comfort, to name a few. Non-energy effects can also be effects on the society as a whole, such as lower or increased levels of emissions from generation facilities and the associated changes in the health and welfare of the populations living within the areas in which these emissions are released. Likewise, NEEs would also include changes in the need for water supply and treatment facilities as a result of changes in water consumption levels resulting from program-installed measures. There may also be NEEs on the companies selling or marketing energy that are a result of the actions taken through an energy efficiency program, including changes in bill collection costs, levels of debt owed by customers, or company productivity or profitability.

In presenting this definition, the reader will notice that NEEs are defined not as "benefits" (as typical in other literature), but as conditions that can change for the benefit or the detriment of the participant, society or the energy provider. This definition is used because NEEs research has documented both negative and positive effects relative to one or more program outcomes.[256] While the vast majority of NEEs are positive, and provide added value, there are cases in which the effects are negative. In order to be objective, it is important for this definition to recognize that NEEs can be positive or negative. This definition fills that requirement.

---

[255] It should be noted that this Framework does not influence the research conducted for California's low-income programs. These programs have their own evaluation decision system that is beyond the jurisdiction of this Framework and which allows for a broader focus on non-energy effects research.

[256] *Non-energy Benefits to Implementing Partners from the Wisconsin Focus on Energy Program* (* Hall et al. 2003).

Why Non-Energy Effects are Included in the Framework

The evaluation and reporting of energy-program-induced NEEs is one of the fastest growing fields of energy program evaluation research. The primary reasons for this growth are:

1. The movement toward Public Goods Charge funded energy programs, in which energy consumers pay a small part of their utility bill towards a "Public Goods Fund" has increased the need to document the range of effects the public receives in exchange for those dollars. The central questions being asked by policy makers are: What is the total value the public is receiving in exchange for their Public Goods Charge dollar? In addition to energy savings, are there other societal benefits that are being achieved with these funds? Because energy programs may provide more than energy savings, public policy and public value accounting requirements suggest that there is a need, where appropriate, to document a wider range of effects the public receives in exchange for their program contribution.

2. Several evaluation studies have concluded that participants in energy programs often implement efficiency measures for reasons other than saving energy. For many participants, the energy savings are not seen as significant enough to make the change, but instead contribute to the value of the total package associated with the change. In many cases, the value of the NEEs to the customer can often be greater than the value of the energy savings. Customers may view their energy savings as a lower-priority byproduct of the change, rather than the primary reason for the change. For example, a California study of schools found that students in well-designed energy efficient daylit classrooms had substantially higher test scores than students in classrooms illuminated by artificial light.[257] The value of the increased learning to the parents and to the government body responsible for education can be priceless, just as the value of the increased learning to the child contributes to a potentially more productive life. Another California study found that retail sales increased significantly in energy efficient stores lit by natural light.[258] In this study, the energy efficient lighting system led to a 30 to 40 percent increase in sales (controlling for other variables), dwarfing the effects of the energy savings. In another study on the effects of weatherization measures, researchers found that participants reported that they were more comfortable and that the increased level of perceived comfort was significantly more valuable to them than the energy savings.[259]

3. Program managers can focus their marketing and outreach efforts on promoting the NEEs of a program in order to motivate customers to participate in their programs or to take actions. As a result, evaluators are being asked to provide

[257] *Daylighting in Schools: An Investigation into the Relationship Between Daylighting and Human Performance* (Heschong Mahone Group 1999).

[258] *Skylighting and Retail Sale: An Investigation into the Relationship Between Daylighting and Human Performance*. (* Heschong Mahone Group 1999)

[259] California Low Income Public Purpose Test. (TecMRKT Works 2001).

additional information on NEEs for use in the program's marketing and sales efforts. Program personnel that do not understand the range of effects associated with energy programs are at a disadvantage when approaching a potential participant. Non-energy effects research can help improve the effectiveness of energy programs by demonstrating a wider range of benefits that can be used to increase participation and build program impacts.

4. Increasing environmental awareness of the general public, state legislatures, and the regulatory community have motivated people to look beyond the direct effects of programs and to investigate the indirect effects of energy programs. In particular, there has been a renewed interest in promoting energy efficiency programs as a solution to local and regional air quality problems and as a way to reduce greenhouse gas emissions.[260]

These four reasons have focused considerable attention in some jurisdictions on evaluating the NEEs associated with energy programs. Taken together, the need to understand the full range of NEEs in the commercial, industrial, institutional, and residential markets can be important to regulators and organizations implementing Public Goods Charge funded programs. It can also be beneficial to program managers and sales staff who need to "sell" their programs.

Accordingly, there may be a need to be able to evaluate these effects under the Framework.

Evaluation Timing

Non-energy effects evaluations can be conducted at any time within the program design and implementation cycle, and in many cases after the cycle. However, there are some key considerations for the timing of these evaluations associated with the Framework. These include:

1. If the evaluation is being conducted to help support program enrollment and participation goals, the research should be conducted in time that the results can be used by the program manager to increase enrollments or installed measures. The results should be provided in time for program marketing materials to be developed, for staff to be trained on the use of the information, and for the effective use of the information to increase enrollments. Accordingly, the NEEs evaluation may need to be conducted in the early months of the program, but after enough participation has accumulated to conduct the evaluation. For example, an evaluation of an energy efficiency program that saves water heating costs by reducing hot water use will need reliable information on the amount of water and water cost savings the program provides in order to use these data in program outreach materials.

[260] *Scenarios for a Clean Energy Future*. (Inter-Laboratory Working Group 2000), and "Opportunities for promoting energy efficiency in buildings as an air quality compliance approach." (Vine 2003).

2. If the evaluation is conducted to support a CPUC policy to better understand NEEs, then the evaluation will need to be conducted before the date on which the CPUC needs the information. This timeline may, or may not need to be coordinated with the program's implementation cycle. For example, if the CPUC issued a policy directive that the Commission needs to understand whether California waste streams are influenced by the energy efficiency changes to industrial operations, the study would need to identify an appropriate timeframe in which to assess the changes in waste steam levels relative to the programs that influence those levels. This effort would require coordination with the program implementation timelines. However, if the CPUC needs to quantify the greenhouse gas reductions of these same programs, then an evaluation of the emissions reductions may not need to be conducted during the program cycle. This type of an evaluation may want to conduct a time dependant valuation[261] of the emission levels associated with energy production and use within different parts of the State and then model the time dependant load effects of those programs.[262]

Information on the NEEs needing to be quantified for a particular program may also be available from other sources. For example, a NEEs evaluation may use information (i.e. contact information) from other programs or program activities that best match the effect(s) that need to be researched. As an example, if program managers wanted to design information materials on the value of improved levels of comfort associated with a residential program, the evaluators may want to interview participants from past programs that provided similar types of measures to similar types of customers.

The above examples show that the timing of specific NEEs research may need to be tailored to the specific effects that need to be quantified or to meet specific policy requirements. There is no simple rule concerning when NEEs research should be conducted. Rather, each project needs to set a research timeline to match the effect being measured, the relationship between the effect and the program implementation efforts, and the needs of the information consumer.

Skills Needed to Conduct Non-Energy Effects Research

The non-energy effects being researched and the evaluation approach considerably influence the skills needed to conduct the NEEs research. These studies can require an assortment of skills available within the evaluation community and can cut across the skills normally associated with individual firms or evaluation professionals. While there is no general rule that can be applied to the skills needed to conduct NEEs research, a range of the skills and experiences associated with NEEs research include:

- Expertise in designing objective research approaches that match the information available to the researcher and the research budgets,

[261] *Time Dependent Valuation (TDV) Economics Methodology*. (Heschong Mahone Group 2002).
[262] "Wisconsin's Public Benefits Approach to Quantifying Environmental Benefits: Creating Different Emissions Factors for Peak/Off-Peak Energy Savings," (Sumi et al. 2002).

- Understanding of, and experience with, the application of experimental and quasi-experimental research designs,
- Expertise in data analysis methods, software and analysis approaches,
- Experience in hypothesis and alternative hypothesis testing for causal relationships or indicators of causal relationships,
- Experience in standard evaluation research techniques, including:
  - Survey design and administration techniques,
  - In-depth interview design and implementation,
  - Secondary data mining and analysis.
- Experience in reporting and presenting evaluation research findings, including:
  - Ability to develop clear evaluation reports that describe the evaluation approach, the reasons for the selected approach, the strengths and weaknesses of the approach, the analysis conducted in enough detail that it can be replicated, and a presentation of the evaluation results.
- Experience in and an understanding of in the subject area being studied (e.g., indoor air quality, productivity, greenhouse gas emissions, etc.):

Evaluation Planning and Approach Tools

All NEEs evaluations start with a detailed evaluation plan that presents the researchable issue(s) being investigated, the evaluation approaches considered, the evaluation approach selected, and the strengths and weaknesses of the selected approach. The plan should also include details of the analysis to be done, including the data to be obtained and analyzed, sampling approaches and sample size selection criteria, levels of precision and confidence intervals, and a presentation of the anticipated levels of uncertainty associated with the research approach and the data assessment efforts to be employed.

There is not a standard approach for quantifying the levels of NEEs associated with energy programs. These studies examine a wide range of potential effects and require flexibility in the evaluation approach. The NEEs evaluation may employ one or more of the following research methods as appropriate to match the research goals:

- Surveys: telephone, mail, Internet, e-mail,
- Interviews: in-depth, small group,
- Focus groups,
- Site visits and effects verification visits,
- On-site metering or monitoring of effects or anticipated effects,
- Secondary data acquisition, review and analysis,
- Billing and payment data analysis,
- Economic or impact modeling tied to primary data collection or secondary data analysis,
- Pre- & post-program participation measurements and assessments,
- The use of comparison groups so that the net effects of the program can be estimated, and
- Other methods as appropriate for the effects being studied.

Research Methods and Approaches

As indicated above, there are many approaches that can be used to conduct evaluations of NEEs. This section of the non-energy effects chapter presents a small sample of the research approaches presented in the literature that have been used to estimate the presence or absence of specific NEEs, and in most cases, report their estimated value. This presentation is not meant to be exhaustive, but illustrative of the range and type of approaches used in this field. The examples presented in Table 11.1 include the results of low-income, residential, commercial and industrial program evaluations in order to provide a range of approach examples.

Table 11.1: Examples of Non-Energy Effects and Research Designs

Program	Effect investigated	Methodology used	Reference
Wisconsin's Public Benefits Charge Funded Business Programs	Presence and value of increased productivity from installed energy projects.	In-depth interviews with participating partners who installed measures incentivized by the program.	[263]
Several Different Programs and Case Studies	Increased productivity in the industrial sector from energy efficient improvements	Economic modeling of the results of several case studies and impact reports with productivity assessments	[264]
Ohio's Weatherization Program	Value of reduced emissions from weatherization measures.	Application of traded values for SO_2, and application of NO_x control costs for burners and over-fire technologies to the estimated levels of reduced emissions.	[265]
Detroit Edison's low-income energy education program	Actions and behaviors taken as a result of energy program training.	Survey of participants 6 months to 1 year after participation compared to a matched control group.	[266]

[263] *Non-Energy Benefits to Implementing Partners from the Wisconsin focus on Energy Program.* (* Hall and Roth 2003).

[264] *Productivity benefits of industrial energy efficiency measures*. (Worrell et al. 2001).

[265] "The Environmental Benefits of Low-Income Weatherization,"(Hill et al. 1999).

[266] *Residential Low-Income Energy Management Program: Process Evaluation Report*. (Hall et al. 1997).

Table 11.1: Continued

Program	Effect investigated	Methodology used	Reference
Iowa's State Weatherization Program	Increased local-area economic activity as a result of program spending.	Economic "input – output" model of economic effects.	[267]
Residential Energy Efficiency Programs	Influence of energy bills on real estate value.	Regression analysis of American Housing Survey Data on housing value and energy consumption. In-depth interviews with high efficiency homebuilders.	[268, 269]
Missouri Gas Energy's, Washington State, and Oregon State Weatherization Programs	Reduced arrearage levels for program participants.	Pre and post program billing data analysis of arrearage levels and payment data using matched comparison groups of non-participants.	[270, 271]
Colorado Public Service's Energy Savings Partners Program	Reduced number of emergency calls as a result of home weatherization.	Uses the difference in the frequency of emergency calls to the utility between participants and non-participants.	[272]
Detroit Edison's Residential Energy Management Program	Changes in the number of bill payments made on time as a result of program participation.	Analysis of billing and payment data for program participants before and after program participation.	[273]
U.S. Army's Energy Efficient Military Barracks Construction	Incidents of increased respiratory disease in newer, more energy efficient housing.	Comparison of medical records of Army recruits living in older, leaky, uninsulated barracks with those living in newer energy efficient insulated barracks.	[274]

[267] *An Evaluation of Iowa's Low-Income Weatherization Efforts*. (Pigg 1994).
[268] *Evidence of Rational Market Valuation for Home Energy Efficiency.* (Nevin et al. 1998).
[269] *Non-Energy Benefits Cross Cutting report: Year 1 Efforts*. (* Hall et al. 2003).
[270] *Process and Impact Evaluation of Missouri Gas Energy's Pilot Weatherization Program.* (Hall et al. 1998).
[271] "Effects of Weatherization Programs on Low-Income Customer Arrearage." (Khawaja et al. 1992).
[272] "Evaluation of Non-Energy Benefits from the Energy Savings Partners Program." (Magourik 1995).
[273] *Detroit Edison's 1995 Residential Energy Management Program Arrearage Analysis*. (RLW Analytics 1997).
[274] *Chemical, Microbiological, Health and Comfort Aspects of Indoor Air Quality – State of the Art*. (Knoppel et al. 1992).

Partial Listing of Non-Energy Effects Presented in the Literature

The NEEs literature includes a significant list of documented and anticipated effects from energy programs. The following table presents the effects that are currently found in the literature that have some level of documentation of the effect as a result of an energy program. The authors make no assumptions about the accuracy or validity of these effects, but present them to suggest the potential range of effects. In reviewing Table 11.2, it should be noted that the reverse effects should not be eliminated as a potential program result when establishing evaluation methods, tools, or in conducting the analysis.

Table 11.2: Partial Listing of Non-Energy Effects

Potential Effect	Definition
Ability to pay utility bills and reductions in utility collection costs	Program-induced energy savings can lower energy bills and influence the ability of participants to pay utility bills, leading to a reduction of utility collection costs.
Changes in features/options of energy efficient measures	Program measures may have more or less features and options than the less efficient measures being replaced.
Economic stimulus impacts	Program spending may have a net local economic effect from the program's materials and labor. Participant savings may have an economic impact as participants spend their "savings" on other goods and services. Net job creation may lower unemployment costs. As trade allies participate in programs, they may increase their product sales resulting in job creation and retention. Energy efficiency improvements may allow manufacturing plants to retain or hire more workers due to increased productivity or profitability associated with the energy efficient technologies.
Fewer customer callbacks	Trade allies that participate in programs may experience fewer customer callbacks related to problems with less efficient equipment or building features that may not be as well-designed, constructed or installed. Alternatively, new energy efficient technologies may have increased callback rates compared to standard equipment.
Greater market share for vendors and increased sales	Vendors who offer energy efficient products and services can capture greater market share, and increase sales if they are able to take business away from competitors that do not offer energy efficient products and services.

Table 11.3: Continued

Potential Effect	Definition
Improved comfort - draftiness, thermostat setting	Program measures, such as insulation and weather-stripping, may reduce drafts or the need to increase thermostat settings, thereby improving comfort.
Improved energy and energy-related perceptions, recognition, choices, or behaviors	A program may provide educational components such as written materials and on-site demonstrations. These features may improve participant knowledge and allow them to make better energy efficiency decisions. Programs may improve energy-related perceptions, ENERGY STAR brand recognition, and their energy choices and behaviors.
Improved home appearance/aesthetics	Program measures, such as energy efficient appliances and equipment may improve the condition or appearance of the home, adding value.
Improved learning	Students may learn more when in an environment that is more energy efficient and more aesthetically productive to the learning environment.
Increased customer loyalty	Vendors and utilities that offer energy efficient products and services may be able to retain customers more readily.
Increased customer satisfaction	Vendors and utilities that offer energy efficient products and services may experience increased levels of satisfaction from their customers.
Increased product or service quality	Energy efficiency improvements are often associated with improved product or service quality.
Increased productivity	Energy efficient technologies and practices may also increase productivity.
Increased profitability	Trade allies that participate in programs may experience increased profitability.
Increased worker satisfaction and morale	Energy efficient technologies and practices can result in greater worker satisfaction.
Lower job-related injuries or illness	Energy efficient technologies and practices may result in fewer workplace injuries and illnesses.
Noise reduction - inside and from outside the home	Program measures such as insulation and weather-stripping may reduce noise being transmitted from the outside to inside the home.
Public relations or image	Energy efficiency may improve corporate image leading to customer attainment or attraction.
Reduced air emissions	Energy efficiency can reduce emissions and pollutants, contributing to global warming and acid rain, or to human health improvements.

Table 11.4: Continued

Potential Effect	Definition
Reduced indoor poisonous gases such as Carbon Monoxide (CO) and Carbon Dioxide (CO_2)	When $CO_{(2)}$ reduction measures are included in a program, health and safety effects can be accrued to the residents in terms of lower hospitalization costs and health-related expenses or avoided illnesses.
Reduced product losses	Energy efficient technologies may reduce product spoilage or result in fewer manufacturing defects or errors.
Reduced use of materials and/or waste and/or increased recycling	Energy efficient technologies and practices may reduce product and material waste, land filling and hazardous material handling costs.
Self sufficiency and improved household economics	Program-induced energy savings lowers energy bills and increases the economic security of participating households. This may lead to reductions in debt.
Water and waste water savings	Low-flow showerheads save energy and water, reducing water resource needs and aquatic impacts. Similarly, programs reducing water consumption via energy efficiency measures reduce the need for water treatment and sewage treatment facilities.

Making Decisions and Selecting Methods

This section presents the evaluation decision roadmap for determining if a non-energy effect evaluation is needed or can be funded for a specific program.

Steps in the Non-Energy Effects Evaluation Roadmap

Figure 11.4 describes the non-energy effects evaluation roadmap and helps to guide readers through the steps involved in making decisions about when NEEs evaluations can be conducted for California energy programs using Public Goods Charge funds. For programs operated through investor-owned utilities, the utility company administrating the programs typically makes these decisions. For programs operated through independent contractors, such as third-party programs, the lead program administrator typically makes these decisions.

Because California Public Goods Charge dollars are typically not spent evaluating NEEs, it is recommended that prior to conducting a NEEs evaluation the program administrator obtain permission from the CPUC to allocate program dollars to this research. Once permission is granted and an evaluation plan developed, the CPUC can approve or request modifications to the evaluation plan as appropriate. In budgeting the NEEs evaluation, the administrators will need to consider the sampling requirements and the threats to validity discussed in the Framework's Sampling and Uncertainty chapters (12 and 13, respectively).

In using the NEEs evaluation roadmap, decision makers should move through all of the decision steps in the roadmap to determine if a NEEs evaluation should be conducted. As a general rule, the NEEs roadmap suggests that programs needing NEEs documentation to increase program participation and related energy savings can conduct NEEs evaluations. However, these studies are not required by the CPUC. If the CPUC identifies one or more NEEs that need to be evaluated in order to document program effects, the CPUC can require specific non-energy program evaluations.

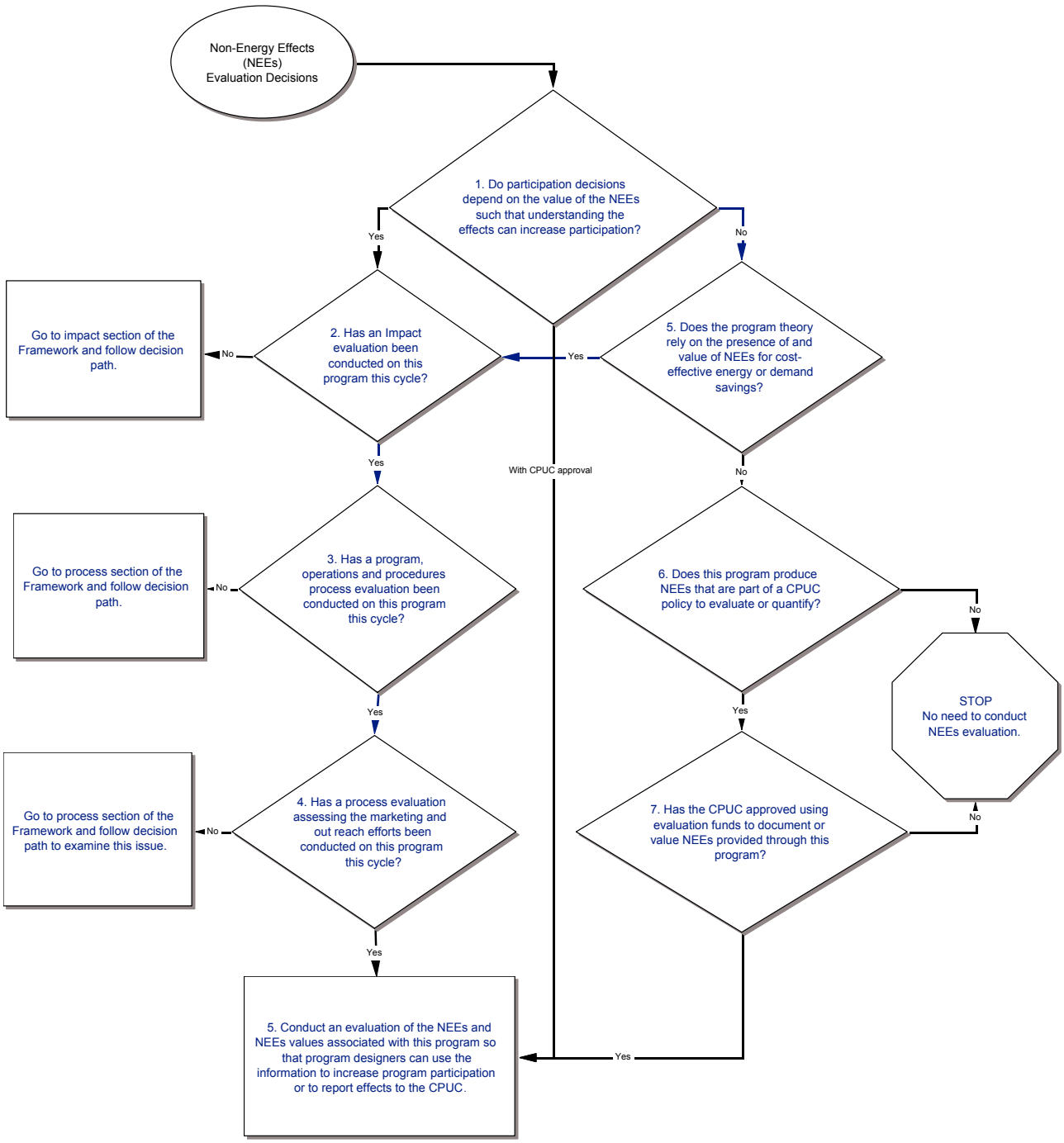

Figure 11.4: Non-Energy Effects Evaluation Framework Decision Path

The steps in the NEEs evaluation roadmap (Figure 11.4) are described below.

Step One

The first step is to determine if an evaluation documenting the presence and value of a non-energy effect can be used to help the program gain participants to increase the program's energy effects. Because the State of California, through the CPUC, is currently responsible for administrating the Public Goods Charge funded programs and approval for the procurement programs, there is an inherent need that all programs provide accurate, reliable information in their outreach and marketing efforts. In addition, the CPUC realizes that there may be programs in which participation decisions rest more with the value of the non-energy effect than with the value of the energy impact associated with the program. While the primary goal of these programs is to achieve energy resources (directly or indirectly), it is realized that promotional information may need to be used that focuses on program benefits other than the energy savings. For this reason, programs that rely on the value of the NEEs to achieve participation can use evaluation resources to fund studies on the NEEs needed to increase participation.

The decisions associated with step one are either "no" (the program does not rely on the value of the NEEs to gain participation in order to accomplish its energy saving goals); or "yes" (the program does rely on the value of the NEEs). For programs that rely on the value of the NEEs, the program theory should clearly identify the importance of the NEEs in gaining energy impacts. In cases where the NEEs are used to accomplish energy savings goals, assessing the program theory and designing the evaluation accordingly become more important to be able to understand and document the linkages between the program intervention, the NEEs, the resulting behavior/actions obtained, and the resulting energy savings. (See the Framework section on program theory and logic model in Chapter 4 for more information and references.)

If the decision from this step is "yes," then a NEEs evaluation can be a worthwhile addition to the evaluation effort. As a general rule the results from the step 1 decision will be yes or no. The paths from these two decisions are presented in the roadmap, and extend off the left or right of the roadmap's step 1 decision box. A "yes" to a step 1 decision moves the decision maker into a consideration for an impact or a process evaluation. This direction is meant to reflect the often higher priority placed upon impact and process evaluation if there are evaluation budget limitations that require a choice be made between the types of evaluations that can be conducted in any program cycle. However, in many cases it may be desirable to conduct the non-energy effect evaluation prior to the impact or process evaluations. In cases where the administrator thinks the program can increase the program's energy impacts beyond their stated goals, or they will not be able to reach their energy impact goals because of a lack of NEEs information, the administrator can request the CPUC to approve the use of funds to conduct the research and not wait until the impact and process evaluations are completed. For this reason a third decision branch is added to the step 1 decision that allows the administrator to move directly to a NEEs evaluation with CPUC approval.

Step Two
If the decision from step two is "yes," and the program has energy impact goals, then the administrator may still need to conduct an energy impact evaluation of the program before funds are spent conducting a NEEs evaluation. If the program does not have specific energy impact goals, then an energy impact evaluation is not needed. This step is placed in the roadmap to make sure that the impact evaluation is placed in the highest priority position for the use of evaluation resources. If the program has energy impact goals and is in need of an impact evaluation (as a result of the Chapter 5 Umbrella Roadmap and current CPUC guidelines), then the impact evaluation takes precedent for the evaluation dollars. If an impact evaluation is not needed or has been conducted consistent with the impact evaluation roadmap (see Chapter 6: Impact Evaluation), then the process moves on to step three.

Step Three
This step is similar to step two except that the step three decision is associated with a process evaluation rather than an impact evaluation. The CPUC's second highest priority for spending evaluation resources is for process evaluations. For this reason, the process evaluation is suggested for most programs. Step three asks if the program has conducted a process evaluation according to the requirements presented in the Chapter 5 Umbrella Roadmap and current CPUC guidelines. In this case, the process evaluation identified in step three is one that examines the program's design, management, operation and implementation procedures to help the program improve its energy impacts or to become more cost-effective. If a process evaluation has not been conducted for this program, then a process evaluation is suggested prior to a NEEs evaluation. If a process evaluation has been completed in accordance with the process decision roadmap (see Chapter 8), then a NEEs evaluation may be conducted. (It should be noted here that the process evaluation roadmap suggests a number of conditions in which an administrator may not need or want to conduct this type of a process evaluation.)

Step Four
This step is similar to step three, but focuses on a decision directly associated with a specific kind of process evaluation. The step four decision focuses on whether a process evaluation has been conducted on the program's marketing and outreach efforts. For most programs, the success of the program can be directly linked with the success of the program's marketing efforts. If a program is not marketed well, then participation can suffer, causing difficulties in achieving the desired energy impacts. This step is placed in this roadmap to bring attention to the potential need for the process evaluation to examine and assess the primary program characteristics that have a direct bearing on the success of the program. The marketing and outreach process evaluation focuses on how the program is marketed, including the targets selected for the marketing approach, the materials used, and the results of these efforts. If the process evaluation conducted for the program included an assessment of the program's marketing and outreach efforts, then a NEEs evaluation may be conducted. If the process evaluation did not assess the marketing and outreach efforts, then the program may want to conduct this type of process evaluation prior to conducting the NEEs evaluation. (It should be noted here that the process evaluation roadmap suggests that if a program is achieving its marketing and

enrollment related goals the program administrator may not need or want to conduct this type of a process evaluation.)

Step Five

This step is contingent on the decision from step one. If the step one decision is that "no," the program does not depend on the value of the NEEs in order to increase participation and enrollment rates, then the decision process moves to step five. In this step the decision relates to whether the program theory relies on the presence of and value of one or more NEEs in order to achieve cost-effective energy or demand savings. While step one asks about reliance on the non-energy effects for gaining participation in the program, the step five decision rests on whether the program theory indicates that cost-effective energy or demand savings rests on or is heavily influenced by the value of the NEEs. This decision presents a somewhat different approach to determining if the value of the non-energy effect is a critical or important step in the ability of the program to achieve energy savings, regardless of whether enrollment or participation in the program is influenced by the presence of the non-energy effect. While enrollment in a program may be consistent with the expectations of the program designers, this does not mean that all appropriate energy-saving technologies associated with that program will be adopted by participants. In some cases, enrollment may be high, but technology adoption for one or more measures may be lower than expected because of lack of information about the full benefits of the technology. If the program expects that a NEEs evaluation can produce information that could be used to increase adoption rates such that the program would be more cost-effective at obtaining its energy impact goals (excluding the value of the NEEs), then a NEEs evaluation may be appropriate. If the program theory indicates that it relies on the value of the non-energy benefits, then the decision process moves back to step two. If the step five decision is that "no," the program theory does not rely on the presence of the non-energy effect to be successful, then the decision moves to step six.

Step Six

This step recognizes that the CPUC may identify NEEs that are in need of evaluation. This step allows that event to occur without influencing the energy impact evaluation or the program process evaluation efforts. If the CPUC has identified a specific non-energy effect that it wants evaluated, it will convey that need to the program administrators, contractors or evaluators as appropriate.

Step Seven

This step is a "safety step" in the non-energy benefit process that requires the CPUC to specifically approve the use of evaluation funds for the purpose of researching CPUC identified NEEs. This allows the CPUC to have more control over the dollars spent for non-energy research to support a CPUC information need and direct control over the type of research to be conducted. If the CPUC has approved the use of evaluation funds to conduct the non-energy effect research, then a NEEs evaluation can be planned and conducted.

Conducting the Non-Energy Effects Evaluation

If any of the above detailed decision steps identify the need for a NEEs evaluation, the program can budget and plan for the implementation of the evaluation by an independent program evaluator that meets the skills requirements presented in this chapter and as is appropriate for the type of NEEs research to be conducted.

Chapter 12: Uncertainty

Introduction and Key Issues

This chapter discussed three topics that are especially relevant to evaluation contractors and policy makers alike:

- Bias and statistical precision,
- Techniques for assessing the savings in a portfolio of programs, and
- Allocation of evaluation resources to programs.

Periodically these topics have been discussed in energy efficiency evaluation but they are a not a standard part of program evaluations or evaluation planning.
Since evaluation is designed to reduce uncertainty, this chapter addresses several issues that are pivotal to successful evaluation. It applies to all types of evaluations: impact, process, market, education, etc. Consideration of areas of uncertainty, level of precision or confidence in the results, and potential bias occurs in all high quality evaluation plans and within the analysis and reporting of all evaluation results.

This chapter is especially important for policy makers. Evaluators and evaluation reviewers also need to be familiar with this material. The sections discussing "How Bias Arises" and "Reporting Guidelines" will give program managers and implementers an overview of some of the issues that arise in evaluation studies.

The chapter makes the case that the research design and data collection methodology are just as important to the quality and accuracy of an evaluation study as the size of the sample and the choice of analytical techniques. A sound evaluation study depends on all these elements but budget limitations may force tradeoffs. Unfortunately, it is generally difficult to quantify these tradeoffs so effective evaluation requires a combination of science and experience.

Skills Required for the Uncertainty Analysis Tasks

This chapter is written from a statistical science perspective to understand relative precision and bias in a quantitative manner. Either making assessments or conducting quantitative analysis in this manner generally requires knowledge of basic graduate statistics or equivalent study and experience with a mentor in this area.

Mitigating the risk of bias requires an understanding of research or evaluation design. Relevant graduate courses or similar training by an experienced professional could be from within many of the social sciences (sociology, economics, psychology, policy

analysis, public health, etc.). Reducing potential measurement bias and response bias in surveys also requires training and experience in survey and instrument design.[275]

Bias and Statistical Precision

An evaluation study is designed to assess one or more parameters such as the gross and net annual savings of all projects undertaken in a program. A sound evaluation study should produce estimates that are free of any substantial bias and should accurately characterize the statistical precision of the results.

Most evaluators are familiar with the most commonly used techniques for reporting statistical precision: error bounds, confidence intervals, and relative precision. Some of us may not realize, however, that these techniques are very misleading if there is substantial bias in the findings. For example, it is misleading to report that the savings of a program have been measured within ±10% at the 90% level of confidence if there is a plausible risk that the results are biased by 25%.

Unfortunately, it is usually extremely difficult to objectively quantify the magnitude of the bias or even its direction. Occasionally, there is an opportunity to carry out a special methodological study designed to assess the risk of bias from one or more particular factors. But generally the best the users of an evaluation study can do is to look for certain aspects of the implementation of the study that could increase the risk of bias – such as

Bayesian Statistics

This chapter has been developed from a *classical statistics* paradigm. At present, classical statistics is the foundation for most evaluation work. *Classical statistics* builds on the frequency school of probability and statistical inference. *Bayesian statistics* presents an alternative paradigm to classical statistics that builds on subjective probability.

When samples are large and there is little danger of bias, classical confidence intervals are generally equivalent to Bayesian inference. The authors have deliberately steered clear of classical concepts, such as p-values, that are disputed under the Bayesian approach. Therefore this chapter's discussions should be of value from either a classical or Bayesian point of view.

Classical and Bayesian statistical methodology generally diverge when there is limited experimental information. For example, a Bayesian approach might be used to develop a subjective assessment of the amount of bias. This type of analysis would extend the methodology discussed in this chapter. The subjective foundation of the Bayesian methodology is its major potential disadvantage in evaluation applications. In conveying the findings of an evaluation study, it is important to clearly identify any subjective inputs used in the analysis.

[275] There are many sources of information on survey design. The Survey Research Methods section of the American Statistical Association has links to university courses on the subject as well as an introductory set of pamphlets called "What is a Survey?" (American Statistical Association 1995). Another source of information is the Research Methods Knowledge Base (Trochim 2001). Other references are: *Survey Methodology* (Groves et al. 2004); and *Methods for Testing and Evaluating Survey Questionnaires* (Presser et al. 2004).

poor response rates in a survey, deliberate exclusions of sample projects, poor measurement techniques in a field study, instability of coefficients in a regression analysis, etc.

In a high quality evaluation those implementing the study would strive to mitigate the risk of bias and to honestly report any circumstances about the study that might increase the likelihood of bias. Unfortunately, it usually takes extra time and money to reduce the risk of bias, and the usual measures of the statistical precision of the results may not be improved at all. For example, in order to reduce the risk of non-response bias in a telephone survey, a substantial investment may be needed in more extensive training for the surveyors, more call backs, and perhaps to offer a financial incentive to each respondent.

It may be tempting to accept a higher non-response rate and divert these resources to a larger sample size since this strategy will almost certainly give a narrower confidence interval. This strategy can seriously compromise the integrity of a study. To make appropriate judgments in planning and executing sound evaluation studies and in interpreting their results, evaluators, reviewers, and those using evaluation results need to understand what bias is, how it can arise, and how it can undermine an evaluation study.

Basic Definitions

Let μ denote a particular parameter of interest such as the true average savings per project in the population of projects to be evaluated and let $\hat{\mu}$ denote the estimate of μ to be produced by a particular evaluation study. Before the evaluation study has been completed, $\hat{\mu}$ can be regarded as a random variable. For any given value of μ, $\hat{\mu}$ will have a probability distribution that reflects sampling variability, measurement error, response error, variation due to randomness in the model assumed to generate the data, etc. This is called the sampling distribution of $\hat{\mu}$.

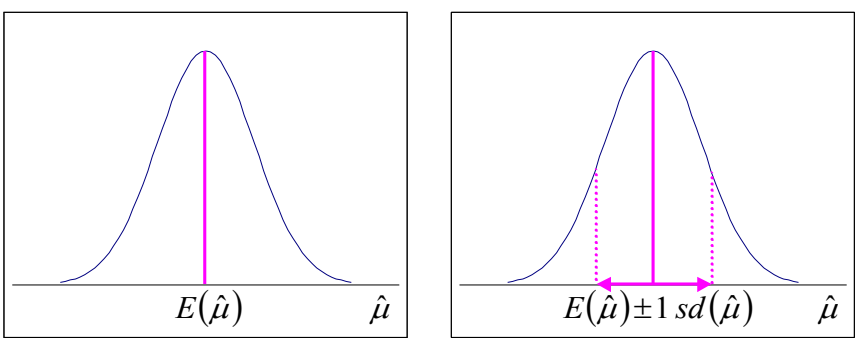

Figure 12.5: The Sampling Distribution of an Estimator

The two key parameters of the sampling distribution of $\hat{\mu}$ are the expected value of $\hat{\mu}$, denoted $E(\hat{\mu})$, and the standard deviation of $\hat{\mu}$, denoted $sd(\hat{\mu})$. Figure 12.5 shows an example of the sampling distribution of an estimator $\hat{\mu}$, together with the expected value

and standard deviation. In this example, the sampling distribution has been assumed to follow the normal probability distribution. The left side of Figure 12.5 shows the expected value of $\hat{\mu}$, $E(\hat{\mu})$. The right side of Figure 12.5 shows the one standard deviation interval around the expected value. If the sampling distribution of $\hat{\mu}$ follows the normal probability distribution,[276] then the probability is about 68% that $\hat{\mu}$ will be in the interval $E(\hat{\mu}) \pm 1\ sd(\hat{\mu})$, i.e. in the interval denoted by the two-headed arrow shown in the right hand side of Figure 12.5.

Assuming a normal probability distribution, the probability is about 90% that $\hat{\mu}$ will be within 1.645 standard deviations of the expected value of $\hat{\mu}$, i.e., in the interval $E(\hat{\mu}) \pm 1.645\ sd(\hat{\mu})$. This is the usual basis for a 90% confidence interval. In particular, the quantity $1.645\ sd(\hat{\mu})$ is called the expected error bound at the 90% level of confidence. Assuming that μ is not equal to zero, the expected relative precision at the 90% level of confidence is defined to be the expected error bound divided by μ: i.e.,

$$rp = 1.645\ \frac{sd(\hat{\mu})}{\mu}$$

The estimator $\hat{\mu}$ is said to be an unbiased estimator of the population parameter μ if and only if the expected value of $\hat{\mu}$ is equal to μ, i.e., if and only if $E(\hat{\mu}) = \mu$. The left side of Figure 12.6 illustrates an unbiased estimator. In this case, although any particular realization of $\hat{\mu}$ will probably be either larger or smaller than μ, in repeated sampling the average value of $\hat{\mu}$ will be equal to μ.

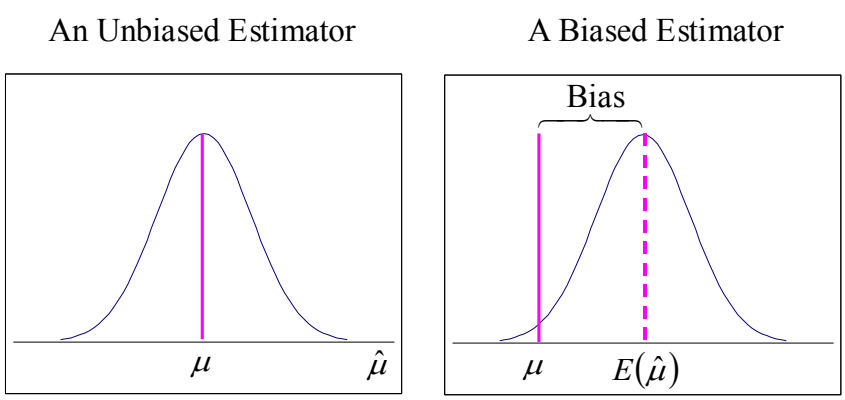

Figure 12.6: Unbiased and Biased Estimators

[276] This will generally be approximately true for a well-designed evaluation study. The Central Limit Theorem and its various extensions suggest that if the sample is large the sampling distribution should generally be approximately normal. Even if the sample is moderate or small, this is generally true if an efficiently stratified sample design has been used. If a regression analysis is carried out with a small sample, it may be necessary to use special techniques, such as a Box Cox transformation, to improve normality. All of the material in this section can be found in most standard upper level undergraduate and graduate statistics textbooks.

If $\hat{\mu}$ is a biased estimator, the bias of $\hat{\mu}$ is the difference between $E(\hat{\mu})$ and μ. The right side of Figure 12.6 illustrates a biased estimator. In this example, $\hat{\mu}$ is so badly biased that the probability is high that $\hat{\mu}$ will be greater than μ. In other words, the evaluation study is very likely to overstate the actual value of μ.

A small amount of bias is usually not of concern, but the bias is serious if it is large compared to the standard deviation $sd(\hat{\mu})$. Assuming that μ is greater than zero, the relative bias is defined to be:

$$rb(\hat{\mu}) = \frac{E(\hat{\mu}) - \mu}{\mu}$$

The bias is serious if the relative bias is large compared to the expected relative precision. In this case, the relative precision and the error bound are likely to be misleading because they fail to reflect the bias.

When there is serious bias, a theoretically more suitable measure of statistical precision is the root mean square error of $\hat{\mu}$, given by

$$rmse(\hat{\mu}) = \sqrt{(E(\hat{\mu}) - \mu)^2 + sd(\hat{\mu})^2}$$

The root mean square error combines both the bias and the standard deviation, and is often used in specialized studies, especially simulation studies. If the root mean square is available, then the generalized error bound can be defined to be 1.645 $rmse(\hat{\mu})$. If in addition $\mu > 0$ then the generalized relative precision can be defined to be 1.645 $rmse(\hat{\mu})/\mu$. In most evaluation studies, however, it is difficult to quantify the bias so the root mean square error is not reported. Generally, the reported error bound and relative precision ignore any possible bias.

Table 12.5: Example of a Biased Estimator

True Value	1000	
Expected Value	1250	
Standard Deviation	100	
Bias	250	1250 - 1000
Relative Bias	0.25	250 / 1000
Root Mean Square Error	269	sqrt($100^2 + 250^2$)
Generalize Error Bound	443	(1.645) (269)
Generalized Relative Precision	0.44	(1.645) (269) / 1000
Reported Error Bound	165	(1.645) (100)
Reported Relative Precision	0.16	165 / 1000

The example given in Table 12.5 illustrates some of these ideas. In the example it is assumed that the true parameter of interest is 1,000 units. The estimator produced by the

study is assumed to have a normally distributed sampling distribution with an expected value of 1,250 units and a standard deviation of 100 units. In this case, the probability is about 90% that the estimator will be in the interval $1{,}250 \pm 165$, i.e., from 1,085 to 1,415, although the true value is 1,000. So the estimator is likely to be too high.

This estimator is expected to perform so badly because it has a bias of 250 units. In this case the root mean square error is 269 units, which would give a generalized error bound of 443 units and a generalized relative precision of 44%. If, as is the common practice, the bias were ignored in planning the study or in reporting the statistical precision, the expected error bound would be 165 units and the expected relative precision would be 16%. In short, if bias is ignored, the estimator would be expected to be within 165 units of the true parameter, whereas in actuality, the estimator is likely to be between 85 and 415 units higher than the true parameter.

How Bias Arises

Most evaluation studies are free of substantial bias, as long as the underlying assumptions are accurate. But bias can arise when the assumptions break down in the implementation of the study. This breakdown of assumptions can be spotted, but only by thoroughly understanding the underlying assumptions and being close to the details of the study. Thus the principal investigators are generally in the best position to discuss any risks of bias.

Assessing all of the various risks is a challenge, even to the most experienced study director. The violation of almost any underlying assumption can lead to bias. So there are as many potential sources of bias as there are assumptions in a study.

Some of the most common sources of potential bias in energy efficiency evaluation include the following:

- **Non-response and other forms of selection bias.** As discussed in the Sampling chapter, a key assumption in statistical sampling is that the sample is randomly selected from a complete and accurate sampling frame following a suitable sampling plan. Selection bias can arise if the sampling frame or sampling plan excludes a particular part of the target population. Non-response bias can occur if any of the designated sample projects are dropped or replaced for almost any reason, such as refusal to participate, technical difficulties, etc.[277] [278] In some engineering field studies, the sample is sometimes subjectively selected. This practice not only raises the risk of bias but also makes it nearly impossible to assess the statistical precision of the estimator.

[277] Occasionally, the target population is a subset of the sampling frame that can only be identified from the study itself. This should not be confused with non-response.

[278] All estimators commonly used in survey sampling will generally be essentially unbiased as long as the case weight applied to each sample project is equal to the reciprocal of the probability that the project is included in the sample, i.e., the true inclusion probability, considering all factors including the sample design, any non-response, technical exclusion, etc.

- **Measurement bias.** Any consistent difference between the value of a measurement and the underlying property can lead to measurement bias. There are as many causes of measurement bias as there are measurement techniques, e.g., poorly worded survey questions, inability or reluctance of the respondent to provide the desired information, inappropriate scoring of survey responses, oversimplified engineering models, inaccurate calibration of measurement instruments, etc.
- **Erroneous Specification of the Statistical Model.** An erroneous statistical model can threaten a statistical evaluation study just as seriously as an erroneous engineering model or badly implemented sampling plan can threaten a field study. Many evaluation studies make heavy use of statistical regression analysis. Any violation of the assumptions of regression analysis can introduce the risk of seriously misleading findings.
- **Choosing an Appropriate Baseline.** In many evaluation studies, there is more uncertainty about the baseline energy consumption than about the post-participation energy consumption. In evaluating an AC rebate program, for example, it is necessary to decide whether the program participant would have retained the existing air conditioner in the absence of the program or replaced it anyway. If the latter, how efficient would the replacement unit have been in the absence of the program? Similarly, in evaluating a new construction program, is it appropriate to assume that the building would have been built to a particular energy code?
- **Self Selection of Program Participants.** Most common methods of assessing free ridership compare the behavior of program participants to a sample of non-participants. The non-participants are generally assumed to reveal what the behavior of the participants would have been in the absence of the program. These methods can provide biased estimates of free ridership if the underlying assumptions are not accurate, e.g., if there is a correlation between interest in conservation and program participation.
- **Misinterpretation of Association as Causal Effects.** Most introductory texts on regression and econometrics discuss the danger of interpreting a regression coefficient as an unbiased estimator of the causal effect of the explanatory variable on the dependent variable. This type of bias is always a danger when regression and other statistical techniques are used to analyze data that has not been produced in a controlled experiment. These concerns are relevant in those impact evaluation studies that seek to estimate the true impact of a program on energy consumption by analyzing the tracking and billing data collected in the program itself.

The research design literature sometimes discusses various threats to the validity of a study using terms such as the following:[279]

[279] Two examples of this literature are: *Applied Social Research* (Rubin 1983) and *Evaluation* (* Weiss 1998).

- **Construct validity.** The degree to which the inferences are valid based upon the way in which the underlying construct being tested is operationalized in the research design.[280]
- **Statistical validity.** Using an incorrect statistical methodology in the analysis, e.g., using ordinary least squared regression analysis when there is substantial random error in an explanatory variable.[281]
- **Internal validity**. Inadequate controls in the research setting (or quasi-experimental design), instrumentation and selection biases.
- **External validity**. Can the evaluation findings be applied to the entire treatment group? [282]

These issues are closely related to the sources of bias that have been discussed such as non-response, measurement error, error in specifying the statistical model, etc.

Special Techniques to Assess Bias

Since bias can be associated with the breakdown of almost any underlying assumption, it is difficult in a typical evaluation study to quantify all potential sources of bias or even to discuss all the different ways that bias might arise. The analysis itself is guided by the underlying assumptions and it is usually difficult to carry out the analysis that would be appropriate in the absence of these assumptions or under alternatives to them. Moreover, alternative assumptions usually lead to different conclusions.[283] Usually the director of an evaluation study must decide which assumptions are most plausible and to emphasize the corresponding results.

However, when it is feasible, a high quality study would report the results obtained under all plausible alternative assumptions. This is generally more practical in a regression study than in a survey or field study. This often takes the form of reporting regression models with alternative choices of explanatory variables and perhaps alternative ways of dealing with outliers.

In surveys and other applications of sampling, it usually is difficult to assess how much bias might be caused by non-response since there usually is limited data for the non-respondents. If simple random sampling has been used, it sometimes is useful to compare the tracking estimates of savings of the respondents to the population since a

---

280 See *The Research Methods Knowledge Base* (Trochim 2001).

281 Bayesian statisticians would include the reliance on p-values. This methodology may lead to false conclusions that can appear quite differently in Bayesian analysis. For example, see the article: *Facts versus Factions: the use and abuse of subjectivity in scientific research* (Matthews 1998).

282 A classic graduate text dealing with research design is *Experimental and Quasi-Experimental Designs for Research* (Campbell and Stanley 1963). They present extensive discussion on various quasi-experimental designs and also discuss eight types of threats to internal validity and four to external validity.

283 One indicator of a serious problem is if plausible alternative assumptions yield confidence intervals will little or no overlap.

significant difference would indicate that the sample is biased. Another approach to investigating the effect of non-response bias is to carry out a special study of a sample of the non-respondents. In practice, however, it usually is difficult to collect meaningful information from non-respondents.

If stratified ratio estimation is used to estimate the program realization rate (the ratio of evaluation estimated savings to program tracking estimated savings, the degree to which estimated savings are "realized"), then it is generally difficult to estimate the effect of non-response bias. In theory, if an efficiently stratified sample design is used, if the case weights are the reciprocal of the true inclusion probabilities, and if there is negligible measurement bias, the stratified ratio estimator will be essentially unbiased. Bias can arise if the case weights fail to accurately reflect the inclusion probabilities due to non-response or other selection factors. In principle, one or more statistical models could be developed for the probability of response, the response probabilities could be factored into the case weights, and the sensitivity of the estimates to the changes in the case weights could be explored. But this work would divert resources from the evaluation study itself. Usually it seems better to devote these resources to minimizing non-response in the first place.

The Sampling chapter discusses the use of a ratio model to assist in the development of an efficient sample design. It is important to note that the stratified ratio estimator is still generally unbiased even if the assumed model is inaccurate – provided that the sample design has been followed. Moreover, the calculated statistical precision will generally be a good guide to the statistical precision of the stratified ratio estimator. In other words, the stratified ratio estimation approach is designed to provide better protection against model-specification bias than econometric methods – provided that the sample design is accurately followed.

However, these results assume large samples whereas in many evaluation studies stratified ratio estimation is used with small or moderate sample sizes. So these results may not be entirely applicable. A special simulation study could be carried out to assess whether stratified ratio estimation techniques are seriously biased when used with moderate sample sizes, especially when the ratio model is not accurate. In practice, however, it is often hard to draw general conclusions from this type of simulation study.[284]

Sometimes a special study can be carried out to assess measurement bias. If a particular aspect of a field study is challenged, an independent study can sometimes be implemented to validate the data collection and analysis. For example, short-term metering might be used to validate the accuracy of reported schedules of use in a small sample of projects. Often the study is carried out in a sub-sample of the original sample so that the new results can be compared to the information collected in the original study.

[284] One large simulation study showed that stratified ratio estimation had little or no bias in several load research applications. See *Sample Designs for Load Research: The Bootstrap Comparison Procedure*, (McCarthy, Thibodeau and Galford, 1985).

Ledyard reported an example in which end-use metering was used to verify the operating hours of lighting measures reported in an onsite survey.[285]

An excellent analysis of measurement error can be found within a billing analysis performed by Rick Ridge of Ridge & Associates for Pacific Gas & Electric Company. In a statistically-adjusted engineering (SAE) billing analysis, the estimated realization rate can be biased if the tracking system estimates of savings used in the analysis have random error and systematic error.[286] This paper cited earlier work[287] that demonstrated that measurement error in a single explanatory variable would cause the coefficient to be downwardly biased. With multiple explanatory variables, the coefficients can be biased in unknown directions. A Monte Carlo simulation was done utilizing specific systematic error of plus and minus 10% and several forms of random error.

This work verified that the SAE methodology worked well with systematic error; the coefficient adjusted for the systematic error as expected. The work showed, however, that with random error of 15% or greater, the SAE analysis performed worse than using a dummy variable in the billing analysis, evidently because the dummy variable is free of error. The importance of this was shown in the billing analysis of PG&E's 1994 Commercial HVAC program. The SAE realization rate was 0.35 using the tracking estimates of savings. When the tracking estimates were restricted to those expected to have less measurement error, the realization rate was found to be 0.50. However, with the dummy variable approach that is free of measurement error, the realization rate was 0.92. Given these results, Ridge recommended that using a dummy variable may be preferable to the SAE methodology for HVAC impact analyses unless the engineering priors are free of significant random error. The dummy variable coefficient may not be as precise but it is free of the bias associated with the SAE coefficient so that it appears to be more reliable.[288]

Another interesting assessment of the importance of bias versus precision can be seen in a simulation study on this topic conducted by Richard Sonnenblick and Joseph Eto from Lawrence Berkeley Laboratory in 1995. Their work found that "imprecision in the cost of conserved energy was significant for programs with mean TRC test ratios close to one, while higher ratios guarantee cost-effectiveness even with considerable estimate imprecision." They stated that "a 90/10 criteria for precision seems excessive for most programs". On the other hand, bias in savings estimates had a greater impact where bias could threaten cost-effectiveness for programs with ratios approaching 2.0. From this, they concluded that "much of the contemporary concern with precision should be redirected to examine bias in evaluation estimates".[289]

---

285 Tom Ledyard (2003).

286 SAE billing analysis is described in the Billing Analysis section of Chapter 6: Impact Evaluation.

287 *Econometrics: Basic and Applied* (Johnson et al. 1967).

288 "Errors in Variables: A Close Encounter of the Third Kind." (* Ridge 1997).

289 "Calculating the Uncertainty in Estimates of DSM Program Cost-effectiveness." (* Sonnenblick et al. 1995). Quote on page 767.

Reporting Potential Bias

Under ideal conditions, all common evaluation approaches yield essentially unbiased estimators – but conditions are rarely ideal. Therefore it seems likely that most evaluation studies have some amount of bias. Unfortunately, it is extremely difficult to objectively quantify the possible bias. Given that bias is generally immeasurable; evaluation reports generally focus on precision, which is measurable.

The effect of bias, however, can be quite important if it exists. You could be quite precise about the answer, but it would be the wrong answer. Given this, potential bias is critically examined in any high quality evaluation planning and evaluation analysis. Greater evaluation resources spent for larger sample sizes to achieve greater precision will only provide more misleading results if the study is biased. In planning and implanting evaluation studies, addressing sources of bias is as important as increasing precision.

As indicated above, the principle investigators are generally in a better position than anyone else to discuss the risk of bias. Therefore, the evaluators are best able to help the reader assess the danger of bias. A sound evaluation would at a minimum discuss all relevant sources of bias and potential threats to validity – non-response, measurement, model specification, self-selection, etc. When appropriate, the report would also discuss the results obtained under plausible alternative models. [290]

Reducing the Standard Deviation of an Estimator

The standard deviation of an estimator depends on all aspects of a study including the general experimental plan, the sampling plan, the sample size, the data collection and verification techniques, and the statistical analysis methodology. In planning an evaluation, it is necessary to consider all of these factors.

For example, in the chapter on sampling it is noted that the standard deviation of a stratified ratio estimator is often much smaller than the standard deviation of the sample mean of a simple random sample of observations. The improvement in statistical precision comes partly from changing the focus from the average actual savings per project to the realization rate, and partly from taking advantage of the supporting information provided by the tracking system to develop an effective sample design. Stratified ratio estimation can often be carried out at little added cost compared to simple random sampling, but with a substantial improvement in statistical precision or reduction in the sample size. The reduction in the sample size, in turn, can allow greater attention to mitigating bias.

Similarly, if a regression analysis of billing data is to be used to evaluate a retrofit program, it is necessary to consider all aspects of the study, e.g., how many months of

[290] A good discussion of reporting requirements is provided in "Quality Assurance Guidelines for Statistical, Engineering, and Self-Report Methods for Estimating DSM Program Impacts." (* CADMAC 1998).

billing data should be required before and after the retrofit, how closely the non-participants should be matched to the participants, how much the billing data should be cleaned prior to analysis, how should weather data be used in the analysis, whether a survey should be carried out to provide information about other factors affecting energy consumption during the study period, etc. All of these issues are likely to affect the standard deviation of the resulting estimate of savings. In particular, to the extent that variation can be controlled without increasing the risk of bias, the statistical precision can often be substantially improved from an effective experimental approach, often with little added cost.[291]

Integrating the Results from Multiple Evaluation Studies

One of the paramount objectives of evaluation is to provide an accurate assessment of the total savings achieved by a set of programs. For example, policy makers may want to talk about the actual savings and cost-effectiveness of:

- All conservation programs taken together compared to supply-side options,
- Residential programs compared to non-residential programs,
- New construction programs compared to retrofit programs, and
- Third-party programs compared to utility-delivered programs.

To make these comparisons[292] it is necessary to assess the aggregate savings and the corresponding statistical precision of various portfolios of programs and to assess whether there is a significant difference in the savings of programs. To do this, it is usually necessary to combine the results of two or more independent evaluation studies. In addition, policy makers may have occasion to contrast and combine the results of independent studies of the same program. The following techniques can be used individually or in combination to address most situations.[293]

- Estimating the Total Savings of a Portfolio of Programs,
- Contrasting the Results of Two Independent Studies,
- Pooling two Statistically Independent Estimators of the Same Parameter, and
- Chaining the Results of Two Evaluation Studies.

---

[291] Special techniques may be needed to correctly estimate the standard deviation of a complex estimator when there are several sources of variation such as sampling, modeling, etc. Chapter 11 of *Model Assisted Survey Sampling* (Sarndal et al. 1992) discusses several methods that are sometimes useful in evaluation, including balanced half sample methods, the Jackknife technique, and Bootstrap analysis. There is a rich literature on methods such as these.

[292] However, any such comparison should take into consideration that some energy efficiency programs may be handicapped by participation requirements. These and other factors may need to be taken into account to ensure an appropriate comparison, i.e., an "apples to apples" comparison.

[293] These and other methods were discussed in the context of evaluation in "A Methodology For Integration of Evaluation Studies" (Wright et al. 1993).

Each of these techniques will be discussed in the following sections. Throughout this discussion, it is assumed that each study has provided an unbiased estimate of the parameter of interest and an accurate error bound or relative precision. A later section in this chapter will also show how to calculate the error bound or relative precision when a p-value is given.

Estimating the Total Savings of a Portfolio of Programs

Suppose the total savings from a set of programs is desired. Consider the following assumptions:

1. There are no interactions between the programs,
2. Each of the individual programs has been evaluated independently,
3. Each evaluation has provided an unbiased estimator of the actual savings of the corresponding program, and
4. Each evaluation has provided the corresponding error bound.

Then the total savings of the portfolio can be estimated as the sum of the savings of each program. This is an unbiased estimator of the total savings. The corresponding error bound is the square root of the sum of the squares of the error bounds of each program.

Table 12.6 provides an example. In this hypothetical example, a portfolio of four programs is being examined. Programs A and B are relatively small, and each have estimated savings of 50,000 units. Program C is twice as large as A and B, with an estimated savings of 100,000 units. Program D is five times larger than program C, with an estimated savings of 500,000 units. It is assumed that there are no interactions between these programs so the total savings of the portfolio can be estimated as the sum of the estimated savings of the four individual programs, i.e., 700,000 units.[294] [295] The evaluation contractors also need to have accurately reported the error bound of the estimated savings of each of the four programs at the 90% level of confidence, as 1.645 times the actual standard deviation of the estimated savings. These results are shown in the last column of Table 12.6. Another important assumption is that the four estimates are mutually independent random variables.[296] The corresponding relative precision is shown for completeness.

[294] Two programs could interact, for example, if they both provided the same measures to the same participants. A more subtle form of interaction could occur if one program affected the freeridership of another program by making a technology more familiar in the market.

[295] The method of pooling generally can be used for both gross and net savings.

[296] The result is a simple consequence of (a) the fact that the standard deviation of a sum of statistically independent random variables (e.g., the estimated savings of each program) is the square root of the sum of the squares of the standard deviations of each of the random variables, and (b) the error bound being defined as 1.645 times the standard deviation.

Table 12.6: Pooling the Savings of a Portfolio of Programs

Program	*Estimated Savings*	*Relative Precision*	*Error Bound*
A	50,000	13%	6,264
B	50,000	10%	4,852
C	100,000	6%	6,264
D	500,000	4%	18,792
Total	700,000	3%	21,334

In this example, the error bound associated with the total estimated savings is easily calculated as the square root of the sum of the squared error bounds of each of the four programs: $\sqrt{6{,}264^2 + 4{,}852^2 + 6{,}264^2 + 18{,}792^2} = 21{,}334$. The 90% confidence interval for the savings of the portfolio is $700{,}000 \pm 21{,}334$ units. The relative precision of the total portfolio savings is $21{,}334/700{,}000 = 3\%$.

Note that in this example, the relative precision of the total savings is better than the relative precision of any of the four programs. This is a direct consequence of our assumption that the four estimators are mutually independent and free of bias.

Contrasting the Results of Two Independent Studies

A common question is whether there is a significant difference between two programs. For example, a policy maker may wish to compare the realization rate achieved by two different programs. Assume that there are statistically independent and unbiased estimates of the realization rates of each of the two programs, and that the associated error bounds have been provided. Then the error bound associated with the difference between the two realization rates can be calculated as the square root of the sum of the squared error bounds of each of the two programs. If the observed difference is greater than the error bound associated with the difference, then the difference is statistically significant at the 90% level of confidence. Otherwise the observed difference is not statistically significant.[297]

Table 12.7 shows an example. In this example, independent evaluation studies have been carried out for two programs, A and B. Program A was found to have a realization rate of 0.80 whereas program B was found to have a realization rate of 0.70. Each of these results is assumed to be unbiased, and the error bounds are 0.12 and 0.10 respectively. Given these results, do the two programs have a *significantly* different realization rate? In other words, is the observed difference statistically significant?

[297] This result is a simple consequence of (a) the fact that the standard deviation of the difference between two statistically independent random variables (e.g., the estimated savings of each program) is the square root of the sum of the squares of the standard deviations of each of the random variables, and (b) the error bound is 1.645 times the standard deviation.

To address this question, the error bound associated with the difference between the two realization rates is calculated as the square root of the sum of the squared error bounds of each of the two programs: $\sqrt{0.12^2 + 0.10^2} = 0.16$. This implies that the 90% confidence interval for the difference in the two realization rates would be 0.10 ± 0.16 or from -0.06 to +0.26. Since the error bound is greater than the difference itself, it should not be concluded that the difference is statistically significant in this example.

Table 12.7: Contrasting Program A and B

Program	*Realization Rate*	*Relative Precision*	*Error Bound*
A	0.80	15%	0.12
B	0.70	14%	0.10
Difference	0.10		0.16

Table 12.8 shows a second example. In this case the realization rate of Program A is being compared to that of Program C. As the table shows, the difference in the two realization rates is 0.25 and the error bound is 0.18. In this case the difference is greater than the error bound and so the difference is statistically significant.

Table 12.8: Contrasting Program A and C

Program	*Realization Rate*	*Relative Precision*	*Error Bound*
A	0.80	15%	0.12
C	0.55	25%	0.14
Difference	0.25		0.18

The preceding methods can also be used to contrast the results of two statistically independent studies of the same program. This type of application is discussed next.

Pooling Two Statistically Independent Estimators of the Same Parameter

Sometimes, two statistically independent evaluation studies are carried out for the same program. For example, the first study might use billing analysis techniques and the second study might use on-site audits and engineering analysis of a sample of projects selected from the program. The first step is to check to see if there is a statistically significant difference between the results of the two studies, using the method previously discussed.

If the difference is not statistically significant, it may be plausible to assume that both studies have provided statistically unbiased estimates of the savings of the program.[298] In this case, the results may be combined into a single estimate.

[298] If there is a statistically significant difference between the results, it is implausible that both estimators are unbiased. In this case it is more reasonable to assume that one or both of the studies is biased. In this case, a substantial effort may be required to understand the source of the bias.

Assume that both studies have provided an unbiased estimator of a given parameter, such as the average savings per project, and that the error bound is known for each of the two estimates. What is the best way to combine the two estimates into a single estimate, and what is the associated error bound? The answer lies in the following rule:

If two estimators are both unbiased estimators of a given parameter, then any weighted average of the two estimators is also an unbiased estimator. To provide the minimum error bound, each of the two estimators should be weighted in proportion to the reciprocal of the squared error bound of the estimator. The error bound of the result is the square root of the reciprocal of the sum of the weights.

It may be helpful to express this rule algebraically. Let $\hat{\mu}_1$ and $\hat{\mu}_2$ be two independent, unbiased estimators of the average savings per project μ of a given program. Let eb_1 be the error bound of $\hat{\mu}_1$ and let eb_2 be the error bound of $\hat{\mu}_2$. The weights are defined as: $w_1 = 1/(eb_1)^2$ and $w_2 = 1/(eb_2)^2$. Then the pooled estimator of μ is calculated as the weighted average

$$\frac{w_1\ \hat{\mu}_1 + w_2\ \hat{\mu}_2}{w_1 + w_2}.$$

The corresponding error bound is $\sqrt{\dfrac{1}{w_1 + w_2}}$.

Table 12.9 and Table 12.10 give a numerical example. In this example two independent studies have yielded estimates of the average savings per project of a given program. The first study indicated that the savings were 600 units with an error bound of 100, and the second study indicated that the savings were 500 units with an error bound of 75. Before pooling the two results, it is generally wise to examine the difference between the two estimates, as shown in Table 12.9. The difference between the two estimates is 100 units with an error bound of $\sqrt{100^2 + 75^2} = 125$, so the difference is not statistically significant.

Table 12.9: Contrasting Two Estimators

Estimator	*Average Saving*	*Relative Precision*	*Error Bound*
1	600	17%	100
2	500	15%	75
Difference	100		125

Therefore it is plausible to assume that the two results are both unbiased estimators of the true average savings of the program. Under this assumption, it is useful to pool the two estimates, as shown in Table 12.10. To carry out this calculation, the weight associated

with each of the two estimators is calculated as $w_1 = 1/100^2 = 0.0001$ and $w_2 = 1/75^2 = 0.0001778$. Notice that the second weight is larger than the first since the second estimator has a smaller error bound.

Then the pooled estimate is calculated using:

$$\frac{w_1\ 600 + w_2\ 500}{w_1 + w_2} = 536 \text{ units.}$$

Finally the associated error bound is calculated as $\sqrt{\frac{1}{w_1 + w_2}} = 60$ units. Since the pooled estimate is combining the information from the two studies, the error bound of the pooled estimate is smaller than the error bound of either of the two individual estimators.

Table 12.10: Pooling Two Estimators

Estimator	*Average Saving*	*Relative Precision*	*Error Bound*	*Weight*
1	600	17%	100	0.0001000
2	500	15%	75	0.0001778
Pooled	536	11%	60	0.0002778

Chaining the Results of Two Evaluation Studies

The method of chaining is relevant if the estimators of two factors are independent and unbiased, such as the gross realization rate and the net-to-gross ratio, and the product of the two factors is an estimate of interest. The relative precision of each of the two estimators is used to obtain the relative precision of the chained result. The relative precision associated with the product of the two factors is approximately equal to the square root of the sum of the squared relative precisions of each of the two factors.[299]

Table 12.11 shows an example. Assume that the gross savings of a program has been estimated to be 100,000 units with a corresponding relative precision of 15% at the 90% level of confidence. The net-to-gross ratio (NTGR) has been estimated in an independent study to be 0.8 with a relative precision of 15%. Then the net savings of the program is estimated to be 80,000 units. The corresponding relative precision is $\sqrt{0.10^2 + 0.15^2} = 0.18$. The corresponding error bound is $80{,}000 \times 0.18 = 14{,}422$ units.

[299] This equation can be derived by using Taylor's theorem to find a linear approximation to the product of the two factors near the expected values of each of the two factors, as explained further in our discussion on the propagation of uncertainty, below. (Most graduate and advanced statistics or econometrics textbooks use and refer to Taylor's theorem as a method to derive a linear approximation of a nonlinear function. For example see *Elements of Econometrics* (Kmenta 1971), pages 399-400.

Table 12.11: Chaining

Factor	*Estimate*	*Relative Precision*	*Error Bound*
Gross Savin	100,000	10%	10,000
NTGR	0.8	15%	0.12
Net Savings	80,000	18%	14,422

Converting a P-Value to a Relative Precision or Error Bound

When regression analysis is used to estimate the savings of a program, a so-called p-value is sometimes used to assess the statistical precision of the estimated savings. Most standard regression software packages report the estimate itself, the standard error, a statistic called the t-value, and the p-value. If the p-value is less than 0.10, the corresponding estimate is usually regarded as statistically significant at the 90% level of confidence.

To ensure that the results can be easily compared to the evaluation results for other programs, the statistical precision should also be reported as an error bound and relative precision at the 90% level of confidence. The error bound can usually be calculated as 1.645 times the reported standard error. The relative precision can be calculated as the error bound divided by the absolute value of the estimate.[300]

Table 12.12: Converting a P-Value

Variable	*DF*	*Parameter Estimate*	*Standard Error*	*t Value*	*Pr > |t|*		*Error Bound*	*Relative Precision*
Intercept	1	493	163	3.03	0.0028		268	0.54
Pre-use	1	0.89	0.02	48.22	<.0001		0.03	0.03
Business	1	4,321	984	4.39	<.0001		1,618	0.37
Program	1	-536	199	-2.69	0.0078		328	0.61

Table 12.12 shows an example.[301] In this example, a multivariate linear regression model has been used to relate the post-retrofit annual energy consumption of 200 customers to their pre-retrofit annual energy use (pre-use), the change in business conditions for the customer (business), and participation in the program (coded 1 if yes, 0 if no).[302] The column labeled Parameter Estimate shows the estimated intercept and the regression coefficients associated with each explanatory variable. The remaining columns show the corresponding standard error, t-value and p-value.

[300] Sometimes, the estimate of the impact of the program is based on more than one coefficient of the regression model. In this case special techniques may be needed to determine the appropriate standard error.

[301] Although the data used to create this example was simulated, the authors believe that this example is representative of many actual applications of regression analysis in impact evaluation.

[302] In this example, one half of the 200 sample customers are program participants. The regression model explains 92% of the variance, i.e., $R^2 = 0.92$.

The regression coefficient associated the program indicator variable is the primary variable of interest. If the usual assumptions of regression analysis are satisfied,[303] this coefficient should be an unbiased estimate of the average reduction in annual energy consumption for each program participant. In this case, the regression results suggest that the program has a savings of about 536 units per participant. Moreover, since the p-value associated with this coefficient is very small (0.0078), an evaluator might conclude that this result is highly significant.

However, the error bound and relative precision suggest a different conclusion. Table 12.12 has two additional columns to show the error bound and relative precision of each parameter estimate. In the case of the program indicator variable, the error bound is calculated as $1.645 \times 199 = 328$ and the relative precision as $328/536 = 0.61$. Therefore a 90% confidence interval for the annual savings is 536 ± 328 units per participant. The error bound and relative precision statistics indicate that the statistical precision is rather poor, despite the excellent p-value of 0.0078. These results show that the p-value can give a misleading indication of the statistical precision of the results of a regression analysis.

Allocation of Resources to Evaluation

Policy makers must decide how much to spend on evaluating a portfolio of programs and how to allocate the spending to the individual programs in the portfolio. This is a complex issue and many factors can influence the analysis, including:

- The amount of savings expected from each program,
- Whether the program is expected to grow or shrink in the future,
- The uncertainty about the savings,
- How long it has been since the last evaluation and how much the program has changed in the interim,
- The unit cost of evaluating each sample project in the program,
- The variability of savings in the population of projects in the program, as measured by the coefficient of variation or the error ratio, as discussed in the chapter on sampling.

[303] The authors have presented the linear regression model to simplify the exposition. The data were simulated assuming that the logarithm of post-use is linearly related to the logarithm of pre-use together with the business and program variables. Therefore our linear regression model is actually mis-specified. The conclusions are essentially the same if the properly specified regression model is used, with the logarithm of post-use as the dependent variable and the logarithm of pre-use together with the business and program variables as the explanatory variables. Under this second model, the standard assumptions are fully satisfied since the model accurately reflects the procedure used to simulate the data. The estimated coefficient of Program is -0.07. This indicates about a 7% reduction in energy consumption due to program participation. The p-value is 0.0396, which is highly significant. But the error bound is 0.06 and the relative precision is 0.79 so the statistical precision is not as good as might be suggested by the p-value.

The objective for this section is to discuss some statistical methodology that can help guide policy makers in these issues.

Bayesian Decision Theory and Power Analysis

Bayesian decision theory provides a structured methodology for determining the value of information and for making decisions between options for collecting information. For example, decision-tree analysis was used by Eric Hildebrandt of the Sacramento Municipal Utility District (SMUD) to assess the value of improved information from impact evaluation. Hypothetical scenarios were examined using cost and impact data for a typical commercial and industrial retrofit program. The decision-tree analysis looked at canceling or redesigning a program versus continuing a program given the information available and chance nodes for energy savings, demand impact, free riders, and the accuracy of the evaluation results. Including program costs and expected benefits allowed this analysis to assess the value of information. A scenario of a program with a break-even benefit-cost of 1.0 found the expected annual value of additional information to range from $8,000 to $182,000. The decision-tree framework was recommended "as a powerful and objective framework for setting impact evaluation budgets and priorities".[304]

In principle, Bayesian decision theory methods seem to be widely applicable to allocating resources to evaluation studies. However, with a few exceptions these methods have not been commonly applied in evaluation nor is there much experience in using this methodology within energy efficiency evaluation. Our assessment is that these methods may have potential value for evaluation but more research and development is needed to realize this potential.[305]

Statistical power analysis is another approach to resource allocation that may be useful in some circumstances. Power analysis is a way of assessing the likelihood that a study will yield statistically significant results. The power of a statistical test is the probability of obtaining sample results that will lead to rejection of a particular null hypothesis given that the null hypothesis is false. For example, if the null hypothesis is that a program has a benefit-cost ratio of one, the statistical power is the ability of a study to demonstrate

[304] "The Value of Improved Information: Using Decision Analysis to Quantify the Value of DSM Impact Evaluation." (* Hildebrandt 1995). Quote on page 126.

[305] Benefit-cost analysis is a classic case where there are several input variables each with their own sources of uncertainty. Often uncertainty in benefit-cost analysis has been handled by using scenarios analyses or sensitivity analysis. Woods and Khawaja have suggested an alternative method, multi-attribute valuation, to simplify this process and avoid assessing the propagation of uncertainty for more complex market transformation or hard to quantify programs. See "Multi-Attribute Valuation for Cost Effective Evaluation of Market Transformation and Other Hard to Quantify Programs." (* Woods et al. 2000).

that the benefit-cost ratio is significantly greater than one, given an assumed value of the true benefit-cost ratio.[306]

Propagation of Uncertainty

A closely related approach is to study how uncertainty propagates from underlying factors to the actual savings of a program or a portfolio. For example, if the total savings of a portfolio of programs is being estimated, this chapter has shown how to calculate the error bound of the total savings from the error bounds of the savings of each program. This is a simple illustration of the propagation of error, in this case from the savings of individual programs to the total savings of the portfolio of programs. In this section more complex relationships between the total savings and underlying factors are examined.

The general approach is to express the total savings as a function of two or more underlying factors. Then the amount of uncertainty in each factor is related to the uncertainty in the total savings. The results are used to assess the impact on the total savings of reducing the uncertainty in each of the factors. Finally the results are used to allocate resources to studying each factor.

This example again uses the chaining concept discussed above and examines net savings of a program, denoted NS. Net savings is calculated here as $NS = GS \times NTGR$ where GS denotes the gross savings of the program and $NTGR$ denotes the net-to-gross ratio of the program. Then, using Taylor's equation, this is approximated by: $NS \cong E(GS) \times NTGR + GS \times E(NTGR) - E(GS) \times E(NTGR)$. With this approximation, and assuming that GS and $NTGR$ are statistically independent random variables, the standard deviation of NS is approximately equal to

$$\sqrt{E(GS)^2 \times sd(NTGR)^2 + sd(GS)^2 \times E(NTGR)^2}$$

Using the added fact that $E(NS) = E(GS)\,E(NTGR)$, the relative precision of NS, denoted $rp(NS)$, is approximately equal to $\sqrt{rp(GS)^2 + rp((NTGR))^2}$ as discussed in the previous section on chaining.

Table 12.13 shows how the relative precision of the gross savings and the relative precision of the net-to-gross ratio affect the relative precision of the net savings. In this case it can be seen that it is generally more effective to reduce the larger relative precision than the smaller relative precision. For example if $rp(GS) = .10$ and $rp(NTGR) = .25$, then, other considerations aside, it is better to reduce $rp(NTGR)$ to .20, giving $rp(NS) = .22$ than to reduce $rp(GS)$ to .05, giving $rp(NS) = .25$.

[306] For an introductory discussion of statistical power analysis, see *Elements of Statistical Reasoning* (Minium et al. 1982). One of the classic textbooks on power analysis is *Statistical Power Analysis for the Behavioral Sciences*. (Cohen 1990).

Table 12.13: The Propagation of Error

		Relative Precision of NTGR				
		5%	***10%***	***15%***	***20%***	***25%***
Relative Precision of Gross Savings	***5%***	7%	11%	16%	21%	25%
	10%	11%	14%	18%	22%	27%
	15%	16%	18%	21%	25%	29%
	20%	21%	22%	25%	28%	32%
	25%	25%	27%	29%	32%	35%

An example of the use of the propagation-of-error method in energy efficiency evaluation is available in an article by Violette et al. which summarizes work performed by Xenergy for Boston Edison Company in 1992.[307] This paper discusses the use of engineering priors within billing analysis and the variety of data sources and combination of elements involved in engineering savings estimates. The propagation-of-error method is used as a planning tool to assess where the uncertainty can be reduced most cost-effectively. In their example they compare the use of light loggers to obtain operating hours versus post-retrofit inspections to eliminate uncertainty on whether a measure is installed or not.

Another example is work performed at Wisconsin Power & Light Company by Kurt Kiefer. Partial derivatives are used to estimate the individual parameter error and then the contribution to overall uncertainty. This work was part of a evaluation planning process. The work showed the importance of good documentation of savings algorithms and data sources, providing ranges around typical point estimates for each parameter and thorough screening of existing data. This strategic work separated out the "nice to knows" from the "need to knows".[308]

The Fixed Relative Precision Criterion

This section examines the question of how evaluation resources should be allocated to each of the programs in a portfolio of programs in order to maximize relative precision. The traditional criterion is to measure the total savings of each program with the same specified relative precision.

To analyze this approach, the following assumptions are made:

1. The objective is to estimate the savings of each program with a fixed relative precision,

[307] "Statistically-Adjusted Engineering Estimates: What Can the Evaluation Analyst Do About the Engineering Side of the Analysis?" (Violette et al. 1993). Correcting typographical errors in Equation b on page 659:

$$\Delta S = \left[\sqrt{\left(\frac{\partial f}{\partial x}\Delta x\right)^2 + \left(\frac{\partial f}{\partial y}\Delta y\right)^2 + \left(\frac{\partial f}{\partial z}\Delta z\right)^2}\right]_{x=\bar{x},\, y=\bar{y},\, z=\bar{z}}$$

[308] "A Framework for Strategic Evaluation Planning." (* Kiefer 1993).

2. A efficiently stratified sample of projects will be evaluated from each program and an essentially unbiased estimate of the actual savings of each project will be determined using ratio estimation,
3. The strength of the tracking estimates available in each program is measured by a parameter called the error ratio of the program (defined in the chapter on Sampling), and
4. Each program has a relatively large number of projects.[309]

Under these assumptions, the sample size that is required for each individual program can be calculated using the methods discussed in the Sampling chapter. Since the number of projects is assumed to be large, the sample size can be determined by the equation

$$n = \left(\frac{1.645\ er}{D} \right)^2$$

Here D is the desired relative precision and er denotes the assumed error ratio of each program. For example, if $D = 0.10$ and $er = 1.0$ then

$$n = \left(\frac{1.645 \times 1.0}{0.10} \right)^2 = 271$$

Table 12.14 shows an example for the four programs previously considered in Table 12.6. The third column shows the assumed value of the error ratio of each program. The column labeled Sample Size shows the required sample size calculated as above. The expected error bound for each program has been calculated as the product of the expected savings and the desired relative precision. Finally the pooling method previously discussed has been used to calculate the error bound and the expected relative precision of the total savings across all four programs.

Table 12.14: Sample Sizes for a Fixed Relative Precision of 10% at 90% Confidence

Program	*Expected Savings*	*Error Ratio*	*Desired Relative Precision*	*Sample Size*	*Expected Error Bound*
A	50,000	1.0	10%	271	5,000
B	50,000	0.6	10%	97	5,000
C	100,000	0.5	10%	68	10,000
D	500,000	0.9	10%	219	50,000
Total	700,000		7%	655	51,478

The equal relative precision rule may not provide a suitable allocation of evaluation resources among the programs since the sample size does not reflect the size of each program or the particular information that is needed for each individual program. For

[309] If the number of projects is small or moderate, a finite population correction can be added, as discussed in the chapter on sampling.

example, although Program A has only 7% of the total savings, it has been allocated over 40% of the total sample under this rule.

Optimal Allocation for the Overall Savings of the Portfolio

A different criterion is to allocate the sample size to each program to estimate the total savings of the portfolio with the best possible statistical precision. To explain this approach, the preceding assumptions are replaced with the following:

1. The objective is to estimate the total evaluation savings of the entire portfolio with the smallest possible error bound,
2. A sample of projects will be evaluated from each program and an unbiased estimate of the actual savings of each project will be determined,
3. The cost of evaluating each sample project is about the same across the projects of the portfolio (In practice this implies that similar methods are being used in each program of the portfolio),
4. The strength of the tracking estimates available in each program is measured by a parameter called the error ratio of the program (defined in the Sampling chapter), and
5. Each program has a relatively large number of projects.

Under these assumptions the overall sample size should be allocated to each program in proportion to the product of the expected savings of each program and the error ratio of the program. Table 12.15 provides an example. To make the new results comparable to the preceding example shown in Table 12.14, it has been assumed that a total sample of 655 projects is to be used across all four programs.

As before, the third column shows the error ratio, which measures the uncertainty in the relationship between actual savings and the tracking estimate of savings. The column labeled "Allocator" is the product of the expected savings of each program and the error ratio. In the column labeled "Optimal Sample Size", the calculations for the sample size that should be allocated to each program are provided. In the case of program A, for example, the sample size is calculated as $\left(\frac{50{,}000}{580{,}000}\right)655 = 56$ projects.

Table 12.15: Optimal Allocation for the Overall Precision

Program	*Expected Savings*	*Error Ratio*	*Allocator*	*Optimal Sample Size*	*Expected Relative Precision*	*Expected Error Bound*
A	50,000	1.0	50,000	56	22%	10,946
B	50,000	0.6	30,000	34	17%	8,479
C	100,000	0.5	50,000	56	11%	10,946
D	500,000	0.9	450,000	508	7%	32,837
Total	700,000		580,000	655	5%	37,280

Once the sample size has been specified for each program, the expected relative precision can be calculated from the equation $rp = 1.645 \frac{er}{\sqrt{n}}$ assuming that there are a relatively large number of projects in each program, as discussed in the chapter on Sampling. Then, the expected error bound associated with each program can be calculated as the product of the expected savings and the expected relative precision, as shown in the last column of Table 12.15. The final step is to calculate the expected error bound and relative precision of the set of four programs taken together, as was discussed in the section on pooling. If it seems desirable to change the overall error bound or relative precision, the total sample size, 655, can be increased or decreased until the results seem appropriate. Finally, the funds allocated to each program can be calculated from the sample size, assuming that the unit cost per sample point is known.

Comparing Table 12.14 to Table 12.15, it can be seen that in this example without increasing the total sample of 655 projects, the statistical precision associated with the overall savings can be improved from ±7% to ±5% using the second criterion.

The allocations of the total sample are very different in Table 12.14 and Table 12.15. In Table 12.14 the fraction of the total sample that is allocated to each program is proportional to the square of the error ratio, without regard to the expected savings of the program. In Table 12.15, by contrast, the fraction of the total sample that is allocated to each program is proportional to the product of the expected savings and the error ratio of the program. Under the latter approach, the size of the program and the uncertainty of savings both have an important effect on the allocation of the evaluation resources.

The preceding methods can be extended to relax the assumption that the cost of evaluating each sample project is about the same across the projects of the portfolio. The first step is to assess the variable cost per project for each program. Then redefine the allocator to be the product of the expected savings of each program and the error ratio, divided by the square root of the unit cost of each program. The new allocator is used to divide up the total assumed sample size across all programs. The total assumed sample size can be adjusted until the total allocation costs are acceptable.

Table 12.16: Optimal Allocation with Different Unit Costs in each Program

Program	*Expected Savings*	*Error Ratio*	*Fixed Eval. Cost*	*Unit Cost / Project*	*Allo-cator*	*Opt. Sam. Size*	*Exp. Rel. Prec.*	*Exp. Error Bound*	*Exp. Eval. Cost*
A	$50,000	1.0	$30,000	$100	5,000	138	14%	7,013	$43,754
B	$50,000	0.6	$30,000	$100	3,000	83	11%	5,432	$38,252
C	$100,000	0.5	$100,000	$1,000	1,581	43	12%	12,472	$143,494
D	$500,000	0.9	$100,000	$1,000	14,230	391	7%	37,415	$491,444
Total	$700,000				23,811	655	6%	40,424	$716,944

shows an example. Two additional columns show the fixed cost of evaluation and the unit evaluation cost per sample project for each program. Then the allocator is calculated as the product of the expected savings and the assumed error ratio divided by the square root of the unit evaluation cost per project. For example, for program A, the allocator is $\frac{50{,}000 \times 1.0}{\sqrt{100}} = 5{,}000$. This provides an optimal sample size for program A as $\left(\frac{5{,}000}{23{,}811}\right) 655 = 138$. Then the expected relative precision for program A is calculated as $1.645 \frac{1.0}{\sqrt{138}} = 14\%$ and the expected error bound as $14\% \times 50{,}000 = 7{,}013$. The expected evaluation cost for program A then becomes: $\$30{,}000 + (\$138 \times 100) = \$43{,}754$. The expected error bound for the total savings of all four programs was calculated as

$$\sqrt{7{,}013^2 + 5{,}432^2 + 12{,}472^2 + 37{,}415^2} = 40{,}424.$$

The expected relative precision for the total savings of all four programs was calculated as $\frac{40{,}424}{700{,}000} = 6\%$. In summary, the last column of shows the optimal way to allocate a total evaluation expenditure of $716,944 among the four programs under the assumptions given in the first four columns. The total assumed sample of 655 projects can be increased or decreased if desired, to fit the available budget and to provide the desired overall statistical precision.

The preceding method takes into account the following four factors:

- The amount of savings expected from each program,
- The uncertainty about the savings,
- The cost of evaluating each sample project in the program,

- The variability of savings in the population of projects in the program, as measured by the coefficient of variation or the error ratio.

The planner may also want to consider other factors such as:

- Whether the program is expected to grow or shrink in the future, and
- How long it has been since the last evaluation and how much the program has changed in the interim.

In considering factors such as these, the planner may want to adjust the optimal sample sizes shown in for one or more of the programs. The adjusted sample sizes can be used to recalculate the expected relative precision, error bound and evaluation cost for each program and across all programs. By working interactively with a spreadsheet similar to , a suitable plan can be developed. [310]

[310] David Jump, Devan Johnson and Linda Farinaccio have developed a tool for assessing several sources of uncertainty, analyzing the propagation of uncertainty, and selecting the most cost-effective M&V plan option. The authors observe that more expensive M&V plans might yield less risk. See "A Tool to Help Develop Cost-Effective M&V Plans." (* Jump et al. 2000).

Chapter 13: Sampling

Introduction and Key Issues

All evaluation studies collect data from participants, non-participants, or the market to provide information for evaluation analysis. Unless all relevant members of a group have data collected from them (a census), some type of sampling is used. This means that sampling is used in most evaluation studies.

After the introduction, the chapter is divided into three major parts. The first part provides the basic statistical background involved in simple random sampling. This material can be found in most graduate statistics textbooks.[311] The second part explains the basic ideas and methods of stratified ratio estimation. Most of this material is based on relatively recent literature in finite population sampling and will not be found in most statistics textbooks. However, these methods have been used extensively in impact evaluation studies and have proven to be effective.[312] [313] The third part of the chapter is the sampling roadmap itself. The sampling roadmap describes one recommended approach to statistical sample design and analysis in a typical evaluation study. The roadmap builds on the terminology and methods of the first two sections but discusses the specific chronological steps that are recommended.

The most important section of this chapter for program implementers is the discussion of the population database. This section raises issues that are best addressed as part of program design and the design of the tracking database. A general understanding of the ratio model and the error ratio in characterizing the accuracy of the tracking estimates of savings could also significantly benefit program designers. They may also obtain some value from an overview of the sampling roadmap.

Many evaluation planners and study directors could find the background provided in the first two parts informative. The roadmap provides a straightforward guide for evaluators and policy makers.

---

[311] A classic text is *Social Statistics* by Hubert M. Blalock, Jr., Stouffer Award recipient from the American Sociological Association for his outstanding contributions in social research and research methodology. (Blalock 1979)

[312] Much of the underlying theory was published in "Finite Population Sampling with Multivariate Auxiliary Information,"(Wright 1983). These methods were used to develop the 1991 impact evaluation plan for Western Massachusetts Electric Company. Wright provided supporting testimony before the Department of Public Utilities of the Commonwealth of Massachusetts in DPU 91-44 on April 19, 1991.

[313] Some recent studies using these methods in California are "SCE Non-Residential New Construction Persistence Study" (RLW Analytics 1998), "Impact Evaluation of PG&E & SCE's 1994 NonResidential New Construction Programs" (RLW Analytics 1996), " Southern California Edison 1998 Non-Residential New Construction Evaluation" (RLW Analytics et al. 1999), and "Final Report - 1999-2001 Building Efficiency Assessment (BEA) Study: An Evaluation of the Savings By Design Program" (RLW Analytics 2003). All of these reports can be found on the CALMAC website at http://www.calmac.org/.

Skills Required for Sampling

The population database work and simple random sampling (or a census) does not require an advanced statistics background. Stratified ratio estimation is somewhat more complex but builds on the more familiar concepts and techniques of simple random sampling. The sampling roadmap provides a step-by-step approach. Yet, it probably still requires someone to have basic training and/or experience in statistics to ensure that it is understood and applied correctly. Given this, persons conducting and reviewing this work would need to have at least basic graduate statistics or equivalent experience with a mentor in this area.

Key Issues

In a typical evaluation study, data are collected and analysis is conducted for a sample of units, usually projects or customers, selected from a given population. By following statistical sampling methods, the data collection and analysis usually can be limited to a relatively small sample. For example, in an impact evaluation study, project-specific measurement and verification (M&V) analysis might be carried out for a sample of 50 to 300 projects selected from the 1,500 projects implemented in the program in a given year.

When it is used effectively, sampling can improve the overall quality of an evaluation study. By limiting resource-intensive data collection and analysis to a relatively small fraction of all projects, more attention can be devoted to each sample project. The goal of the sampling and research design of an impact evaluation, for example, is a sound, defensible, unbiased determination of the actual gross and net savings for the overall program. Some measurement error is acceptable for each sample project – especially if the measurement error is small relative to the sampling variability. But measurement bias should be minimized since it will propagate through the analysis. It is equally important to minimize other sources of bias such as non-response, self-selection, deliberate substitution, etc.[314]

If the sample projects are selected following an efficient sample design, and if the data collection and site-specific analysis is free of substantial bias, then the statistical analysis can provide an essentially unbiased estimate of each population characteristic of interest and a good measure of the achieved statistical precision. The statistical analysis can also provide measures of population variability to guide future evaluation studies.

If a suitable sample design is followed in selecting the sample projects, the estimates of population characteristics developed from the sample data will generally be close to the true values that would have resulted if the same data collection and project-specific M&V analysis had been carried out for all projects in the population. Moreover, an error bound can be calculated to assess the statistical precision of the results.

[314] See Chapter 12 on Uncertainty for a further discussion of potential bias issues that need to be addressed within the evaluation planning process and in analysis and reporting.

However, the estimates calculated from the sample may be seriously biased and the statistical error bound may be misleading if (a) the final sample is substantially different than the primary sample due to non-response, refusals, or substitutions or (b) the research design, data collection, or analysis is seriously biased. Therefore it is sound practice to reinvest some of the resources saved by sampling into efforts to (a) promote high response rates and minimize substitutions, and (b) minimize bias in the data collection, research design, and analysis.

General Steps in a Study

This chapter summarizes the sampling methodology that is generally followed in an evaluation study. As shown in Figure 13.7, a study usually moves through the following seven sequential stages:

1. **Objectives** – identify the objectives of the study,
2. **Population Database** – create the list of all units in the target population,
3. **Data Collection Approach** – plan the basic methodology that will be used to develop the desired information for the sample units,
4. **Sample Design** – develop the approach to selecting the sample units,
5. **Sample Selection** – select the sample units following the sample design,
6. **Data Collection and M&V** – develop the desired information for each sample unit,
7. **Statistical Analysis** – extrapolate the sample data up to the population, calculate error bounds, and develop the information needed to guide future sample designs, and
8. **Final Report** – summarize the methodology, findings and recommendations of the study.

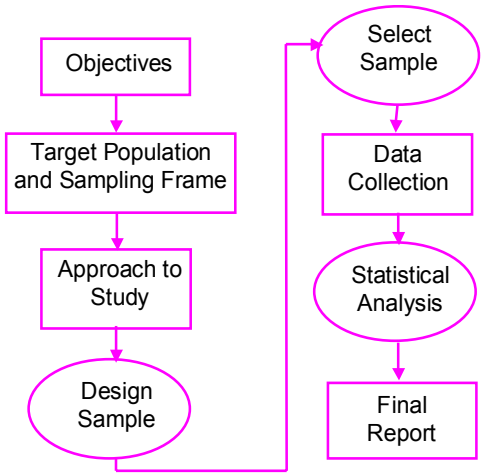

Figure 13.7: Stages in a Study

However, the conceptual foundations cannot be discussed effectively in this linear fashion. For example, the sample design and statistical analysis methodologies have to be considered together as a coherent strategy for developing information. Moreover, it is easier to develop a thorough conceptual understanding of the relatively simple approach

of simple random sampling before embarking on the more complex methodology of stratified ratio estimation. Therefore the conceptual discussion given in the first two parts does not strictly follow this sequence shown in Figure 13.7. This chapter will, however, return to this chronological sequence in the sampling roadmap given in the final section of this chapter.

Population Database

The starting point of any discussion of finite population sampling is the target population itself. For example, in an impact evaluation study, the population may be the set of projects implemented in the program in a given year. In a baseline study the population may be the set of all non-residential customers served by a given utility or the set of all vendors of a given appliance in a specified region.

The population database is a list of each unit in the population, with relevant information for each unit. If the target population is a set of projects implemented in a particular program in a given year, then the population database is a list of each of these projects, together with supporting information for each project, developed from the program tracking system.

In the context of simple random sampling, the minimum requirement is the information needed to identify each project so that the desired data can be collected for the projects that fall into the sample. The population database generally should also have a suitable measure of the size of each unit in the population. This usually is either the tracking estimate of the savings of each project in the program, or the annual energy usage of each unit in the population. For more information about constructing the population database, please see the sampling roadmap given at the end of this chapter.

Simple Random Sampling

In simple random sampling, a sample of a given size, denoted n, is selected from the projects in the population following any randomized procedure in which all possible subsets of n projects are equally likely to be selected. For example, assume there is a complete computerized list of all of the projects in the population.[315] This is called a sampling frame. A simple random sample is selected by assigning a random number to each project, sorting the sampling frame according to the random number, and designating the first n projects in the sorted list to be the sample.

In evaluation, simple random sampling can be very effective if the projects in the population do not vary too much in size. This may be the case for some residential programs and certain non-residential programs if larger projects are suitably divided into

[315] The population must be unambiguously defined. Although most programs are ongoing, it is not appropriate to allow the population to be constantly changing. Therefore it is common to define the population for an evaluation study to be the set of projects paid in a given program in a given year.

subprojects that can be independently evaluated. The statistical theory of simple random sampling is also of interest because it is the foundation of all other methods of sampling.

Basic Definitions

Some technical terminology and notation is needed to discuss the methodology of simple random sampling. This notation will also be useful in discussing the more complex sampling methods to follow. Let N denote the number of projects in the population and assume that the projects are labeled $i = 1, \ldots, N$. Let y denote any measurable variable of interest, such as gross or net savings. Then let y_i denote the value of y for project i. Y will denote the true total of y for all N projects in the population, i.e., $Y = \sum_{i=1}^{N} y_i$. The population mean of y is denoted as μ:

$$\mu = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i .$$

The population variance of y is denoted by σ^2:

$$\sigma^2 = \frac{1}{(N-1)}\sum_{i=1}^{N}(y_i - \mu)^2 .$$

The population standard deviation of y is $\sigma = \sqrt{\sigma^2}$ and the population coefficient of variation of y is $cv = \sigma/\mu$. These are all parameters of the population; in practice they are generally unknown.

Now assume that a random sample of n projects has been selected and that the sample projects are labeled as $i = 1, \ldots, n$. Assume that the value of y_i has been observed [316] [317]

[316] In discussing sampling, it is common to neglect measurement error. In the context of impact evaluation, for example, the population mean μ is regarded to be the mean of the values of y_i that would be hypothetically observed if all projects in the population were subjected to the same data collection and analysis techniques planned for the sample projects. Only minor changes in our discussion would be needed to accommodate random measurement error. Most of our results would be unchanged with the exception of the finite population correction factor.

[317] Measurement bias is much more serious than random measurement error. If the planned data collection and analysis techniques give a biased estimate of the true savings of each project that is analyzed, then μ itself will be biased in the sense that it will not be equal to the population mean of the true savings of all projects. In this case, although the procedures discusses in this chapter will yield an unbiased estimate of μ with measurable statistical precision, it is generally very difficult to estimate the difference between μ and the population mean of the true savings of all projects. See the chapter on Uncertainty for a more complete discussion of bias.

[318] for each sample project i. Corresponding to the population parameters that have been defined above, the sample statistics are defined as:

- The sample mean of y is $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$,
- The sample variance of y is $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$,
- The sample standard deviation of y is $s = \sqrt{s^2}$, and
- The sample coefficient of variation of y is $\hat{c}v = s/\bar{y}$.

In the context of simple random sampling, each of these sample statistics can be regarded as an estimator of the corresponding population parameter. The key issue is the performance of statistics as estimators. This issue bears on the choice of sample size and the statistical precision to be expected of an estimator. There is generally particular interest in the statistical precision of the sample mean $\bar{y}$ as an estimator of the population mean μ.[319] The central concept needed to address these issues is the sampling distribution of $\bar{y}$.

Sampling Distribution of the Sample Mean

The preceding section discussed the procedure of selecting a random sample of a fixed size from a fixed population, observing the value of a particular variable y for each sample project, and calculating the sample mean, $\bar{y}$. Now consider the following conceptual experiment. Suppose the procedure just described is repeated a large number of times. The distribution of values of $\bar{y}$ that would be observed in repeated sampling is called the sampling distribution of the sample mean.[320]

The sampling distribution can be used to define the expected value, variance, and standard deviation of $\bar{y}$. Under simple random sampling, the following results are known:

1. The expected value of $\bar{y}$ in repeated sampling is equal to μ, the population mean of y. In other words, $\bar{y}$ is an unbiased estimator of μ.

---

[318] One form of measurement error can arise when some of the sample projects are so large that only part of the total space or measures can be explicitly evaluated. In this situation it is important to avoid measurement bias arising from the selection of the spaces or measures to be studied.

[319] Analogous methods are also readily available for a sample proportion as an estimator of a population proportion. These methods are commonly used to help plan market research and other qualitative studies, especially if the units in the population are fairly homogeneous in size.

[320] This idea of repeated sampling and its use in statistical inference comes from the frequency interpretation of probability that is the basis for the classical theory of statistics. The alternative Bayesian approach relies upon the subjective interpretation of probability. See *Bayesian Statistics for Evaluation Research: An Introduction* (Pollard 1986). For a brief discussion of the Bayesian approach and added references please see the chapter on Uncertainty.

2. The standard deviation of $\bar{y}$ in repeated sampling, sometimes denoted as $sd(\bar{y})$, is equal to $\frac{\sigma}{\sqrt{n}}$, provided that the population size N is large relative to the sample size n. Here σ is the population standard deviation of y. In general, $sd(\bar{y}) = \sqrt{1-\frac{n}{N}}\frac{\sigma}{\sqrt{n}}$ where $\sqrt{1-\frac{n}{N}}$ is called the finite population correction factor.

3. If the sample size n is large but substantially smaller than the population size N, then the sampling distribution of $\bar{y}$ can be approximated by the normal probability distribution.

All three of these results can be combined into the following statement: In repeated sampling, the probability is about 90% that the sample mean $\bar{y}$ will fall within $\pm 1.645\, sd(\bar{y})$ of the population mean, μ. This key result is the basis for calculating the expected statistical precision for a specified sample size, for choosing the sample size to provide a desired statistical precision, for estimating the statistical precision from the sample data that have been collected, and for estimating the population parameters needed to plan future studies.

Expected Statistical Precision

Refer again to the key statement: In repeated sampling, the probability is about 90% that the sample mean $\bar{y}$ will fall within $\pm 1.645\, sd(\bar{y})$ of the population mean μ. This statement is used to define the two most important measures of statistical precision: the error bound and the relative precision. The quantity $1.645\, sd(\bar{y})$ is called the expected error bound of $\bar{y}$ as an estimator of μ at the 90% level of confidence[321]. This is denoted as $eb(\bar{y})$. Assuming that μ is greater than zero, the quantity $eb(\bar{y})/\mu$ is called the expected relative precision at the 90% level of confidence and is denoted as rp.

Using result 2 from the preceding section, the expected error bound of $\bar{y}$ is given by the equation $eb(\bar{y}) = 1.645\frac{\sigma}{\sqrt{n}}$ provided that the population size N is large relative to the sample size n. In general, $eb(\bar{y}) = 1.645\sqrt{1-\frac{n}{N}}\frac{\sigma}{\sqrt{n}}$. Similarly, the relative precision is

[321] The 90% level of confidence has become conventional in evaluation work. The authors recommend that the 90% level of confidence be used consistently to avoid needless confusion.

given by the equation $rp = 1.645\frac{cv}{\sqrt{n}}$ if N is large, and $rp = 1.645\sqrt{1-\frac{n}{N}}\frac{cv}{\sqrt{n}}$ in general.[322]

These results show that the expected statistical precision depends on the following three factors:

1. The sample size n,
2. The population size N, especially if the latter is relatively small, and
3. The variability of y in the population, measured either by σ or by cv.

Choice of Sample Size

The preceding results can be used to guide the choice of the sample size. The analysis requires an assumption concerning the variability of y in the population, measured by cv. In many applications, the cv of y can be estimated adequately by calculating the coefficient of variation of the measure of size that is available in the population database. Of course the number of units in the population N is also known from the population database.

Given N and the estimated value of cv, several techniques provide guidance for the choice of sample size. One approach is to create a table showing the expected relative precision for different choices of sample size n, using the equation $rp = 1.645\sqrt{1-\frac{n}{N}}\frac{cv}{\sqrt{n}}$.

Another approach is to set the desired relative precision and calculate the required sample size. The desired relative precision is usually denoted D. Assume initially that the population size N is large relative to the sample size n. Using the equation $rp = 1.645\frac{cv}{\sqrt{n}}$, the required sample size can be estimated as a function of the population cv:

$$n = \left(\frac{1.645\ cv}{D}\right)^2$$

---

[322] In evaluation, the focus is sometimes on the population total $Y = \sum_{i=1}^{N} y_i$. When simple random sampling is used, the population total is estimated using the statistic $\hat{Y} = N\ \bar{y}$. Since N is not subject to sampling variability, the standard deviation of $\hat{Y}$, denoted $sd(\hat{Y})$, is equal to $N\ sd(\bar{y})$. As usual, the error bound of $\hat{Y}$ is $eb(\hat{Y}) = 1.645\ sd(\hat{Y}) = N\ eb(\bar{y})$. The expected relative precision of $\hat{Y}$ is identical to the expected relative precision of $\bar{y}$.

For example, if $D = 0.10$ and assuming $cv = 0.75$, then the preceding equation gives

$$n = \left(\frac{1.645 \times 0.75}{0.10} \right)^2 = 152$$

Table 13.17 shows the results of this type of calculation for various values of cv and D. The table shows clearly that both of these parameters have a dramatic impact on the required sample size. It will be shown later in this chapter that the relevant population variability can often be reduced very substantially by using stratified ratio estimation instead of simple random sampling.

Table 13.17: Required Sample Size Assuming a Large Population

		Coefficient of Variation, cv						
		0.5	***0.75***	***1.0***	***1.5***	***2.0***	***2.5***	***3.0***
Desired Relative Precision D	***0.25***	11	24	43	97	173	271	390
	0.20	17	38	68	152	271	423	609
	0.15	30	68	120	271	481	752	1,082
	0.10	68	152	271	609	1,082	1,691	2,435

If there are a small or moderate number of projects in the population, then a finite population correction can be applied to the preceding method. The first step is to use the preceding equation to calculate an initial, unadjusted estimate of the required sample size, denoted n_0:

$$n_0 = \left(\frac{1.645\, cv}{D} \right)^2$$

Then the following finite population correction is applied:

$$n = \frac{n_0}{1 + n_0 / N} \ .$$

For example, if the population has 500 projects so that $N = 500$, and as before $D = 0.10$ and $cv = 0.75$, then the required sample size can be calculated in two steps as follows:

$$n_0 = \left(\frac{1.645 \times 0.75}{0.10} \right)^2 = 152 \text{, and}$$

$$n = \frac{152}{1 + 152/500} = 117 \ .$$

After the finite population adjustment, the sampling fraction is $\frac{n}{N} = \frac{117}{500} = 0.23$.

Table 13.18 illustrates the second step of this calculation for different population sizes, assuming $n_0 = 152$. The table illustrates an important fact: if the population is large, then the finite population correction has little effect on the sample size n, but if the population is small, then the finite population correction is important..

Table 13.18: Finite Population Adjustment for Different *N*

N	100	500	1,000	5,000	10,000
n	60	117	132	148	150
n/N	0.60	0.23	0.13	0.03	0.01

The table illustrates a second important fact that has important policy implications. If the population is large, then the sample size usually is a small fraction of the population but if the population is small, the sample can be a substantial portion of the population. This can have an important impact on the cost of evaluating small programs. In other words, the traditional criterion of obtaining ±10% relative precision at the 90% level of confidence for each program would impose a disproportionately larger burden on small programs than large programs.[323]

In place of this traditional criterion, the desired relative precision can be relaxed for smaller programs. Table 13.17 shows that this can reduce the required sample size substantially. Alternatively, the authors recommend that the evaluation budget allocated to each program be guided by the resource allocation principles discussed in Chapter 12 on uncertainty.[324]

Statistical Analysis

It is useful to summarize some of the results discussed so far. It has been shown that in repeated sampling, the probability is about 90% that the sample mean $\bar{y}$ will fall within $\pm 1.645\, sd(\bar{y})$ of the population mean μ. The quantity $1.645\, sd(\bar{y})$ has been called the expected error bound at the 90% level of confidence and is denoted $eb(\bar{y})$. The quantity

[323] The original CADMAC Protocols specified that if the program population was small, a census should be attempted. For example, if a program had 100 projects, an attempt was made to evaluate each one of the 100 projects. Suppose that a number of projects were dropped from the evaluation, leaving a convenience sample or self-selected sample of perhaps 60 projects. A casual reading of Table 13.18 might suggest that this would be adequate in this example since 60 is the required sample size under the ±10% relative precision criterion. However, while a random sample of 60 projects will provide an unbiased estimate of the total savings of all 100 projects with measurable precision, a convenience or self-selected sample is vulnerable to bias that is difficult to quantify.

[324] The resource allocation approach tends to focus on the expected error bound of the estimator of the population total, $\hat{Y}$, denoted $eb(\hat{Y})$. Since $eb(\hat{Y}) = rp\, Y$, the expected error bound of the total can always be calculated by multiplying the expected relative precision by the expected value of Y.

$rp = eb(\bar{y})/\mu$ has been called the expected relative precision at the 90% level of confidence. Under simple random sampling, recall that $sd(\bar{y}) = \frac{\sigma}{\sqrt{n}}$ if the population size N is large relative to the sample size n, and in general $sd(\bar{y}) = \sqrt{1 - \frac{n}{N}}\frac{\sigma}{\sqrt{n}}$.

Now assume that a specific random sample has been selected of n projects labeled as $i = 1, \ldots, n$, and that the value of y_i has been observed for each sample project i. The sample data is used to calculate the following sample statistics: the sample mean of y, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample variance of y, $s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(y_i - \bar{y})^2$, the sample standard deviation of y, $s = \sqrt{s^2}$, and the sample coefficient of variation of y, $\hat{c}v = s/\bar{y}$. Here $\bar{y}$ is an estimator of the population mean μ, s is an estimator of the population standard deviation σ, and $\hat{c}v$ is an estimator of the population coefficient of variation cv.

These statistics, in turn, can be used with the equations just given to estimate the standard deviation of $\bar{y}$, as well as the corresponding error bound and relative precision. When the standard deviation of a statistic has been estimated, the result is called a standard error. The standard error of $\bar{y}$ is $se(\bar{y}) = \frac{s}{\sqrt{n}}$ if the population size N is large relative to the sample size n, and $se(\bar{y}) = \sqrt{1 - \frac{n}{N}}\frac{s}{\sqrt{n}}$ in general.[325] Then the error bound can be calculated as $eb(\bar{y}) = 1.645\ se(\bar{y})$ and the relative precision as $rp = eb(\bar{y})/\bar{y}$.

The 90% confidence interval for μ is defined to be $\bar{y} \pm eb(\bar{y})$. In the hypothetical context of repeated sampling, the probability is about 90% that the confidence interval (which will vary from sample to sample) will contain the true value of μ.[326]

Applicability to Impact Evaluation

In impact evaluation, an important objective usually is to assess the true annual energy savings of all projects undertaken in a program. One way to do this is to select a random sample of projects from the population of all projects and then assess the annual energy savings of each sample project using suitable data collection and analysis techniques. Under the assumptions of simple random sampling, the average of the annual energy savings of the sample projects is an unbiased estimator of the average annual energy

---

[325] If the population total is of interest, the standard error of $\hat{Y}$ is equal to N times the standard error of $\bar{y}$, i.e., $se(Y) = N\ se(\bar{y})$.

[326] Sometimes it is useful to calculate a one-sided confidence interval. For example, a one-sided confidence interval could be used to state that the savings are 'no smaller than …'.

savings of all projects in the population from which the sample has been drawn. Thus far this chapter has described procedures for assessing the statistical precision of this estimator and results that can guide the choice of the sample size.

The assumptions of simple random sampling are the key to these results. It is assumed that the sample is randomly selected from the entire target population and that the actual annual savings are observed for all projects in the sample. So the possibility of participant refusal or other difficulties has been excluded that would lead to replacement of a sample project by an alternate. The methods here also assume that the actual annual savings is observed for each sample project with little or no measurement error or bias compared to the expected relative precision of the results.

If, contrary to these assumptions, there is substantial bias, perhaps due to self-selection, non-response, deliberate substitution of sample projects, or measurement bias, then the methods presented here can be seriously misleading. For example it is misleading and counterproductive to report that the average savings has been estimated with a relative precision of ±10% at the 90% level of confidence if there is a serious risk that the results might be in error by 25% due to bias. Bias is particularly insidious because it usually is extremely difficult to assess its magnitude and often even its direction.

Consequently, in a sound evaluation study, every practical effort should be made to minimize all of the likely sources of bias such as non-response, self-selection, deliberate substitution, and measurement bias.[327] One of the important goals of sampling is to limit the sample size so that there will be adequate resources for sound data collection and analysis.

The statistical precision provided by a simple random sample is strongly influenced by the variability of annual savings from project to project in the target population. For example, Table 13.17 shows that if the coefficient of variation is three, if the desired relative precision is ±15%, and if the population is large, then a sample size of 1,082 is needed with simple random sampling. A sample this large would usually be infeasible.[328] However, this does *not* imply that statistical sampling is impractical whenever there is large variability of savings in the target population – only that simple random sampling is impractical. The stratified random sampling techniques discussed in the next section generally address this situation very effectively, especially if the program provides good tracking information for each project.

The sample size is also affected by the size of the target population. Assuming that the sample size is chosen to provide a fixed relative precision, the sample will generally be a small fraction of the population if the population is large, but it will be a larger fraction of the population if the population is small. Thus sampling will be inherently more cost-effective for large programs than for small programs.

---

[327] See the chapter on Uncertainty for a more detailed discussion of potential bias, and the need to assess and report potential bias and threats to the evaluation's validity.

[328] Moreover a very large sample generally precludes the steps needed to mitigate bias.

For this reason, if there are a number of small programs, it may be desirable to consolidate the programs for the purpose of evaluation. In this context, the population would be taken to be the entire set of projects undertaken in a portfolio of programs. Using the methods of this section, a simple random sample would be selected from the population of all projects in the portfolio of programs, the annual savings of each sample project would be assessed using appropriate data collection and analysis techniques, and the sample results would be used to estimate the average savings per project for all projects in the portfolio. If there are a sufficiently large number of projects within the portfolio, sampling will generally be cost-effective and there should be adequate resources for sound data collection and analysis.[329]

Under this approach, the consolidated sample data can be used to provide an unbiased estimate of the average savings per project for each individual program. However, the average savings for the individual programs would generally have poorer statistical precision than the average savings for the portfolio since the individual-program results would be based on the sample projects that were drawn from that program. Moreover, using the simple random sampling approach, the number of sample projects from each program would not be controlled but would vary randomly from one possible sample to another. Nevertheless, the information from the consolidated impact evaluation, in conjunction with other types of information, may still provide adequate information about each program.

Stratified Ratio Estimation

Stratified ratio estimation combines a stratified sample design with a ratio estimator.[330] Both stratification and ratio estimation take advantage of supporting information available for each project in the population. As an example, suppose that an impact evaluation study is being undertaken to assess the annual energy savings of the projects undertaken in a given program. Suppose that the program tracking system provides an estimate of the annual energy savings of each project in the population. Suppose, furthermore, that a substantial fraction of the projects have comparatively small tracking savings but a relatively small number of projects have very large tracking savings. In this case, the coefficient of variation of the tracking savings will often be quite large, e.g., three or larger, and it can be expected that the population coefficient of variation of the

[329] Consolidation will usually increase the coefficient of variation of annual savings if the average savings per project is different across the individual programs. Under simple random sampling, the larger coefficient of variation could compromise the statistical precision. These problems can usually be addressed through appropriate use of stratified ratio estimation, to be discussed in the next section.

[330] Statisticians have developed many other approaches to sample design and estimation, including sequential sampling, cluster sampling, multi-stage sampling, stratified sampling with mean per unit estimation, stratified sampling with regression estimation, etc. See, for example, *Sampling Techniques*, by W. G. Cochran, Wiley, 3rd Edition, 1977. Any of these methods may be appropriate in a particular application. The authors have found that stratified ratio estimation is generally effective in both impact and process evaluation studies, especially when (a) there is substantial variation in the size of projects in the program, and (b) the tracking system provides fairly accurate estimates of the savings of each project. These conditions are frequently true for energy conservation programs.

actual savings is also large. In this case, the simple random sampling methods described in the preceding section would not be practical.

This problem can be partly mitigated by using the tracking estimate of savings as a stratification variable. Stratifying by the tracking savings generally reduces the coefficient of variation of actual savings in each stratum thereby improving the statistical precision.[331] Moreover, the sampling fraction can be varied from stratum to stratum to further improve the statistical precision. In particular, a relatively small sample can be selected from the projects with small tracking savings, but the sample can be forced to include a higher proportion of the projects with larger tracking savings. In particular, the largest projects can, if desirable, be included in the sample with certainty.

The tracking estimates of savings can also be used in ratio estimation. In impact evaluation, one ratio of interest is the realization rate, i.e., the ratio between the total gross annual savings of all projects in the population and the total tracking savings.[332] To understand the potential advantage of ratio estimation, suppose hypothetically that the actual savings of each project in the population is directly proportional to the savings recorded in the tracking system as illustrated in Figure 13.8.

In the extreme example illustrated in Figure 13.8, the actual savings of each project is 0.8 times the tracking estimate of savings. In other words, the tracking system systematically overstates the saving of each project by 20%. The realization rate, 0.8, is the slope of the line relating the actual savings to the tracking for every project. If the realization rate is known, then the true savings of all projects can be accurately estimated by multiplying the total tracking savings by the realization rate. Moreover, in this extreme case, the realization rate can be assessed perfectly by measuring the actual savings of any one project in the population.

[331] In this case, however, the coefficient of variation of tracking savings within each stratum usually does not provide a meaningful estimate of the coefficient of variation of actual savings within each stratum. Therefore added information is needed to estimate the expected statistical precision and to choose the sample size, e.g., from a prior sample or from a model characterizing the relationship between tracking and actual savings.

[332] The net-to-gross ratio is another ratio of interest. Our experience has been that ratio estimation can be used to estimate essentially all parameters of interest in evaluation.

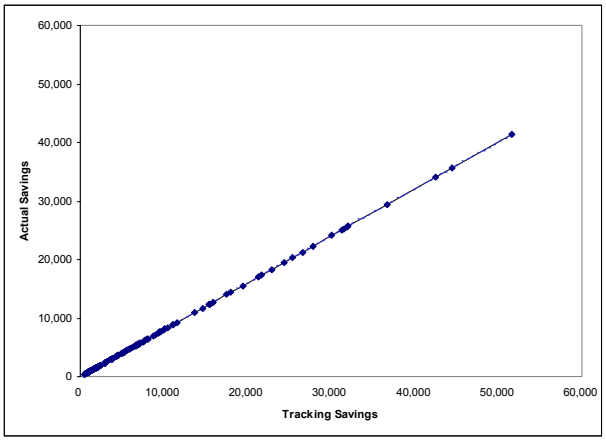

Figure 13.8: Ideal Case for Ratio Estimation

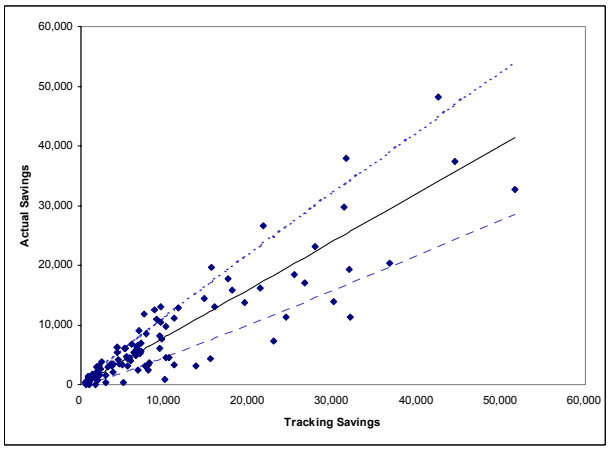

Figure 13.9: More Typical Relationship between the Actual and Tracking Savings

In practice, of course, there is always some random error in the association between the actual and tracking savings. Figure 13.9 illustrates a more typical situation. In this case the tracking estimate of savings is a good but not perfect predictor of the actual savings of each project. Nevertheless, the statistical precision can be greatly improved by using stratified ratio estimation to estimate the realization rate rather than by using simple random sampling to assess the average actual savings as discussed in the prior section.

Goals of the Section and Basic Definitions

This section will provide the tools needed to use stratified ratio estimation effectively in evaluation. The goal is to explain the underlying concepts in enough detail for users to be comfortable with the methodology. Specifically, this section will explain:

- How to estimate the population parameters of interest and to calculate the associated confidence intervals,
- How to characterize the population variation when efficiently stratified ratio estimation is to be used,

- How the expected statistical precision is related to the population variation and to the planned sample size assuming that efficient stratification is used,
- How to estimate the required sample size to achieve a desired relative precision,
- How to construct an efficiently stratified sample design, and
- How to estimate the relevant population variation from the sample for use in planning future studies.

Much of the notation needed to discuss the methodology of stratified ratio estimation is retained from the earlier discussion of simple random sampling. Let N denote the number of projects in the population and assume that the projects are labeled $i = 1, \ldots, N$. Let y denote any measurable variable of interest, such as gross or net savings and let y_i denote the value of y for project i. Y denotes the true total of y for all N projects in the population, i.e., $Y = \sum_{i=1}^{N} y_i$, and μ_y denotes the population mean of y,

$$\mu_y = \frac{Y}{N} = \frac{1}{N}\sum_{i=1}^{N} y_i .$$

Stratified ratio estimation focuses on the relationship between y and a second variable, denoted x. The value of x is assumed to be known for each project in the population,[333] and to avoid minor notational inconveniences, x is assumed to be greater than zero for each project in the population. In the impact evaluation context, x is usually the tracking estimate of the savings of each project. X denotes the total of x for all N projects in the population, i.e., $X = \sum_{i=1}^{N} x_i$ and μ_x denotes the population mean of x,

$$\mu_x = \frac{X}{N} = \frac{1}{N}\sum_{i=1}^{N} x_i .$$

The key population parameter of interest is the ratio between the population total of y and the population total of x, which is denoted B and defined by the following equation:

$$B = \frac{Y}{X} = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} .$$

Of course, B is also equal to the ratio between μ_y and μ_x, i.e., $B = \frac{\mu_y}{\mu_x}$.

[333] Stratified ratio estimation can also be used when the denominator of the ratio is unknown. For example this methodology can be used to estimate the net-to-gross ratio. In this case, a different variable, usually the measure of size in the tracking system, is used for stratification.

Stratified sample design uses knowledge about the population to add efficiency to the sample design. A stratum is any subset of the projects in the population that is based on known information. A stratification of the population is a classification of all units in the population into mutually exclusive strata that span the population. Under a stratified sample design, simple random sampling is used to select a chosen number of projects from each of the pre-established strata.

Added notation is needed to discuss stratified sampling. Let L denote the number of strata and assume that the strata are labeled $h = 1, \ldots, L$. Let N_h be the total number of population projects in stratum h. Let n_h be the number of projects to be randomly selected from stratum h. Assume that n_h is greater than zero for each stratum h. Then $\sum_{h=1}^{L} N_h = N$, the total population size, and $\sum_{h=1}^{L} n_h = n$, the total sample size.

Using this notation, the stratified ratio estimator can be defined. For each project i in the sample, the case weight is defined according to the equation $w_i = N_h / n_h$ where h denotes the particular stratum that contains project i. Using the case weights, define the stratified ratio estimator of B, denoted b, as follows:[334]

$$b = \frac{\sum_{i=1}^{n} w_i \; y_i}{\sum_{i=1}^{n} w_i \; x_i}$$

The statistical precision of b can be assessed by calculating the standard error using the following equation:

$$se(b) = \frac{1}{\hat{X}} \sqrt{\sum_{i=1}^{n} w_i \; (w_i - 1) \, e_i^2}$$

Here $\hat{X} = \sum_{i=1}^{n} w_i \; x_i$ and $e_i = y_i - b \, x_i$. Then, as usual, the error bound can be calculated as $eb(b) = 1.645 \; se(b)$ and the relative precision can be calculated as $rp = eb(b)/b$.

---

[334] An equivalent equation is $b = \frac{\hat{Y}}{\hat{X}} = \frac{\sum_{h=1}^{L} N_h \; \bar{y}_h}{\sum_{h=1}^{L} N_h \; \bar{x}_h}$. Technically, the stratified ratio estimator is a biased estimator of the true population ratio. However, Cochran shows that the bias is small if the relative precision of $\sum_{h=1}^{L} N_h \; \bar{x}_h$ is small, pp. 160-167 (* Cochran 1977). In impact evaluation, the bias should be negligible if the population has been appropriately stratified by size as discussed later in this chapter.

Stratified ratio estimation can also be used to estimate the population mean or population total of y from the known population mean or population total of x. The estimator of the mean is $\hat{\mu}_y = b\ \mu_x$ and the corresponding standard error is $se(\hat{\mu}) = \mu_x\ se(b)$. The estimator of the total is $\hat{Y} = b\ X$ and the corresponding standard error is $se(\hat{Y}) = X\ se(b)$.

The Ratio Model

To develop a suitable sample design, it is necessary to characterize the relation between x and y in the population. This is done by assuming a statistical model called the ratio model. The primary equation of the ratio model is $y_i = \beta\ x_i + \varepsilon_i$. Here x_i and y_i denote the value of x and y for each project i in the population, β is an unknown but fixed parameter of the model that is similar to a regression coefficient, and ε_i is similar to the random error in a regression model. As in a regression model, the expected value of ε_i is assumed to be zero for each project i in the population. It is also assumed that $\varepsilon_1, \ldots, \varepsilon_N$ are mutually independent. Then μ_i is defined to be the expected value of y_i given x_i. Under the ratio model $\mu_i = \beta\ x_i$.

Instead of assuming that the standard deviation of ε_i is constant, the standard deviation of ε_i is allowed to vary from project to project. For any project i in the population, the standard deviation of ε_i is denoted as σ_i. This is called the residual standard deviation of project i. The population error ratio of x and y, denoted er, is defined to be

$$er = \frac{\sum_{i=1}^{N} \sigma_i}{\sum_{i=1}^{N} \mu_i}.$$

The error ratio is the key measure of the population variability in the relationship between x and y for stratified ratio estimation. The role of the error ratio in stratified ratio estimation is virtually the same as the role of the coefficient of variation in simple random sampling. Figure 13.10 shows several examples of error ratios ranging from 0.4 (a relatively strong relationship) to 1.0 (a weak relationship).

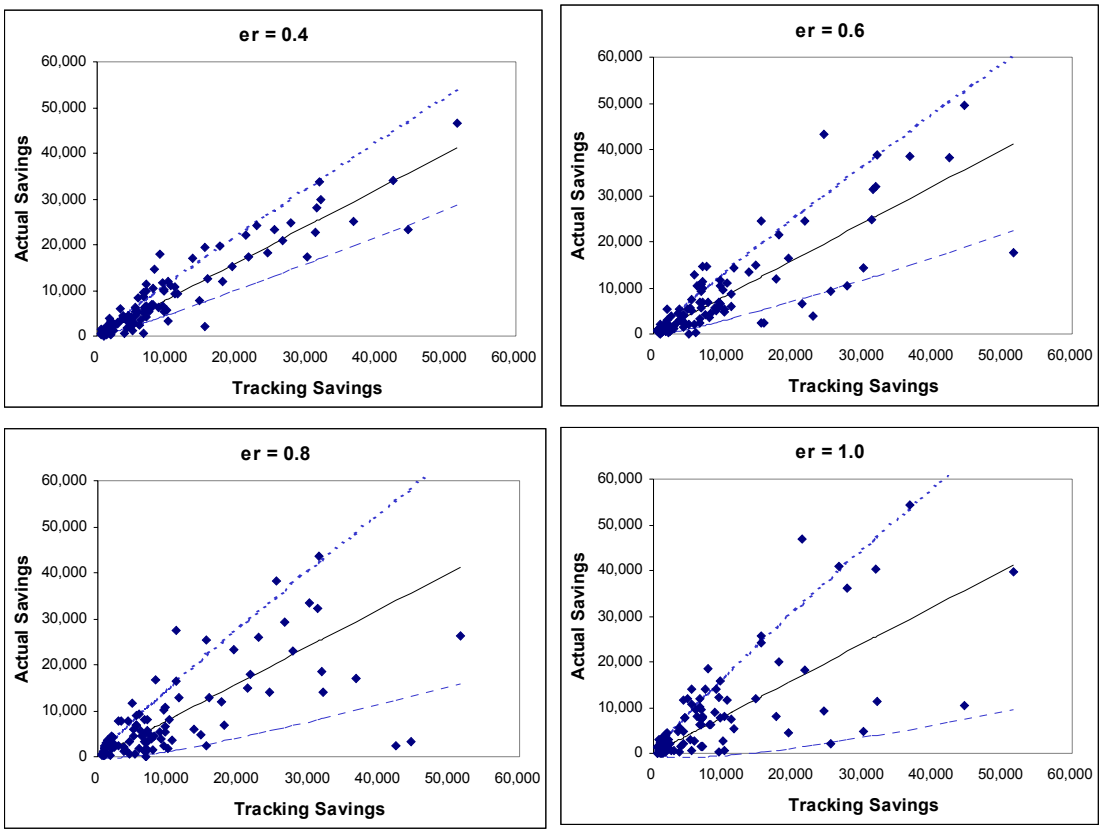

Figure 13.10: Examples of Different Error Ratios

The following specific functional form for σ_i is often assumed: $\sigma_i = \sigma_0 \, x_i^{\gamma}$. This is called the secondary equation of the model.[335] The secondary equation specifies that the residual standard deviation of each project i in the population is proportional to the value of x_i raised to the power γ, pronounced gamma. A common assumption is that $\gamma = 0.8$. This specification is used in constructing efficiently stratified sample designs and to assist in the estimation of the error ratio from a prior sample.

The secondary equation includes a parameter denoted σ_0. This parameter is determined by the error ratio as follows:

$$\sigma_0 = er \, \frac{\sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} x_i^{\gamma}}.$$

[335] Sarndal writes the secondary equation as $\sigma_i^2 = c \, x_i^{\gamma}$ (Sarndal, Swensson et al. 1992), pp. 449.

Sampling Distributions

The simple random sampling section discussed the concept of repeatedly selecting a random sample of a fixed size from a fixed population, observing the value of a particular variable y for each sample project, and calculating appropriate statistics. This concept was used to define the sampling distribution of a statistic such as the sample mean. This same concept of repeated sampling is used in the present discussion with one extension. Instead of regarding y_i as fixed for each project i, y_i is assumed to vary randomly from sample to sample, generated by independent realizations of the ratio model. In other words, the sample is regarded to be randomly determined following the prescribed sample design, and the true values of y_i are considered to be randomly determined for all N units in the population following the ratio model. A more in-depth discussion of this concept can be found in Sarndal.[336]

Expected Statistical Precision and Choice of Sample Size

A key result for stratified ratio estimation is the following: Assuming that the ratio model is accurate, that the sample design is efficiently stratified for the model as described later in this section, that the population size N is large and that the 90% level of confidence is used, then the expected relative precision of the stratified ratio estimator is approximately equal to

$$rp = 1.645 \frac{er}{\sqrt{n}}.$$

This result can be used to guide the choice of the sample size. Suppose that the desired relative precision is denoted D. Under the preceding assumptions, the sample size needed to provide an expected relative precision of D at the 90% level of confidence is approximately

$$n = \left(\frac{1.645\, er}{D} \right)^2.$$

These are the same equations given in the discussion of simple random sampling, but with the coefficient of variation replaced by the error ratio. If N is moderate or small, the finite population correction factor can be used as a first approximation as in simple random sampling. A somewhat more complex but more accurate way of adjusting the large population results for the size of the population will be presented later in this chapter.

For example, if $D = 0.10$ and $er = 0.5$, then the preceding equation gives

[336] (Sarndal, Swensson et al. 1992), pages 448-471.

$$n = \left(\frac{1.645 \times 0.5}{0.10} \right)^2 = 68$$

Table 13.19 shows the results of this type of calculation for various values of er and D. Table 13.19 is similar to Table 13.17 except that in Table 13.19 the error ratio is used since efficiently stratified ratio estimation is being discussed. The sample sizes shown in Table 13.19 are generally much smaller than in Table 13.17 because the error ratio is generally much smaller than the coefficient of variation for a given population.

Table 13.19: Required Sample Size Assuming a Large Population

		Error Ratio er						
		0.4	***0.5***	***0.6***	***0.7***	***0.8***	***0.9***	***1.0***
Desired Relative Precision D	***0.25***	7	11	16	21	28	35	43
	0.20	11	17	24	33	43	55	68
	0.15	19	30	43	59	77	97	120
	0.10	43	68	97	133	173	219	271

Assessing the Error Ratio without a Prior Sample

Figure 13.10, above, illustrated four examples of relationships between x and y. These are typical examples of the type of association expected under the ratio model, assuming various error ratios. In each graph, the solid line represents the expected value of y given x, $\mu_i = \beta\, x_i$, and the broken lines represent the one standard deviation intervals around the mean: $y_i = \beta\, x_i \pm \sigma_i$. In preparing these graphs, the secondary equation $\sigma_i = \sigma_0\, x_i^{\gamma}$ has been assumed with $\gamma = 0.8$.

In most impact evaluation studies, the error ratio can be expected to be in the range 0.4 to 1.0, as illustrated in Figure 13.10. If the tracking system is expected to provide quite accurate estimates of the actual savings of most sample projects in the evaluation study, then the error ratio is likely to be relatively small, e.g., near 0.4. This might be the case, for example, if the program provides energy efficiency retrofits to large commercial buildings, and the tracking estimates of savings are based on a fairly detailed analysis of each project that is undertaken in the program. If the tracking system is expected to provide rather poor estimates of the actual savings of most sample projects in the evaluation study, then the error ratio is likely to be larger, e.g., near 1.0. This might be the case, for example, if the program is an express-style program that requires only a simple application and does not provide any site-specific analysis as part of the program delivery.

Estimating the Error Ratio from a Sample

When stratified ratio estimation is being used to analyze a sample, the sample can also be used to estimate the underlying error ratio for use in future sample designs. Assuming the secondary equation $\sigma_i = \sigma_0 \, x_i^{\gamma}$ with $\gamma = 0.8$, then the error ratio can be estimated as

$$\hat{er} = \frac{\sqrt{\left(\sum_{i=1}^{n} w_i \; e_i^2 / x_i^{\gamma}\right)\left(\sum_{i=1}^{n} w_i \; x_i^{\gamma}\right)}}{\sum_{i=1}^{n} w_i \; y_i}$$

Here, as usual, $e_i = y_i - b \, x_i$.[337]

Model-Based Stratification

The preceding results assume that stratified ratio estimation is used with an efficiently stratified sampling plan. This section will describe how to construct an efficiently stratified sampling plan. The goal is to group the projects into several strata based on the value of x, usually the tracking estimate of savings, and then specify the number of sample projects to be selected from each stratum. The following method is called model-based stratification by size.[338]

The following steps are required:

1. Create a spreadsheet or database listing each project in the population and providing the value of x_i for each project, $i = 1, \ldots, N$.

2. Use the assumed secondary equation of the ratio model to calculate σ_i for each project, $i = 1, \ldots, N$. Typically, $\sigma_i = \sigma_0 \, x_i^{\gamma}$ where γ is a set value, often 0.8. The value of σ_0 can be calculated from the assumed value of the error ratio using the equation given previously. Sort the list by increasing σ_i. For each $i = 1, \ldots, N$ calculate the cumulative sum of the σ_i, $c_i = \sum_{j=1}^{i} \sigma_i$.

3. Choose the desired number of strata L, (usually three to five) and divide the projects in the sorted list so that the sum of the σ_i is approximately equal in each of the L strata. This can be done by calculating $h_i = INT(L \frac{c_i}{c_N} + 0.99999999)$.

---

[337] If it is also necessary to estimate gamma from the sample, a method is available. See "Estimating regression models with multiplicative heteroscedasticity" (Harvey 1976).

[338] Another method of constructing strata is called Dalenius-Hodges stratification by size. The authors have chosen to emphasize model-based stratification because it is known to provide nearly optimal sample designs for stratified ratio estimation. See *Model Assisted Survey Sampling*, (Sarndal, Swensson et al. 1992).

Here the *INT* function rounds the value down to the nearest integer and 0.99999999 has been added to the equation to keep the last project from being assigned to a new stratum.

Once the strata have been constructed as just described, the sample should be allocated equally to each stratum. If the sample size in a particular stratum exceeds the population size in that stratum, the projects in that stratum should be selected with certainty. If desired, the sample may be increased in the remaining strata so that the sample size is closer to the planned value.

In some applications, it may be desirable to stratify the population by a categorical characteristic of the projects as well as by size. For example, the projects might be stratified by building type, technology, contractor, or region. The underlying principle is that the sample size allocated to each categorical stratum should be proportional to the sum of the σ_i within each stratum. Given the definition of the error ratio, a convenient way to determine the sum of the σ_i within each stratum is to multiply the expected actual savings in each stratum by the error ratio assumed in the stratum. This gives the rule: the sample size allocated to each categorical stratum should be proportional to the product of the expected actual savings in each stratum and the error ratio assumed in the stratum.[339]

Once the sample size has been determined within each categorical stratum, the projects within each stratum should be further stratified by size as described above.

The Expected Statistical Precision for Any Sample Design

This section discusses how to assess the expected statistical precision of the stratified ratio estimator when stratified ratio estimation is used with an arbitrary sample design. These results assume that the ratio model is accurate and that the sample design is truly followed without non-response or other similar problems.

To develop the result of interest, a new concept is needed. For any given sample design, define the inclusion probability π_i to be the probability that project i is included in the sample, for all $i = 1, \ldots, N$ in the population. Assume that the inclusion probability is greater than zero for every project in the population, and that sample size n is fixed. There are two useful facts about inclusion probabilities. First, the population sum of the inclusion probabilities is equal to n. Second, for any stratified sample design, the inclusion probability is equal to the sampling fraction in each stratum.

Now the result: Let b be the stratified ratio estimator. Under the ratio model, the expected value of the standard deviation of b in repeated sampling is approximately

[339] This result can be used to allocate evaluation resources among a portfolio of programs, especially if the marginal evaluation cost per sample project is approximately the same for all projects in the portfolio. See the chapter on Uncertainty.

$$sd(b) = \frac{1}{X}\sqrt{\sum_{i=1}^{N}\left(\pi_i^{-1} - c\right)\sigma_i^2}$$

Here c is 1 if the finite population correction is desired, or 0 if not.[340] Under the ratio model, the expected relative precision can be defined to be $rp = sd(b)/\beta$.

The preceding equation can be used to assess the expected relative precision for any stratified sample design under the ratio model. This methodology can be used, for example, to explore the effect of increasing the number of strata. This type of analysis indicates that three to five model-based strata are adequate in most impact evaluation applications. This equation has also been used to explore the effect of using model-based stratification with a set value of gamma that is smaller than the value assumed in the ratio model. In several evaluation applications, it has been shown that there is very little loss in expected statistical precision if the strata are constructed using a gamma of 0.5 when the value in the secondary equation is 0.8. This tends to decrease the sampling fractions in the strata containing larger projects. This can sometimes facilitate recruiting and data collection.

Using the preceding equation, a sample design is said to be optimal under the assumed ratio model if the inclusion probabilities minimize $sd(b)$ for a given sample size n. It can be shown that a sample design is optimal if and only if $\pi_i = n\,\sigma_i \Big/ \sum_{i=1}^{N}\sigma_i$ provided this is not greater than 1. If $n\,\sigma_i \Big/ \sum_{i=1}^{N}\sigma_i > 1$, then project i should be selected with certainty.[341]

Applicability to Impact Evaluation

Stratified ratio estimate also relies on the assumptions that the sample design is followed and that the true savings are measured for the sample projects with little or no bias, as discussed in the section on simple random sampling. Since the sample can generally be smaller with stratified ratio estimation than with simple random sampling, it should be possible to give even more attention to minimizing bias from self-selection, non-response, deliberate substitution of sample projects, or systematic measurement error.

---

[340] For example, the finite population correction might not be suitable if random measurement error is a large contributor to the residual standard deviation of each project.

[341] Under the ratio model, $\sum_{i=1}^{N}\mu_i = \beta\,X$. This result can be used to show that if $\pi_i = n\,\sigma_i \Big/ \sum_{i=1}^{N}\sigma_i$ for all projects in the population and $c = 0$, then $rp = 1.645\,\frac{sd(b)}{\beta} = 1.645\,\frac{er}{\sqrt{n}}$. This justifies our use of the error ratio to calculate the estimated relative precision assuming that a ratio estimator is used with an efficiently stratified sample design and a large population.

Stratified ratio estimation is generally especially effective when simple random sampling is inappropriate. Whenever the coefficient of variation of savings is greater than one, stratified ratio estimation should be considered. Stratified ratio estimation will almost always be more effective than simple random sampling if the program provides good tracking estimates of savings.

Stratified ratio estimation often focuses on the relationship between the tracking estimates of savings and the actual savings. The two key parameters are the realization rate and error ratio. The realization rate is the slope of the trend line. It is the ratio between the average or total value of the actual savings and the average or total value of the tracking estimates. Thus, the realization rate reflects the amount of systematic bias in the tracking estimates of savings.

The error ratio, on the other hand, describes the strength of the association between the tracking estimates of savings and the actual savings, i.e., the variation of actual savings around the trend line associated with the realization rate. The error ratio measures whether the tracking savings are accurate from project to project across the population of projects.

The error ratio is a useful indicator of the quality of the program delivery system. Well-designed and managed programs will tend to have smaller error ratios than programs with poorer control and less attention to detail. Indeed, if the error ratio is found to be higher than expected, it generally indicates that there is a problem with program delivery. Conversely, stratified ratio estimation tends to reward strong programs, i.e., those with relatively small error ratios, by making it possible to carry out an effective impact evaluation using a relatively small sample.

With stratified ratio estimation, the ratio model has been used to assist in the development of a suitable sample design. It is important to understand, however, that the model is only used to develop the sample design. The model is not used to support the statistical analysis of the sample data, except the estimation of the error ratio. If the model is accurate, the achieved statistical precision will be close to the expected statistical precision predicted by the model. If the model is inaccurate, the expected statistical precision may be inaccurate also. But even if the model is inaccurate, the stratified ratio estimator is still free of any material bias and the standard error is still a good guide to the achieved statistical precision.[342]

[342] Sarndal has referred to these methods as model-assisted since, although the analysis does not depend on the accuracy of the model, the model does guide the analysis. (Sarndal, Swensson et al. 1992), pp. 227 and 239. Sarndal provides a much more general model called the generalized regression model which may, in some circumstances, suggest other estimators such as the difference or regression estimators, but the authors have found that the ratio estimator generally is suitable in evaluation.

A Sampling Roadmap

Building on the preceding foundations, a sampling roadmap is now provided, i.e. one recommended way of approaching statistical sample design and analysis in a typical evaluation study. This map can provide effective guidance for most evaluation studies, but occasionally a special circumstance may arise that will call for some other approach.

Figure 13.11 provides an overview of the overall roadmap. Each step will be discussed below in sequence, starting with the population database.

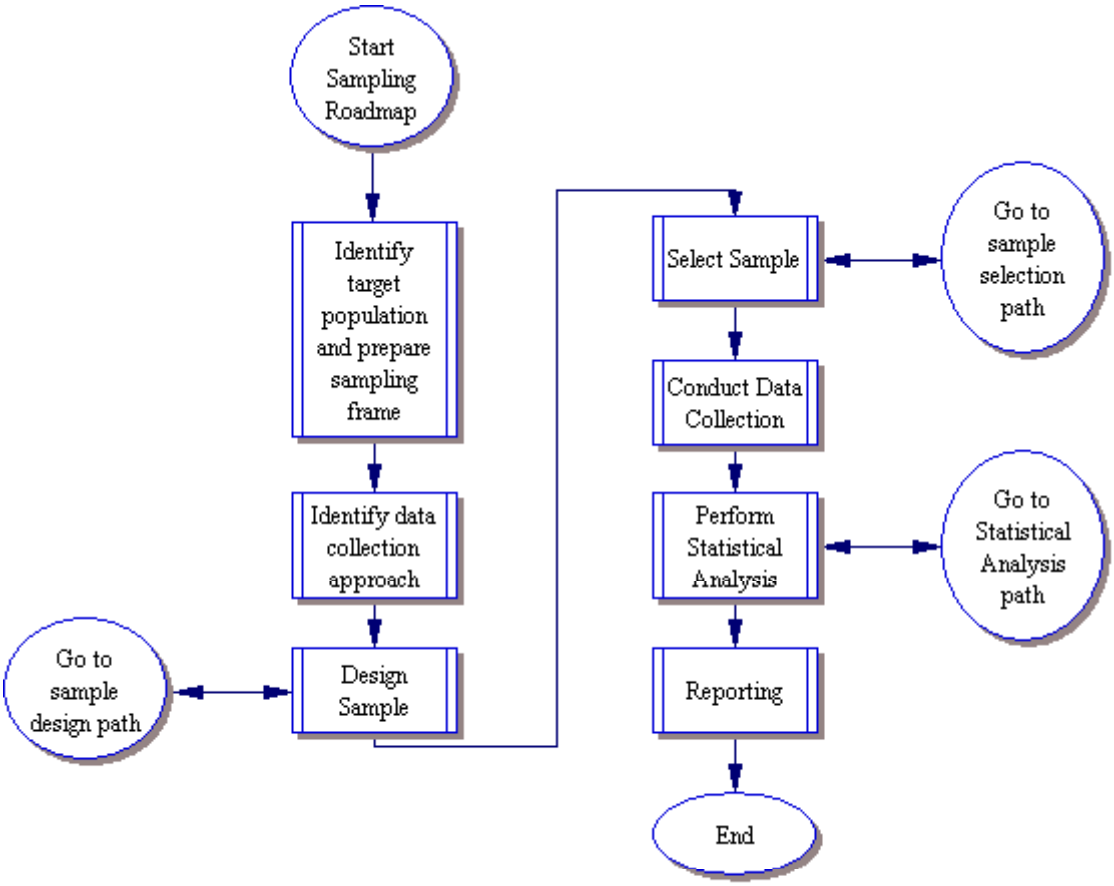

Figure 13.11: Overall Sampling Roadmap

The Population Database

The first step in sample design is usually to develop a population database (sometimes called the sampling frame) listing each unit in the population and providing relevant information for each unit. For example, in an impact evaluation study, the population may be the set of projects implemented in the program in a given year. In a baseline study the population may be the set of all non-residential customers served by a given utility or the set of all vendors of a given appliance in a specified region.

The population database can often be extracted from a program's tracking system[343] or a utility's billing system. The sampling frame can also come from directories of businesses or other lists, as long as these provide a complete representation of the population being studied. The data should be imported into a spreadsheet such as Microsoft Excel or Lotus 123.[344]

Some consolidation or reorganization may be required to make the population database more suitable for sampling, data collection and M&V analysis. The population database should have a single record for each unit – project, customer, vendor, etc., for data collection and analysis. For each unit, the population database should include a unique ID, relevant descriptive information, and available contact information.[345] The population database should also contain a suitable measure of the size of each unit such as the tracking estimate of the kWh savings of the project, the annual consumption of the customer, or the total sales of the vendor.

In addition to the population database itself, there usually is more detailed information for each unit in the population. For example, if the population is a set of projects, the project files should provide descriptive information for each installed measure and its physical location in enough detail that the measure can be located at the site during an onsite audit. This information does not have to be stored electronically but it should be available to the evaluation team for the projects that are included in the sample.

Step 1: Identify the target population and prepare the sampling frame.
Prepare a spreadsheet database listing all units in the population (projects, customers, etc.) with supporting information for each unit. Each row should be a separate unit in the population. Separate columns should provide: ID • Descriptive information • Utility account number (to link the project with utility billing data) • Measure of size (e.g., tracking estimate of savings or annual consumption from billing data) • Contact information • Other relevant information

[343] An adequate program tracking system is essential to effective program delivery and control, in addition to effective evaluation. But the design of a tracking system is beyond the scope of this chapter.

[344] While Excel will be used to illustrate the methods of data management and analysis discussed in this chapter, it is often advantageous to implement the analysis in SAS or another suitable statistical analysis system.

[345] The completeness and accuracy of the contact information will affect the cost of data collection. Depending on the type of program and type of evaluation study, you may need, for example, contact information for the facilities' manager or O&M supervisor, or suitable e-mail addresses if an e-mail survey is planned.

Sample Design

Figure 13.12 summarizes the sample design path of the sampling roadmap. In impact evaluation, there are three primary sampling approaches:[346]

1. A census,
2. Simple random sampling, or
3. Stratified ratio estimation.

Census
A census of all units in the population may be used if the population is very small (i.e., 50 or fewer units) and the unit cost of data collection is so low that the added complexity of sampling is not justified. The use of a census does not mitigate the danger of bias from non-response, measurement error, etc. Therefore the decision to attempt a census should not be made lightly. It is often better to use sampling to limit the data collection to a smaller number of units so that added attention can be devoted to each unit.

Simple Random Sample
A simple random sample may be the most reasonable method if the coefficient of variation of the size of the units in the population is less than one and the unit cost of data collection is so moderate that the added complexity of stratified ratio estimation is not justified.

[346] These approaches are expected to handle most applications. Other methods such as cluster sampling, multi-stage sampling, and sequential sampling may be advantageous in some circumstances.

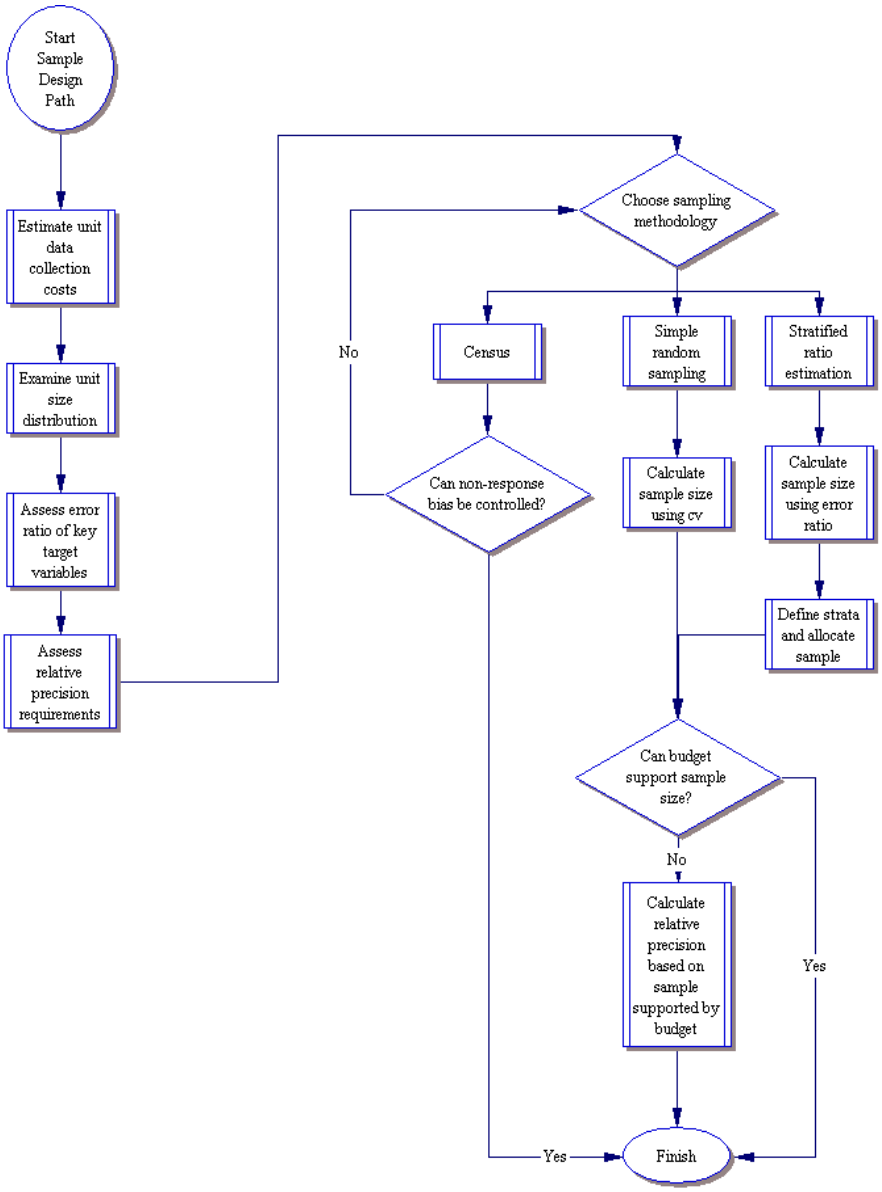

Figure 13.12: Sample Design Path

In this case the primary sample-design decision is the choice of the sample size. For any assumed sample size, the expected statistical precision can be estimated from the coefficient of variation and the number of projects in the population. The sample size can be chosen by considering the unit cost of data collection, the available budget, and the expected statistical precision. If the desired statistical precision is known, the sample size can be calculated directly.

Stratified Ratio Estimation

Stratified ratio estimation will usually provide improved statistical precision or reduced sample sizes compared to simple random sampling, provided that a measure of variability called the error ratio is smaller than the coefficient of variation. In this case, stratified

ratio estimation may be the preferable method unless the reduction in sample size does not warrant the added complexity. The error ratio should be taken from a prior study of the program, from a similar study, or from an assessment of the accuracy of the tracking estimates.

For any assumed sample size, the expected statistical precision can be calculated from the error ratio and the number of projects in the population. As under simple random sampling, the sample size can be chosen by considering the unit cost of data collection, the available budget, and the expected statistical precision. If the desired statistical precision is known, the sample size can be calculated directly.

With stratified ratio estimation, it is necessary to construct strata reflecting the size of each unit in the population, e.g., the tracking estimate of savings or the annual consumption of the customer. If desired, qualitative strata can also be developed.

Steps in Developing the Sample Design

The following discussion describes the sample-design steps shown in Figure 13.12.

Step 1) Develop the general approach to the data collection and project-specific M&V analysis.
a) Develop the most appropriate approach to the data collection and M&V analysis, e.g., telephone survey, onsite audit, site-specific simulation, etc. b) Estimate the average cost for the typical unit in the population. You may need to estimate the sample size as specified below to assess the average cost. c) If there are less than 50 units in the target population and the data collection costs are moderate, a census may be planned and the following steps may be omitted.

Step 2) Examine the distribution of size of the units in the population.
a) Select a suitable measure of the size of each unit. If the population is a set of projects, the tracking estimate of savings should be used. For non-participant surveys and billing analysis, annual consumption may be used. b) Calculate the mean, standard deviation and coefficient of variation of size. The coefficient of variation is the standard deviation divided by the mean. c) If the coefficient of variation is less than 1 and the data collections costs are moderate, then simple random sampling can be used and Step 3 and Decision 1 can be omitted. d) This step may be omitted if stratified ratio estimation is planned.

Step 3) Assess the error ratio for the key target variable(s) to be collected in the study.
a) If possible, use an error ratio from a similar prior study of this program or a similar program. b) If no prior experience, assume an error ratio of 1. c) In most applications, the tracking kWh savings should be used as the x-variable.

Decision 1) Choose the type of sampling methodology, either simple random sampling or stratified ratio estimation.
Criteria: • If the error ratio from Step 3 is smaller than the coefficient of variation from Step 2, use stratified ratio estimation unless the reduction in sample size does not offset the added complexity of sample design and analysis. • Otherwise, use simple random sampling. • If using simple random sampling, complete steps 4 and 5 and then go on to sample selection.

Step 4) Assess the desired relative precision.
Criteria: 1. The size of the program, 2. The error ratio or coefficient of variation in the program, 3. The total evaluation resources available within the portfolio, 4. The optimal allocation of evaluation resources among the programs in the portfolio. Notes: a) It can be helpful to calculate the expected relative precision for various sample sizes. b) If simple random sampling and the 90% level of confidence are used, the expected relative precision is given by the equation $rp = 1.645 \frac{cv}{\sqrt{n}}$ provided that the number of units in the population, denoted N, is large compared to the planned sample size, n. Otherwise $rp = 1.645 \sqrt{1 - \frac{n}{N}} \frac{cv}{\sqrt{n}}$. Here cv is the population coefficient of variation of the target variable of interest, which generally can be estimated as the coefficient of variation of the measure of size in the population. c) If stratified ratio estimation is used, the preceding equation can be used with the error ratio instead of the coefficient of variation. d) This step may be skipped if the evaluation budget is already given for the study being planned.

Step 5) Calculate the required sample size to achieve the desired relative precision.
a) If the evaluation budget is already given for the study being planned, the sample size can be calculated from budgetary considerations. In this case, the expected relative precision should be calculated and documented. b) If the desired relative precision has been specified, the sample size should be calculated. If simple random sampling is used and if D is the desired relative precision at the 90% level of confidence, then the required sample size is $n = \left(\frac{1.645\ cv}{D}\right)^2$ provided that the number of units in the population, denoted N, is large compared to the planned sample size, n. Otherwise, calculate n_0 using the preceding equation, and then apply the finite population correction $n = \frac{n_0}{1 + n_0/N}$. c) If stratified ratio estimation is used, then the preceding equations can be used with the error ratio instead of the coefficient of variation. d) If the indicated sample size is not sufficiently smaller than the population to justify the added complexity of sampling, a census may be used. The use of a census does not mitigate the danger of bias from non-response, measurement error, etc.

Decision 2) If stratified ratio estimation is used, will qualitative strata be used? Yes or No
The target population can, if desired, be divided into two or more qualitative strata based on any information that is available for all units in the population, e.g., building type, type of measures installed, climate zone. Criteria: • Qualitative stratification may provide added control over the distribution of the sample. • Qualitative stratification may improve statistical precision if error ratios vary substantially between qualitative strata. • Sample size in each qualitative stratum should not be too small (e.g., less than 25), so usually only possible if a relatively large sample is planned. • Qualitative stratification usually adds complexity to the analysis. • In the analysis, each qualitative stratum can be considered individually and the results combined across the strata using methods similar to the methods discussed for assessing the total across a set of projects within a portfolio. • Meaningful results can often be developed for some subsets of the sample even if these subsets are not set up as qualitative strata in the sample design.

Step 6) If stratified ratio estimation, develop qualitative strata (optional).
1. Divide the units in the population into unique qualitative strata based on any relevant characteristics (type of building, climate zone, etc.). 2. Tabulate the number of units and total size (e.g., tracking savings) in each qualitative stratum. 3. Assess the error ratio in each stratum (may be equal to the overall error ratio) and allocate the sample to each stratum in proportion to the product of the total size and the error ratio in each stratum.

Step 7) Construct the size strata (required under stratified ratio estimation).
1) Choose the number of size strata, denoted L, to be used overall or within each qualitative stratum (if used). Usually L is in the range 3-5. 2) Use the values of size in the population database to construct the strata overall or within each qualitative stratum (if used). 3) Construct the strata: a) Assume a value for the beta parameter of the ratio model, usually $\beta = 1$ and calculate $\mu_i = \beta\, x_i$ for each case in the population database. Here x_i is the measure of size of case *i*. b) Assume a value for the gamma parameter of the ratio model, usually $\gamma = 0.8$ and calculate x_i^γ for each case in the population database. Here x is the measure of size of each project *i*. c) Calculate the σ_0 parameter of the ratio model using the equation $\sigma_0 = er \dfrac{\sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} x_i^\gamma}$. Here er is the error ratio from step 3. d) Calculate $\sigma_i = \sigma_0\, x_i^\gamma$ for each case in the population database. e) Sort the population database by increasing values of σ_i. f) For each case i in increasing order, calculate the cumulative sigma $c_i = \sum_{j=1}^{i} \sigma_j$. g) Divide the projects in the sorted database so that the sum of the σ_i is approximately equal in each of the L strata. In Excel this can be done by calculating $h_i = INT\left(L \dfrac{c_i}{c_{Ni}} + 0.99999999 \right)$. 4) Allocate the overall sample or the sample allocated to each qualitative stratum (if used) equally among the size strata in each stratum.

Sample Selection

Figure 13.13 summarizes the sample selection path. As in the preceding example, the population database is assumed to be in a spreadsheet such as Microsoft Excel or Lotus 123. Each row of the spreadsheet must be a separate sampling unit such as a project or customer. The sample can be selected by creating a new column with the value of a random number assigned to each row, sorting the rows according to the random variable, and designating the first rows in the sorted database to be the sample.

With Simple Random Sampling

If simple random sampling is used, the following steps can be followed to select the units to be included in the sample and to prioritize the units to be used if any replacements are required.

Step 1) Assign a random number to each unit in the population database.
• Each row in the spreadsheet must be an individual sampling unit. • In Excel, the function RAND() can be used to assign a random number to each row, using a new column. • In Excel, after the random numbers have been created, their values must be frozen using edit - copy and edit - paste special - values.

Step 2) Sort the units in the sampling frame according to the increasing values of the random number created in Step 1.

Step 3) Use a new column to assign the priority to each unit in the sampling frame in the order established in Step 2. The first unit should have priority 1, the second unit priority 2, etc.

Step 4) Designate the units to be included in the primary sample, and designate the backup units.
• Let *n* denote the number of units to be included in the sample. The primary sample is the first *n* units in the sampling frame in the order established in Step 2. In other words the sample is the units with priority from 1 to *n*. • The replacement or backup units are the next *n* units, to be used in the order established in Step 2. In other words, the unit with priority $n+1$ is the first replacement unit, the unit with priority $n+2$ is the second replacement unit, etc. • The primary sample should be used initially. A replacement should only be used when a primary sampling unit cannot be used. Replacements should be minimized to control the risk of selection bias. All replacements should be thoroughly documented.

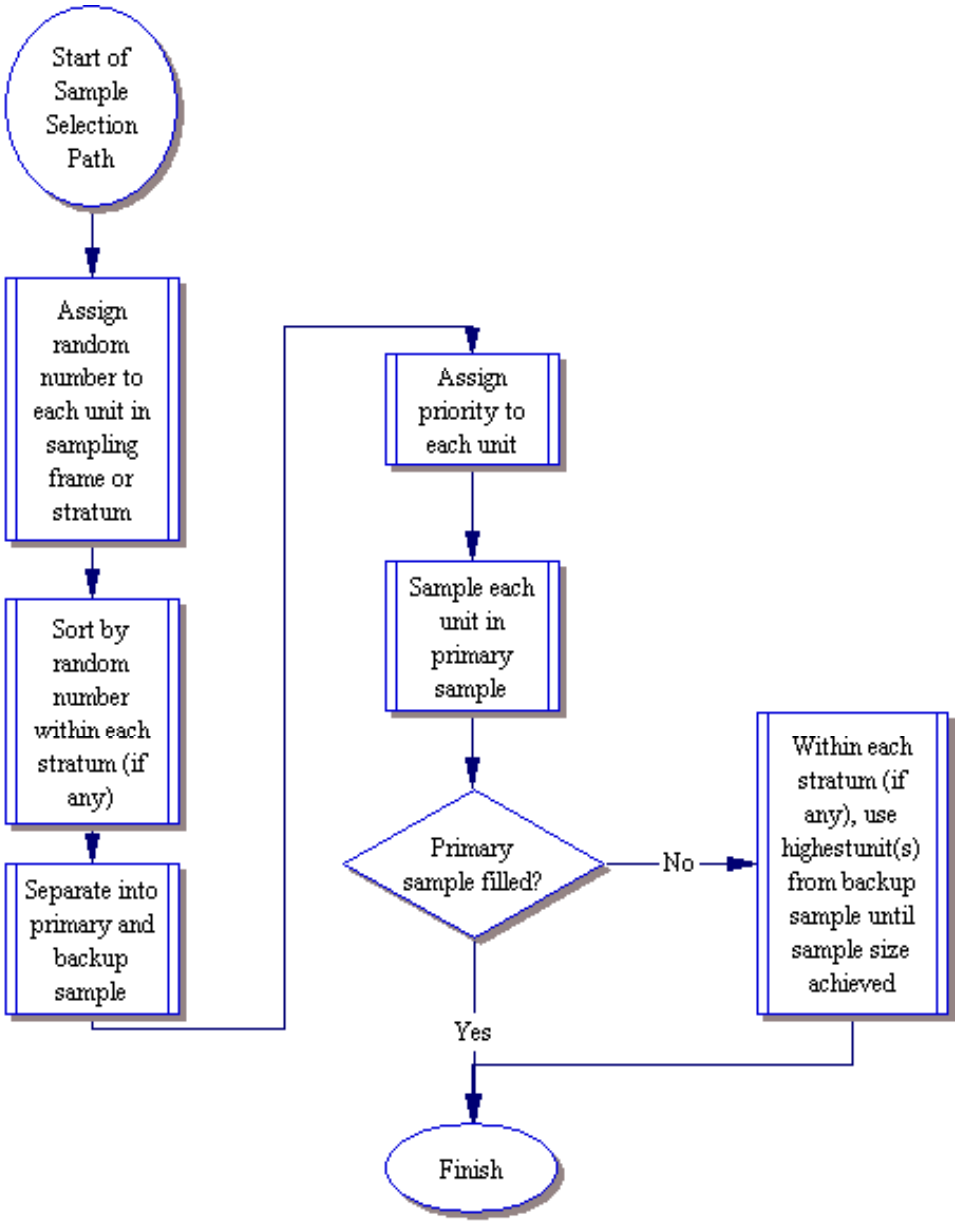

Figure 13.13: Sample Selection Path

With Stratified Ratio Estimation

If stratified ratio estimation is used, the preceding steps can be carried out within each stratum. In particular, the population database should be sorted by stratum and then by the random number within each stratum. In each stratum, the first unit should have priority 1, the second unit priority 2, etc. To the extent possible replacements for any substitutions should be made from within the same strata. If no replacements are available within a given size stratum, the replacement should be taken from the immediately preceding size stratum.

Data Collection and M&V

Once the sample has been selected, the focus switches to the data collection and M&V analysis for each sample unit. The work will vary greatly from one study to another. Some of the data collection and analysis methods include file reviews, telephone surveys, free ridership data collection and analysis, onsite audits, billing data analysis, computer simulation, and short-term monitoring. The most suitable approach will vary depending on the type of evaluation being conducted (impact, process, market) and from study to study.

As discussed in the Uncertainty chapter of this report, most measures of statistical precision assume that the data collection and M&V work is essentially free of both non-response and measurement bias. Therefore it is essential to minimize bias and to report any potential risks of bias in both the recruiting and site-specific data collection and analysis.

Recruiting and Scheduling

If a study involves onsite data collection, the recruiting and scheduling of the sample sites is usually a separate task. This is usually done over the telephone. A special database is generally used to track each attempt to contact the customer. The recruiting staff should be skilled and well-trained in:

- Identifying the person(s) that can provide permission for the work and arrange access to the site,
- Explaining the purpose of the study, the information that is required, and the data collection procedures that will be followed,
- Collecting any added information that may be required such as the location of the measures, the best time for the onsite visit, and the contact person for the visit, and
- Understanding the issue of non-response bias, and how to use the backup sample appropriately.

Using Backups

It is very important to use the backup sample correctly. The most efficient way to recruit a sample of the desired size may appear to be to contact both the primary and backup sample at once and to schedule those sites that are first to respond and agree. But this is generally not sound practice since this approach ensures that the response will be no better than 50%, assuming that the backup sample size is equal to the primary sample size. Instead, the initial recruiting effort should be limited to the primary sample. A backup should be used only if a primary sample site is impossible to contact or refuses to participate.[347]

[347] Sometimes the primary sample site is found to be out of scope. This situation is beyond the scope of this discussion.

The Sample Database

The statistical analysis can often be carried out in a spreadsheet such as Excel or 123. A sample database is created with one row for each unit in the sample and with a separate column for each field of collected data or M&V results.[348] In a typical impact evaluation study, the sample database may include the project ID, the tracking estimate of savings, the gross savings from the M&V analysis, and the net savings from the free ridership analysis.

Quality Control

The goal of quality control is, to the extent possible, to ensure that the sample is free of selection bias and that the collected data are sound and reliable. Three techniques for quality control are especially related to the statistical analysis:

- Monitoring the recruiting, scheduling, and use of backups,
- Reviewing the data that is collected and the M&V analysis that is done for each sample case, and
- Identifying the most influential sample cases.

In order to ensure that problems are detected early enough for timely correction, each of these techniques generally should be done during pre-testing, again soon after data collection has begun, and repeated periodically through the data collection stage of the study.

The primary purpose of monitoring the recruiting, scheduling, and use of backups is to avoid bias from non-response or substitutions and to ensure that the conventional measures of statistical precision are not misleading. As discussed elsewhere, the standard statistical analysis procedures rest on the assumption that the sample is selected strictly following the sample design. Therefore it is important to monitor any deviations from the sample design, e.g., inability to contact the customer due to inaccurate contact information, refusal of the customer to participate in the study, refusal or inability to provide all of the information required for the analysis, etc. If problems are observed, the procedures for implementing the study may need to be adjusted. For example, an incentive may be needed for participating in the study, the recruiting staff may need to be given added training, or the survey may need to be simplified.

The second type of quality control work involves reviewing the data and M&V analysis for each sample case. For example, inconsistencies in the responses to various survey questions may indicate problems with the wording of the question. Similarly, estimates of savings developed in the M&V analysis should be compared to billing data. These issues are discussed further in the M&V section.

It is generally useful to identify the sample cases that have the greatest influence on the results. A close examination of these cases may shed insights into the program itself as

[348] Often, a separate worksheet is used to do the M&V analysis for each sample site and the key results are linked to the sample database.

well as into the data collection and M&V analysis procedures. Two specific techniques can be mentioned – scatter plots and weighted squared errors.

Scatter plots and other graphical techniques can be used to examine the key data elements across the sample. In an impact evaluation, for example, it is generally useful to examine a scatter plot relating the measured savings to the tracking savings. This type of scatter plot can display patterns in the data, particular sample cases that are outliers, and, sometimes, subsets of cases that deviate systematically from the general pattern. These scatter plots can also shed light on the validity of the ratio model that is the foundation for stratified ratio estimation.

With a standard scatter plot, it can be difficult to factor in the case weights that play a role in stratified ratio estimation. The weighted squared error technique addresses this issue. This technique builds on the methods of stratified ratio estimation that are discussed in the following section, but it can be briefly described here. The idea is to calculate the weighted squared error that reflects the contribution of each sample case to the standard error of the ratio estimator, i.e., $w_i(w_i - 1)e_i^2$. The preceding notation is defined in the next section. The sample cases with the largest weighted squared error can be regarded to be the most influential cases. By sorting the sample data according to the value of the weighted squared error, these cases can be easily identified and examined more closely. This work can easily be carried out in a spreadsheet and summarized either in a table or using a suitable graphical technique.

Statistical Analysis

The appropriate statistical analysis will vary depending on the goals and special characteristics of each study. This section will emphasize the methods used to extrapolate the results from the sample to the population and to estimate an error bound for the results. If stratified ratio estimation is used, it is also important to use the sample data to estimate the error ratio to help plan future studies. Figure 13.14 summarizes the steps in the statistical analysis path.

Regardless of the sample design path – census, random, or ratio – the issue of bias should be addressed in the analysis. If a census was attempted and has been essentially complete, or if sample data have been developed for essentially all of the original sample units with almost no refusals or replacements, and if there is no material measurement error in data collection and site-specific analysis, then the statistical analysis is relatively straightforward. Otherwise, the analysis and report should address the possibility of bias.

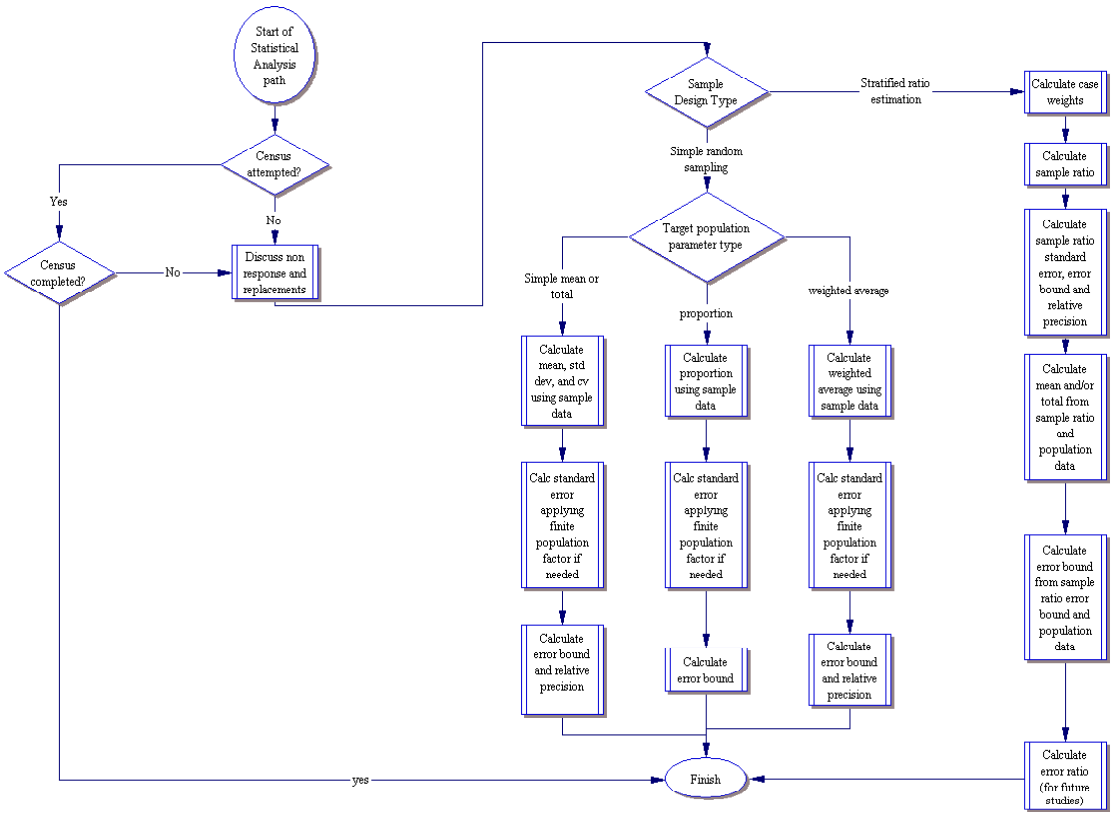

Figure 13.14: Statistical Analysis Path

With a Census

If a census has been taken of all units in the population, the results can be reported with no special statistical analysis. For example, the total evaluated savings and average savings per project can be calculated and reported from the sample data.

If a census was attempted but was not completed due to refusals or other reasons, then the analysis is more complex. The key decision is whether the included units can be regarded as representative of the units that were omitted from the census.

For example, in a persistence study, if a census was attempted but some program participants were excluded from the study because they could not be located, it may be more appropriate to write off the corresponding savings than to assume that these participants are similar to the participants included in the sample.

On the other hand, if the exclusion of the project from the census can be assumed to be statistically independent of the information to be collected, then the projects included in the study may be regarded as if they are a true sample. For example, suppose that in a persistence study, some program participants could not be included in the study because they refused to cooperate. If it seems plausible that the refusal was statistically independent of the persistence of the savings, then the participants that did cooperate can be regarded as a sample of all participants. In this case, either simple random sampling

or stratified ratio estimation should be used to analyze the data, depending on the coefficient of variation and error ratio.

With Simple Random Sampling

If simple random sampling has been used and there have been a limited number of replacements or non-respondents, the analysis is usually relatively straightforward. Otherwise, the disposition of the original sample should be discussed. Any non-respondents or replacements should be explained. The possibility of selection or non-response bias should be discussed in the report.

The methodology used in the statistical analysis depends on whether the parameter to be estimated is a mean or total, a proportion, or a weighted average or ratio. A mean or total is often used to summarize a quantitative variable such as the measured savings of each project in the population. A proportion is often used to summarize a qualitative variable such as whether or not a program participant is satisfied with a particular measure. A weighted average is often more appropriate than a simple mean, especially if the units vary substantially by size. For example the average efficiency of a population of air conditioners is generally defined to be the weighted average of the efficiency of each air conditioner, using a measure of the size of each unit such as tons as the weight. A ratio is often of interest in evaluation. For example, the gross realization rate is the ratio between the total gross savings of all projects in the population and the total tracking savings.

A population mean can be estimated by calculating the associated sample mean. The corresponding standard deviation and coefficient of variation should also be calculated. Then the standard error, error bound and relative precision are calculated in order to characterize the statistical precision of the sample mean as an estimator of the population mean. A population total can be estimated by multiplying both the sample mean and the corresponding error bound by the number of units in the population.

A population proportion can be estimated by calculating the associated sample proportion. Again the standard error and error bound are calculated in order to characterize the statistical precision of the sample proportion as an estimator of the population proportion

A weighted average can be estimated using the corresponding weighted average of the sample data. As usual, the standard error, error bound and relative precision are calculated in order to characterize the statistical precision of the sample weighted average as an estimator of the population weighted average. A ratio can also be estimated using these same techniques.

Decision 1) Is the target population parameter a simple mean or total, a proportion or a weighted average? Go to path 1, 2 or 3 respectively.

Path 1) For a simple mean or total:

- Calculate the mean, sample standard deviation and coefficient of variation using the sample data.
- Calculate the standard error using the equation $se = \frac{s}{\sqrt{n}}$. Here s is the sample standard deviation and n is the number of units in the sample.
- If the sample is more than 5% of the population, multiply the preceding result by the finite population correction factor $\sqrt{1 - \frac{n}{N}}$. Here N is the number of units in the population.
- Calculate the error bound at the 90% level of confidence using the equation $1.645\ se$.
- Calculate the relative precision by dividing the error bound by the mean
- If the population total is to be estimated, multiply the sample mean and the error bound by N, the number of units in the population.

Path 2) For a proportion:

- Calculate the proportion, denoted $\hat{p}$, using the sample data.
- Calculate the standard error using the equation $se = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- If the sample is more than 5% of the population, multiply the preceding result by the finite population correction factor $\sqrt{1 - \frac{n}{N}}$. Here N is the number of units in the population.
- Calculate the error bound at the 90% level of confidence using the equation $1.645\ se$.

Path 3) For a weighted average or ratio:

- Write the weighted average or ratio in the population as $B = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$. For example, if the ton-weighted average efficiency of a population of air conditioners is of interest, x_i is the size in tons of air conditioner i and y_i is the product of x_i and the efficiency of air conditioner i.
- Calculate the corresponding sample weighted average or ratio using the equation $b = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} x_i}$.
- Calculate the standard error as $se = \frac{\sqrt{\frac{\sum_{i=1}^{n} (y_i - b\,x_i)^2}{n(n-1)}}}{\frac{1}{n}\sum_{i=1}^{n} x_i}$.
- If the sample is more than 5% of the population, multiply the preceding result by the finite population correction factor $\sqrt{1 - \frac{n}{N}}$. Here N is the number of units in the population.
- Calculate the error bound at the 90% level of confidence using the equation $1.645\ se$.
- Calculate the relative precision by dividing the error bound by the sample weighted average.

With Stratified Ratio Estimation

Within each stratum, the final sample should be compared to the planned sample and any non-respondents or replacements should be reported and explained. Then the case weight should be calculated in each stratum and assigned to each of the sample projects in the stratum.

Just as with simple random sampling, the stratified ratio estimation path can be used to estimate means and totals, proportions, and weighted averages or ratios. The starting point is the estimation of a population ratio by calculating the case-weighted sample ratio. The standard error, error bound and relative precision are calculated in order to characterize the statistical precision of the sample ratio as an estimator of the population ratio.

Finally, the sample data should be used to estimate the error ratio – the key measure of the strength of the association between *x* and *y* in the population for planning future studies using stratified ratio estimation.

As an example, the preceding methods can be used to estimate the gross realization rate which is defined to be the ratio between the true gross savings (*y*) of all projects in the population and the total tracking savings (*x*). The error bound can be calculated to summarize the statistical precision of the sample realization rate, and the error ratio can be calculated to help plan future evaluation studies.

Once the results have been developed for the ratio, they can usually be used to estimate other parameters that may be of interest. For example, the population mean or total of the variable used in the numerator of the ratio (*y*) can be estimated by multiplying both the sample ratio and the corresponding error bound by the population mean or total of the variable used in the denominator of the ratio (*x*), provided that the latter is known. For example, the average gross savings or the total gross savings of all projects in the population can be estimated by multiplying the gross realization rate and the corresponding error bound by the average tracking savings or total tracking savings in the population.

The stratified ratio estimator is also the key to the estimation of results for a particular subset of the population. Suppose for example that an estimate of the gross realization rate is desired for the projects of a particular vendor or those that fall in a given region or market segment. This can be equated to a ratio in which the y-variable is the gross savings of each project in the subset of interest and zero otherwise, and the x-variable is defined to be equal to the tracking savings of each project in the subset of interest and zero otherwise. With these definitions, the usual stratified ratio estimation methods apply. In addition, the total gross savings of the projects of interest can be estimated by multiplying the estimated realization rate by the total tracking savings of the projects.

The preceding methods can also be used to estimate proportions. For example the proportion of all gross savings that is associated with a particular subset of the population can be estimated by dividing the total gross savings of the projects of interest, estimated as described in the preceding paragraph, by the total gross savings of all projects.

Step 1) Calculate the case weight of each sample project, denoted w_i.
• Tabulate the number of population projects in each stratum. • Tabulate the number of sample projects in each stratum. • Calculate the case weight as the number of population projects divided by the number of sample projects, and apply the results to each unit in the sample.

Step 2) For any population ratio or proportion, use the sample data and case weights to calculate the sample ratio.
• The population ratio or proportion is defined to be of the form $B = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i}$ where x and y are any suitable variables. • The sample ratio is calculated as $b = \frac{\sum_{i=1}^{n} w_i \; y_i}{\sum_{i=1}^{n} w_i \; x_i}$. Here w_i is the case weight calculated in step 1.

Step 3) Calculate the standard error of the sample ratio.
• For each unit i in the sample, calculate $e_i = y_i - b\,x_i$ • Calculate the standard error of the sample ratio using the equation $se(b) = \frac{\sqrt{\sum_{i=1}^{n} w_i (w_i - 1) e_i^2}}{\sum_{i=1}^{n} w_i \; x_i}$. • Calculate the error bound at the 90% level of confidence using the equation $1.645\, se$. • Calculate the relative precision by dividing the error bound by the sample ratio.

Step 4) Calculate the corresponding error ratio for use in future sample designs.
• Assume a value for the gamma parameter of the ratio model, usually $\gamma = 0.8$ • For each unit i in the sample, calculate $e_i = y_i - b\,x_i$ as in step 2. • For each unit i in the sample, calculate x_i^{γ}. • Estimate the error ratio between x and y in the population using the equation $\hat{er} = \frac{\sqrt{\left(\sum_{i=1}^{n} w_i \; e_i^2 / x_i^{\gamma}\right)\left(\sum_{i=1}^{n} w_i \; x_i^{\gamma}\right)}}{\sum_{i=1}^{n} w_i \; y_i}$.

Step 5) If the population mean or total of y is of interest, and the population mean or total of x is known, use the sample ratio to obtain the corresponding estimate and error bound.
• Estimate the population mean or total of y by multiplying the population mean or total of x by the sample ratio developed in Step 1. • Calculate the corresponding error bound by multiplying the population mean or total of x by the error bound developed in Step 2.

Step 6) Sometimes there is an interest in the population ratio for a particular subset S of the population, $B^* = \frac{\sum_{i \in S} y_i}{\sum_{i \in S} x_i}$. Here $i \in S$ is read, "i is an element of S".
• Define an indicator variable I_i to be 1 if i is an element of S, and 0 otherwise. • Define $y_i^* = I_i \; y_i$ and $x_i^* = I_i \; x_i$. • Redefine the population ratio of interest to be $B^* = \frac{\sum_{i=1}^{N} y_i^*}{\sum_{i=1}^{N} y_i}$ • Estimate B^* as specified in steps 2 and 3. • Estimate the population total of y for the subset S as specified in step 5. • Then, if desired, calculate the proportion of the grand total of y that is associated with the subset S.

Reporting Guidelines

There should be some standardization of how results are reported. The present discussion recommends reporting guidelines from the point of view of sampling and statistical analysis. The following guidelines are general and should be tailored to each project.

Findings

This section of the report will generally start by discussing the objectives of the study and characterizing the population using information taken from the population database. For example, in the case of an impact evaluation study, the number of projects and tracking estimates of savings should be described. Then the findings should be presented, with confidence intervals provided for the important results at the 90% level of confidence. Finally, the relevant sample-design parameters such as the coefficients of variation and/or error ratio should be reported for the key results to guide the planning of future studies.

Data Collection and M&V Methodology

This section should describe all data collection and M&V methodology used in the study. All data-collection instruments should be provided, usually in an appendix to the report. An assessment of the risk of measurement bias should be included in this section.

Statistical Methodology

This section should document the sampling path: e.g., census, random, or ratio; give the parameters used to develop the sample design such as the assumed coefficient of variation and/or error ratio; and discuss the expected relative precision at the 90% level of confidence. This section should also discuss any substantial difference between the assumed sample-design parameters and the estimates found from the sample data for the key results. If a stratified sample design is used, each stratum should be described, giving the population statistics for each stratum (number and total size), and the specified number of sample projects in each stratum.

The disposition of the sample should also be discussed, documenting all non-responses, refusals, substitutions, etc. If a stratified sample design was followed, this information should be provided for each stratum. If a census or random sample has been attempted, the report should describe any deviations from the intended sample by size category to help shed light on any selection bias. In any case, the report should discuss the underlying reasons that are believed to have contributed to the incomplete execution of the intended sample selection. Building on this information, the danger of selection bias should be discussed.

Finally the statistical analysis methodology should be described. Any deviations from the methodology in this chapter should be discussed.

Appendix to Chapter 13

Example and Use of Stratified Ratio Estimation for Sampling

This appendix provides an example of the stratified ratio estimating procedure, following the step-by-step sampling roadmap. The description of the methodology and steps are not repeated here in the interest of brevity. This example can be followed alongside the material above.

In this particular program, lighting retrofits were implemented in 1,248 small commercial sites. Under the program, efficient lighting was installed in relatively small non-residential sites. Very few lighting control measures were installed.

The objective of the study was to select a sample of projects from the program, assess the gross and net savings of each sample project, and then extrapolate the findings back to the population of all 1,248 projects in the population.

A population database was developed from the program tracking system. Table 13.20 shows a small portion of the population database. The table shows the first ten projects in the database as well as the last nine projects. For confidentiality reasons, the customer account number and contact information have been deleted from the database.[349]

Step 1 – Plan the General Approach to Data Collection and M&V

Based on prior experience, it was felt that onsite auditing and M&V engineering analysis would be the best approach to evaluating the energy and demand savings of this particular program. In a prior study of this program, short-term monitoring was used quite extensively.[350] In the prior study the hours of use estimated from the monitoring were very highly correlated with reported hours of use developed on a space-specific basis with almost no bias. So in the current project, no end use monitoring was planned.

However, fairly detailed onsite data collection and project-specific M&V analysis were planned. For each project in the sample, the analysis would start with a detailed review of the project file and review of the methods used to calculate the savings. Any apparent errors would be corrected and the savings recalculated. Then each sample site was to be visited, and a detailed onsite audit would be carried out with a focus on the installed lighting equipment. The type of fixtures and the associated wattage would be recorded and compared to the information in the file. The savings would be recalculated if any differences were observed. Then the number of each fixture type that was actually installed would be compared to the reported quantity and the savings would be

[349] The authors are grateful to Lisa Shea and NSTAR Electric & Gas for allowing the use of this example. Minor modifications have been made to simplify the example. But the example accurately reflects an actual study carried out for NSTAR by RLW Analytics.

[350] "Evaluating the Underserved Small C&I Market: Building a Bridge to Implementation," (Ledyard 2003).

recalculated. Finally the hours of use of each type of space served by the installed lighting would be assessed and the kWh savings would be recalculated. The result was taken to be the gross savings.

Table 13.20: The Population Database

Project	Vendor	Trk kWh
1	4	3,117
2	4	3,056
3	2	2,668
4	7	14,710
5	4	11,142
6	4	1,482
7	7	3,974
8	2	49,782
9	2	45,659
10	4	4,206
11	4	1,750
12	2	6,369
1238	4	51,579
1239	2	28,496
1240	4	11,737
1241	4	2,000
1242	4	129,205
1243	4	16,520
1244	4	44,611
1245	4	106,948
1246	4	624
1247	4	18,188
1248	4	52,225

Based on the file review, the onsite visit and an interview with the program participant, the free ridership would be assessed for each measure installed in the project. This information would be applied to the gross savings to calculate the net savings adjusted for free ridership. Finally, all of these sources would be used to identify any efficiency measures taken that were not rebated through the program but were the result of the program, and the savings would be recalculated adjusted for the participant spillover.

Similar information would be developed for the summer and winter (as applicable) demand saving, but these results will not be discussed here due to space limitations.

It was estimated that the preceding M&V work would cost about $1,000 per sample project or about $50,000 if the sample contained 50 projects.

Step 2 – Examine the Distribution of Size of the Projects in the Population
The kWh savings recorded in the tracking system was taken to be the measure of the size of each project. This variable is referred to as the tracking savings. In the database it is labeled as Trk kWh.

Table 13.21 summarizes the tracking savings provided in the population database. The table shows that the 1,248 projects had a total tracking savings of 20,119,315 kWh. They have an average tracking savings of 16,121 kWh per project and a standard deviation of 23,333 kWh per project. The coefficient of variation of the tracking savings is 1.45. The smallest project has 48 kWh of tracking savings and the largest project has 345,350 kWh of tracking savings.

Table 13.21: Summary of Tracking Savings in Population Database

Population Size	1,248	Projects
Total Tracking Savings	20,119,315	kWh
Average Tracking Savings	16,121	kWh per Project
Standard Deviation	23,333	kWh per Project
Coefficient of Variation	1.45	
Minimum Tracking Savings	48	kWh
Maximum Tracking Savings	345,350	kWh

Table 13.22: The Number of Projects by Tracking Savings

Count of Project	
Trk kWh	Total
1-10000	691
10001-20000	269
20001-30000	100
30001-40000	65
40001-50000	38
50001-60000	26
60001-70000	19
70001-80000	7
80001-90000	4
90001-100000	8
100001-110000	9
110001-120000	5
120001-130000	3
150001-160000	1
160001-170000	1
180001-190000	1
340001-350000	1
Grand Total	1248

Table 13.22 shows the population distribution of the tracking kWh savings. The population contains a large number of small projects and a small number of relatively large projects, i.e., the distribution is highly skewed.

Table 13.23: The Skewness of Tracking Savings in the Population

Trk kWh	Projects	Savings
0 - 30,000	85%	45%
30,001 - 100,000	13%	41%
100,001 and up	2%	14%
All Projects	100%	100%

Table 13.23 shows the skewness of savings in the population in another way. The table shows that 85% of all projects have 30,000 kWh or less tracking savings but these projects account for only 45% of all tracking savings. The largest 2% of the projects, those with tracking savings of 100,001 kWh or more, have 14% of all savings. This type of situation is very common in impact evaluation, particularly for commercial and industrial programs. Consequently, stratification by size is generally an extremely important feature of an efficient sample design.

Step 3 – Assess the Error Ratio

Fortunately this program has been offered for several years and was evaluated two years earlier. Although it included short-term end use metering, the prior evaluation used a similar approach to the planned approach. In the prior study the error ratios for gross and net kWh savings were both found to be about 0.4. Since the methodology used in the prior evaluation was comparable and since the program has not changed substantially, an error ratio of 0.4 was assumed to be a good basis for planning the new study.

Choose the Sampling Approach

As noted above, the coefficient of variation of kWh savings was found to be 1.45 and the data collection and analysis costs were expected to be rather high per sample project. Moreover the error ratio was assumed to be 0.4 – much lower than the coefficient of variation. Therefore stratified ratio estimation was clearly preferable to simple random sampling. This decision will be validated in Step 5b.

Step 4 – Assess the Desired Relative Precision

Since stratified ratio estimation is planned, the expected relative precision can be calculated for any given sample size using the equation

$$rp = 1.645\sqrt{1-\frac{n}{N}}\,\frac{er}{\sqrt{n}}\,.$$

In this application, $er = 0.4$ is the assumed error ratio, $N = 1{,}248$ is the number of units in the population, and n is the sample size. For example, if $n = 50$, we calculate

$$rp = 1.645\sqrt{1-\frac{50}{1{,}248}}\,\frac{0.4}{\sqrt{50}} = 0.091\,.$$

This implies that under the stratified ratio estimation methodology, the expected relative precision would be about ±9% at the 90% level of confidence with a sample of 50 projects.

Table 13.24 shows the expected relative precision for various sample sizes. Figure 13.15 shows these results graphically. From these results, it is clearly practical to plan the new study for ±10% relative precision at the 90% level of confidence.

Table 13.24: Expected Relative Precision vs. Sample Size with Stratified Ratio Estimation

Sample Size	*Expected Relative Precison*
30	11.9%
35	11.0%
40	10.2%
45	9.6%
50	9.1%
55	8.7%
60	8.3%
65	7.9%
70	7.6%
75	7.4%
80	7.1%

Step 5a – Calculate the Required Sample Size

The preceding results indicate that a sample of about 40 projects is sufficient to provide the desired ±10% relative precision at the 90% level of confidence. To provide a more complete example, we will illustrate the calculations in greater detail. The required sample size can be calculated using the two equations

$$n_0 = \left(\frac{1.645\, er}{D}\right)^2 \text{ and } n = \frac{n_0}{1 + n_0/N}.$$

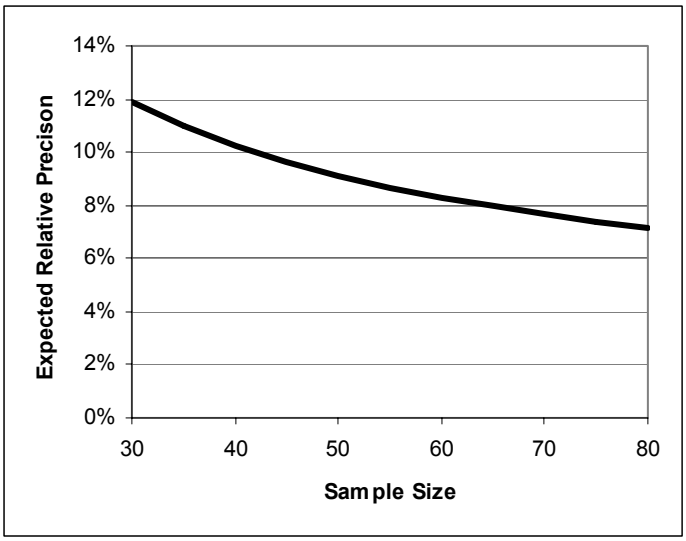

Figure 13.15: Expected Relative Precision vs. Sample Size with Stratified Ratio Estimation

Here $er = 0.4$ is the assumed error ratio, $D = 0.10$ is the desired relative precision at the 90% level of confidence, and $N = 1{,}248$ is the number of units in the population. Thus we calculate

$$n_0 = \left(\frac{1.645 \times 0.4}{0.1}\right)^2 = 43 \text{ and } n = \frac{43}{1 + 43/1{,}248} = 42 \text{ projects.}$$

To provide a conservative margin, a sample of 50 projects was planned.

Step 5b – Compare Stratified Ratio Estimation with Simple Random Sampling
This step is not actually required since the assumed error ratio of 0.4, is substantially smaller than the coefficient of variation of 1.45. In order to provide a more complete example, the comparison to the statistical precision and sample size if simple random sampling is used is still provided.

In these calculations the preceding equations are used but with the error ratio replaced by the coefficient of variation, which was found to be 1.45 in Step 2. Table 13.25 shows the expected relative precision for a range of sample sizes. Figure 13.16 shows these results graphically. These results suggest that to provide ±10% precision at the 90% level of confidence, a sample of almost 400 projects would be needed. In other words, the sample would have to be about ten times as large with simple random sampling instead of stratified ratio estimation. These results are expected given that the projects in the population have a highly skewed size distribution and that simple random sampling ignores the tracking estimates of savings.

Table 13.25: Expected Relative Precision vs. Sample Size with Simple Random Sampling

Sample Size	*Expected Relative Precison*
50	33.0%
100	22.8%
150	18.2%
200	15.4%
250	13.5%
300	12.0%
350	10.8%
400	9.8%
450	9.0%
500	8.2%
550	7.6%
600	7.0%

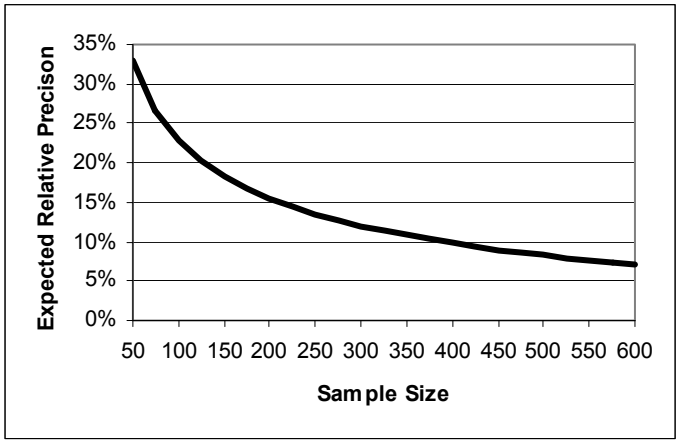

Figure 13.16: Expected Relative Precision vs. Sample Size assuming Simple Random Sampling

Table 13.26 shows a simple comparison of the cost of the two approaches. The table shows that under simple random sampling, 390 projects would be needed to provide ±10% precision at the 90% level of confidence. Assuming an M&V cost of $1,000 per sample project, the total variable M&V cost would be about $390,000. Under simple random sampling, we have assumed that the statistical sample design and analysis cost would be about $10,000, for a total variable cost of $400,000. This neglects the cost of project management, reporting, etc. that can be assumed to be unaffected by the sampling approach.

Table 13.26: Cost Comparison of Simple Random Sampling vs. Stratified Ratio Estimation

Common Parameters	
Desired Relative Precision	10%
Population Size	1,248
Average M&V Cost per Project	$1,000
Simple Random Sampling Approach	
Coefficient of Variation	1.45
Sample Size before Finite Population Correction	567
Sample Size after Finite Population Correction	390
Total Variable M&V Cost	$390,000
Statistical Sample Design and Analysis	$10,000
Total M&V and Statistical Cost	$400,000
Stratified Ratio Estimation Approach	
Error Ratio	0.4
Sample Size before Finite Population Correction	43
Sample Size after Finite Population Correction	42
Total Variable M&V Cost	$42,000
Statistical Sample Design and Analysis	$30,000
Total M&V and Statistical Cost	$72,000

By contrast, under the stratified ratio estimation approach, the required sample is 42 projects so the total M&V cost would be about $42,000. Under stratified ratio estimation, the statistical sample design and analysis are assumed to cost $30,000, for a total variable cost of $72,000. In this example, simple random sampling is not a feasible choice even though stratified ratio estimation requires more statistical expertise. In other applications, this type of comparison may help determine which approach is preferable.

Step 7 – Construct the Size Strata

Stratification by size is required whenever the projects in the population vary by size. Usually three to five strata are required to provide adequate control over size. In this example five strata will be used since there is substantial variation in the population. Table 13.27 summarizes the assumptions used to construct the strata.

Table 13.27: Assumptions Used to Construct the Size Strata

Sample Size	50
Number of Strata	5
Beta	1
Error Ratio	0.4
Set Gamma	0.8

Table 13.28 shows a portion of the worksheet used to calculate the stratum of each case in the population. The column labeled μ shows the value of $\mu_i = \beta\, x_i$. The next column shows the value of x_i^{γ}. Using these columns, the σ_0 parameter was calculated using the equation

$$\sigma_0 = er \frac{\sum_{i=1}^{N} \mu_i}{\sum_{i=1}^{N} x_i^{\gamma}}.$$

This provided a calculation: $\sigma_0 = 3.096$. This was used to calculate $\sigma_i = \sigma_0 \; x_i^{\gamma}$ which is labeled Sigma.

Table 13.28: Constructing the Strata

Project	Trk kWh	μ	x^{γ}	Sigma	Sort Str	Cum Sigma	Stratum
450	48	3117	22	69	1	69	1
449	75	3056	32	98	2	166	1
453	75	2668	32	98	3	264	1
458	75	14710	32	98	4	362	1
462	75	11142	32	98	5	460	1
108	157	1482	57	177	6	637	1
1107	112,340	51579	10,976	33,976	1238	7,610,012	5
1186	114,621	28496	11,154	34,526	1239	7,644,539	5
1030	114,931	11737	11,178	34,601	1240	7,679,140	5
516	117,148	2000	11,350	35,134	1241	7,714,274	5
987	120,889	129205	11,639	36,029	1242	7,750,302	5
488	121,117	16520	11,656	36,083	1243	7,786,385	5
1242	129,205	44611	12,275	37,998	1244	7,824,383	5
840	152,005	106948	13,979	43,274	1245	7,867,657	5
522	164,464	624	14,889	46,089	1246	7,913,746	5
1032	184,577	18188	16,328	50,545	1247	7,964,291	5
213	345,350	52225	26,953	83,435	1248	8,047,726	5

Then the database was sorted by increasing sigma. The column labeled Sort Str identifies the order of the cases after they were sorted, i.e., the sort order for stratification. Cum Sigma was calculated using the equation $c_i = \sum_{j=1}^{i} \sigma_j$. Finally, Stratum was calculated using the Excel equation $h_i = INT\left(L \frac{c_i}{c_{Ni}} + 0.99999999 \right)$.

Table 13.29 shows a summary of the sample design. For example, stratum 1 consists of all projects with tracking savings of 10,128 kWh or less. Stratum 1 contains 694 projects with a total tracking savings of 2,892,887 kWh. The average tracking savings of these projects is 4,168 kWh per project. By contrast, stratum 5 consists of the largest 54 projects. These projects have a total tracking savings of 5,265,839 kWh and an average size of 97,516 kWh per project.

The last column of Table 13.29 shows the desired number of sample projects in each stratum. It is characteristic of this type of design that the sample is equally allocated among the strata. Sometimes, however, the sample design calls for more projects that are in the population in the largest stratum. In this case the sample size is set equal to the number of projects. In other words all projects in this stratum are to be selected with certainty.

Table 13.29: Summary of the Sample Design

	Tracking Savings				*Sample*
Stratum	*Projects*	*Max*	*Total*	*Average*	*Size*
1	694	10,128	2,892,887	4,168	10
2	253	18,981	3,502,474	13,844	10
3	151	35,341	3,979,634	26,355	10
4	96	62,056	4,478,481	46,651	10
5	54	345,350	5,265,839	97,516	10
Total	1,248		20,119,315	16,121	50

Sample Selection

Continuing with the preceding example, Table 13.30 shows a portion of the worksheet used to select the sample. All four steps of sample selection were carried out in this worksheet. Note that the values of "Random" were created using the Excel equation = RAND() and then copied and pasted to replace the equation by the actual values so that they would not be recomputed. Then the cases were sorted according the increasing values of Random and the sort order for selection was recorded (Sort for Selection). The Priority was assigned in each stratum. In each of the five strata, the ten projects with priority 1-10 were designated as the primary sample and the projects with priority 11-20 were designated as backups. If a replacement is required in any given stratum, the first unused project in that stratum should be used as the replacement.

Table 13.31 summarized the characteristics of the primary sample. The table shows the number of accounts in each stratum, the total tracking savings, and the average savings per stratum. The table also shows the total and average savings of all 50 of the sample projects. It should be recognized, however, that the total and average across all strata are not meaningful statistics since the sample projects have not been correctly weighted to reflect the sampling fraction in each stratum.

In each of the five strata, the average savings of the sample projects shown in Table 13.31 should be approximately equal to the average savings of the sample projects shown in Table 13.29. The differences should be due only to sampling variation. In this case, the sample average happens to be smaller then the population average in strata 1 and 4 and larger in the remaining three strata. The overall average size of the sample projects is much larger than the average size of the projects in the population since we have deliberately over-sampled the larger projects. In the analysis, the case weights will be used to avoid any bias in the final results.

Table 13.30: Selecting the Primary Sample and the Backup Projects

Project	Trk kWh	Stratum	Random	Sort for Selection	Priority	Sam Size	Sample Category
55	2,283	1	0.0002	1	1	10	Primary
707	2,632	1	0.0009	2	2	10	Primary
108	157	1	0.0045	3	3	10	Primary
777	978	1	0.0049	4	4	10	Primary
257	157	1	0.0054	5	5	10	Primary
263	6,273	1	0.0057	6	6	10	Primary
418	6,840	1	0.0064	7	7	10	Primary
496	1,577	1	0.0077	8	8	10	Primary
871	5,275	1	0.0102	9	9	10	Primary
113	349	1	0.0113	10	10	10	Primary
462	75	1	0.0130	11	11	10	Backup
1013	7,410	1	0.0162	16	12	10	Backup
1241	2,000	1	0.0186	13	13	10	Backup
503	7,517	1	0.0194	14	14	10	Backup
465	1,786	1	0.0226	15	15	10	Backup
1092	5,045	1	0.0255	12	16	10	Backup
475	4,759	1	0.0259	17	17	10	Backup
879	9,010	1	0.0281	18	18	10	Backup
1152	3,590	1	0.0318	19	19	10	Backup
873	6,180	1	0.0324	20	20	10	Backup

Table 13.31: Summary of the Primary Sample

Stratum	*Number*	*Total*	*Average*
1	10	26,521	2,652
2	10	152,375	15,238
3	10	275,066	27,507
4	10	443,479	44,348
5	10	981,100	98,110
Total	50	1,878,541	37,571

In this example, experienced staff was used to recruit and schedule the sites that were selected for the sample. Table 13.32 shows a portion of the database used to track the recruiting process. The full recruiting database contained all of the available contact information, fields for tracking each attempt to reach the primary contact, and other information used to manage the recruiting process.

The recruiting database listed all 100 selected projects in either the primary or backup sample. The projects were listed in order of priority within each stratum. The status field initially indicated whether the project is in the primary or backup sample. The initial recruiting was limited to the primary sample. When the program participant was contacted, the status field was changed to indicate the result of the contact – in this case either 'refused' or 'scheduled'. If contact could not be made for any reason, e.g., because of inaccurate contact information that could not be corrected from other sources, the

status was changed to 'not found'. Note that in stratum 1, three backups were required to schedule ten projects successfully. The three backups were taken from those on the backup list with the lowest priority, i.e., 11 – 13.

Table 13.32: Portion of the Recruiting Tracking Data

Project	Stratum	Priority	Status	Scheduling Note
55	1	1	Scheduled	
707	1	2	Refused	No lighting installed at actual facility
108	1	3	Scheduled	Meet at Rental Office
777	1	4	Scheduled	
257	1	5	Scheduled	A facility person.
263	1	6	Scheduled	Ring the bell to get let in
418	1	7	Scheduled	
496	1	8	Scheduled	
871	1	9	Not Found	Cannot obtain contact info
113	1	10	Scheduled	Talk to Carrie who will set you up with
462	1	11	Refused	
1013	1	12	Scheduled	Contact will be on vacation so ask for Palmer
1241	1	13	Scheduled	Meet at public buildings dept
503	1	14	Backup	
465	1	15	Backup	
1092	1	16	Backup	
475	1	17	Backup	
879	1	18	Backup	
1152	1	19	Backup	
873	1	20	Backup	

Table 13.33 illustrates some of the data collected in the M&V analysis for the sample projects. For each project in the sample, the following M&V work was carried out:

1. Review the file of each sample project, correct any apparent errors, and recalculate the savings (denoted "Rev kWh"),
2. Visit each sample site, and verify the type and wattage of the lighting equipment actually installed, and recalculate the savings (denoted "Gross Tech kWh"),
3. Verify the quantity of the installed lighting, and recalculate the savings (denoted "Gross Quant kWh"),
4. Assess the hours of use of the lighting in each space, and recalculate the savings (denoted "Gross Hrs kWh"),
5. Interview the program participant, determine whether each measure would have been installed in the absence of the program, and recalculate the savings adjusted for free ridership (denoted "Net kWh FR"),
6. Based on the onsite audit and the participant interview, identify any efficiency measures taken that were not rebated through the program but were the result of the program, and recalculate the savings adjusted for the participant spillover (denoted "Net kWh PSO").

In this study, similar information was developed for the summer and winter demand savings but this information will not be used in the present example.

Table 13.33: Part of the Sample Database

Project	Stratum	Trk kWh	Rev kWh	Grs Tech kWh	Grs Quant kWh	Grs Hrs kWh	Net kWh FR	Net kWh PSO	Case Weight (w)
55	1	2,283	2,282	2,282	2,282	1,173	1,173	1,173	69.4
108	1	157	157	157	157	158	158	158	69.4
113	1	349	349	349	349	350	350	350	69.4
257	1	157	157	157	157	158	158	158	69.4
263	1	6,273	6,275	6,497	6,960	5,435	5,435	5,435	69.4
418	1	6,840	6,845	6,845	6,845	5,086	5,086	5,086	69.4
496	1	1,577	1,577	1,577	1,577	1,704	1,704	1,704	69.4
777	1	978	978	978	978	981	981	981	69.4
1013	1	7,410	7,409	7,409	7,409	15,486	15,486	15,486	69.4
1241	1	2,000	2,001	2,001	2,001	770	770	770	69.4

Periodic review indicated that recruiting was very successful, with few failures due to bad contact information or refusals by the program participants. Table 13.34 summarizes the recruiting effort. Considering all five strata, there were four failures due to inability to locate a contact person for the site and nine failures due to refusal to participate in the study. In total, 63 sites were contacted to recruit the 50 sample sites, so the response rate was 50/63 = 79%. The response rate was at least 71% in each of the five strata.

The recruiting notes were reviewed to understand the reasons for the recruiting failures and whether they might have led to selection bias. There was concern that the sites that were not found might have lower operating hours than the sites that were successfully recruited. If so, the gross realization rate might be overstated.

Table 13.34: Disposition of Sample

Stratum	Not Found	Refused	Scheduled	Total	Response Rate
1	1	2	10	13	77%
2	0	0	10	10	100%
3	1	3	10	14	71%
4	0	2	10	12	83%
5	2	2	10	14	71%
Total	4	9	50	63	79%

Figure 13.17 shows the scatter plot relating the measured gross kWh savings (denoted "Grs Hrs kWh") versus the tracking estimate of the kWh savings, denoted "Trk kWh". Each of the 50 sample projects is a separate point in the scatter plot. In this example, the graph shows that one sample project has much lower gross savings than expected from the tracking information. A closer examination revealed that this was project 840. Project 840 was a maintenance garage for a transportation fleet. The tracking system assumed that the lights operated continuously but the onsite audit showed that a substantial portion of the installed lighting was actually operated for about twelve hours a day for five days a week.

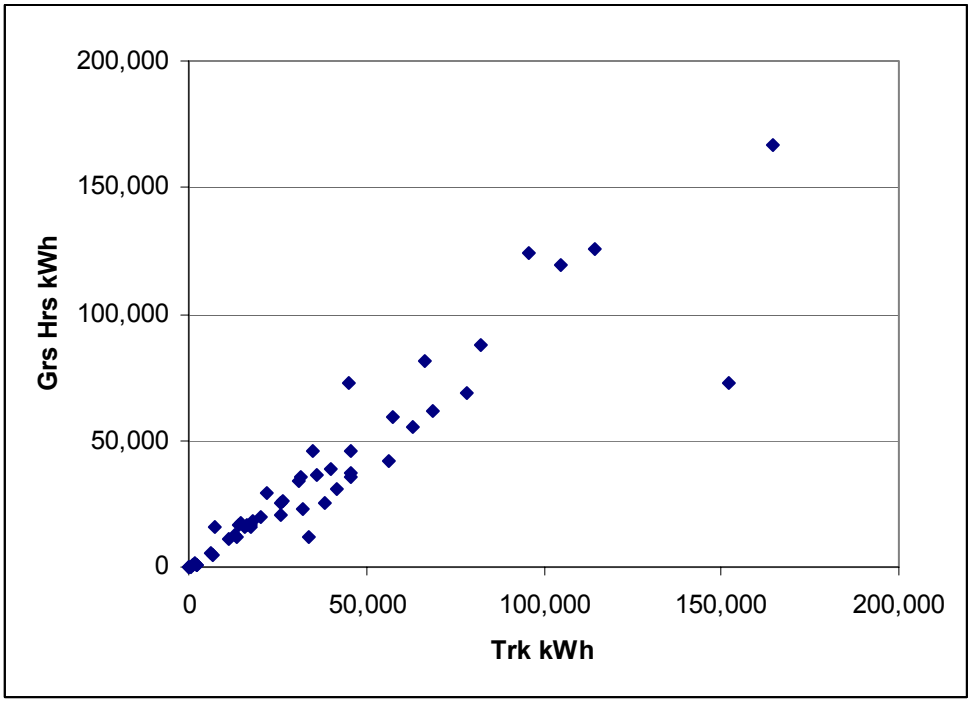

Figure 13.17: Measured Gross Savings versus Tracking Estimates for Sample Projects

Table 13.35 shows the ten most influential sample projects as measured by the weighted squared error (WSE) associated with each project. The most influential case is project 1013. This is a small stratum-one project with tracking savings of 7,410 kWh. The

measures savings of this project were found to be 15,486 kWh. After adjusting the tracking savings for the observed realization rate of 0.981, the error was found to be 8,215 kWh. Although this error is relatively small, each project in stratum one is given a high case weight of 69.4. When this case weight is considered, this project makes a much larger contribution of the overall standard error than any other project.

Table 13.35: Using the WSE to Identify Influential Observations

Project	Stratum	Trk kWh (x)	Grs Hrs kWh (y)	Case Weight (w)	w (w-1) e^2	w (w-1) e2	Sort Influence
1013	1	7,410	15,486	69.4	8,215	320,342,841,430	39
840	5	152,005	72,579	5.4	-76,579	139,338,152,729	36
679	3	33,563	11,704	15.1	-21,230	95,960,582,841	10
204	4	44,967	72,812	9.6	28,687	67,943,968,683	47
550	3	34,944	46,045	15.1	11,755	29,422,331,649	40
677	5	95,833	124,038	5.4	30,000	21,384,148,310	3
1043	4	56,369	42,017	9.6	-13,296	14,595,157,221	6
1036	3	32,037	23,302	15.1	-8,135	14,091,097,698	29
418	1	6,840	5,086	69.4	-1,626	12,546,023,177	17
705	3	22,006	29,209	15.1	7,615	12,347,445,461	25

The second most influential project in the sample was project 840. By contrast with project 1013, this project was the large stratum-five project previously identified in Figure 13.17.

Statistical Analysis

The appropriate statistical analysis will vary depending on the goals and special characteristics of each study. This section will emphasize the methods used to extrapolate the results from the sample to the population and to estimate an error bound for the results. If stratified ratio estimation is used, it is also important to use the sample data to estimate the error ratio to help plan future studies.

Step 1 – Calculate Case Weights

Table 13.36 shows the calculation of the case weights. The case weight is simply the number of projects in the population in each stratum divided by the number of projects in the final sample in the corresponding stratum.

Table 13.37 shows part of the spreadsheet used to assess the gross realization rate. The table shows the ID of each sample project, the tracking savings of the project, the gross savings measured in the M&V analysis, and the case weights determined from the stratum of each sample project.

Table 13.36: Calculating the Case Weights

Stratum	Popu-lation	Sample	Case Weight
1	694	10	69.4
2	253	10	25.3
3	151	10	15.1
4	96	10	9.6
5	54	10	5.4
Total	1,248	50	

Table 13.37: Applying the Case Weights to the Sample Projects

Project	Stratum	Trk kWh (x)	Grs Hrs kWh (y)	Case Weight (w)
9	4	45,659	35,606	9.6
48	5	66,718	81,319	5.4
55	1	2,283	1,173	69.4
74	3	31,791	35,392	15.1
108	1	157	158	69.4
113	1	349	350	69.4
188	2	16,408	16,478	25.3
204	4	44,967	72,812	9.6
209	4	45,544	46,204	9.6
257	1	157	158	69.4
263	1	6,273	5,435	69.4
276	2	17,668	16,271	25.3
283	4	38,085	25,451	9.6
418	1	6,840	5,086	69.4
490	4	57,511	59,482	9.6
496	1	1,577	1,704	69.4
522	5	164,464	166,935	5.4
550	3	34,944	46,045	15.1
557	2	17,920	18,243	25.3
608	3	25,843	25,194	15.1
621	5	82,260	87,470	5.4
628	2	14,883	17,602	25.3
660	2	11,220	11,253	25.3
677	5	95,833	124,038	5.4
679	3	33,563	11,704	15.1

Steps 2 & 3 – Calculate the Ratio and the Standard Error

Table 13.38 shows the results for the gross realization rate. Based on the 50 sample sites, the gross realization rate was found to be 98.1%. In other words, the measured gross savings were estimated to be only 1.9% smaller than the tracking savings across all projects in the population. The standard error was found to be 4.6%. This gives an error bound of ±7.5% at the 90% level of confidence. In other words, the 90% confidence interval for the gross realization rate in the population is from 90.6% to 105.7%. The relative precision is 7.7% at the 90% level of confidence.

Table 13.38: Gross Realization Rate

Realization Rate	0.981
Standard Error	0.046
Error Bound	0.075
Low	0.906
High	1.057
Relative Precision	7.7%

The gross realization rate shown in Table 13.38 was calculated from the sample data illustrated in Table 13.37 using the equation $\hat{B} = \frac{\sum_{i=1}^{n} w_i \; y_i}{\sum_{i=1}^{n} w_i \; x_i}$. This equation was converted into the Excel equation:

=SUMPRODUCT($J$4:$J$53,$I$4:$I$53)/SUMPRODUCT($J$4:$J$53,$H$4:$H$53)

The SUMPRODUCT function in the numerator calculates the sum of the products of the case weights and gross savings. The SUMPRODUCT function in the denominator calculates the sum of the products of the case weights and tracking savings.

Table 13.39 illustrates some of the data used to calculate the standard error of the gross realization rate. For each sample project the error was calculated as $e_i = y_i - \hat{B}\, x_i$ and labeled as e. Then $w_i \left(w_i - 1\right) e_i^2$ was calculated and labeled accordingly. Finally, using another set of Excel functions the standard error was calculated

$$se\left(\hat{B}\right) = \frac{\sqrt{\sum_{i=1}^{n} w_i \left(w_i - 1\right) e_i^2}}{\sum_{i=1}^{n} w_i \; x_i}.$$

Then the error bound at the 90% level of confidence was calculated using the equation $1.645\; se$, and the results were used to calculate the lower and upper boundaries of the 90% confidence interval. Finally, the relative precision was calculated by dividing the error bound by the sample ratio.

Step 4 – Calculate the Error Ratio

The error ratio is the key measure of the association between the y-variable and the x-variable that form the ratio of interest, i.e., between the gross measured savings and the tracking savings of all 1,248 projects in the population. Table 13.40 illustrates part of the spreadsheet used in the calculations. In carrying out this analysis the assumption $\gamma = 0.8$ has been used. This is based on the analysis of many prior evaluation studies.

Table 13.39: Calculating the Standard Error

Project	Stratum	Trk kWh (x)	Grs Hrs kWh (y)	Case Weight (w)	e	w (w-1) e^2
9	2	45,659	35,606	9.6	-9,198	6,984,838,092
48	5	66,718	81,319	5.4	15,850	5,969,398,201
55	4	2,283	1,173	69.4	-1,067	5,407,670,013
74	5	31,791	35,392	15.1	4,197	3,749,693,877
108	5	157	158	69.4	4	62,224
113	1	349	350	69.4	8	299,009
188	1	16,408	16,478	25.3	377	87,442,596
204	2	44,967	72,812	9.6	28,687	67,943,968,683
209	5	45,544	46,204	9.6	1,513	189,063,476
257	3	157	158	69.4	4	62,224
263	1	6,273	5,435	69.4	-720	2,461,033,172
276	2	17,668	16,271	25.3	-1,066	699,122,452
283	3	38,085	25,451	9.6	-11,921	11,732,959,376
418	1	6,840	5,086	69.4	-1,626	12,546,023,177
490	4	57,511	59,482	9.6	3,048	767,084,397
496	2	1,577	1,704	69.4	157	116,430,848
522	4	164,464	166,935	5.4	5,552	732,297,343
550	2	34,944	46,045	15.1	11,755	29,422,331,649
557	4	17,920	18,243	25.3	658	266,534,882
608	2	25,843	25,194	15.1	-165	5,776,173
621	3	82,260	87,470	5.4	6,750	1,082,677,262
628	1	14,883	17,602	25.3	2,998	5,525,726,540
660	4	11,220	11,253	25.3	243	36,334,141
677	4	95,833	124,038	5.4	30,000	21,384,148,310
679	5	33,563	11,704	15.1	-21,230	95,960,582,841

Using this assumed parameter x_i^γ was calculated. Then the ratio e_i^2 / x_i^γ was calculated for each sample project. Then three

SUMPRODUCT functions were used to calculate $\hat{er} = \dfrac{\sqrt{\left(\sum_{i=1}^{n} w_i \; e_i^2 / x_i^\gamma\right)\left(\sum_{i=1}^{n} w_i \; x_i^\gamma\right)}}{\sum_{i=1}^{n} w_i \; y_i}$.

The result is an error ratio of 0.3. This is smaller than the 0.4 assumed in planning this study.

Table 13.40: Estimating the Error Ratio for the Gross Realization Rate

Project	Stratum	Trk kWh (x)	Grs Hrs kWh (y)	Case Weight (w)	e	x^{γ}	e^2 / x^{γ}
9	2	45,659	35,606	9.6	-9,198	5,341	15,840
48	5	66,718	81,319	5.4	15,850	7,234	34,729
55	4	2,283	1,173	69.4	-1,067	486	2,343
74	5	31,791	35,392	15.1	4,197	3,998	4,405
108	5	157	158	69.4	4	57	0
113	1	349	350	69.4	8	108	1
188	1	16,408	16,478	25.3	377	2,355	60
204	2	44,967	72,812	9.6	28,687	5,276	155,979
209	5	45,544	46,204	9.6	1,513	5,330	430
257	3	157	158	69.4	4	57	0
263	1	6,273	5,435	69.4	-720	1,091	475
276	2	17,668	16,271	25.3	-1,066	2,499	455
283	3	38,085	25,451	9.6	-11,921	4,620	30,763
418	1	6,840	5,086	69.4	-1,626	1,170	2,260
490	4	57,511	59,482	9.6	3,048	6,424	1,446
496	2	1,577	1,704	69.4	157	362	68
522	4	164,464	166,935	5.4	5,552	14,889	2,070
550	2	34,944	46,045	15.1	11,755	4,312	32,047
557	4	17,920	18,243	25.3	658	2,527	172
608	2	25,843	25,194	15.1	-165	3,387	8
621	3	82,260	87,470	5.4	6,750	8,554	5,327
628	1	14,883	17,602	25.3	2,998	2,178	4,126
660	4	11,220	11,253	25.3	243	1,738	34
677	4	95,833	124,038	5.4	30,000	9,665	93,118
679	5	33,563	11,704	15.1	-21,230	4,175	107,946

Step 5 – Estimate the Total Gross Savings of the Program

The results for the gross realization rate were used to estimate the total gross savings of the program as illustrated in Table 13.41. The first entry of the table shows the total tracking savings of all 1,248 projects in the population database. The remaining results were obtained by multiplying the total tracking savings by the results for the realization rate shown in Table 13.38. Then the gross program savings were estimated to be 19,742,490 kWh. At the 90% level of confidence, the total gross savings are estimated to be between 18,225,104 kWh and 21,259,876 kWh.

Table 13.41: Gross kWh Savings of Program

Total Tracking Savings	20,119,315	kWh
Estimated Gross Savings	19,742,490	kWh
Standard Error	922,423	kWh
Error Bound	1,517,386	kWh
Low	18,225,104	kWh
High	21,259,876	kWh
Relative Precision	7.7%	

Step 6 – Estimate a Ratio for a Subset

Suppose that the gross savings of the projects implemented by vendor 7 are desired. Table 13.42 illustrates some of the sample data used in the analysis. Note that the indicator variable is the key to the analysis. The indicator is equal to one for vendor 7 and zero otherwise. The tracking and gross kWh are equal to their usual values times the indicator. The rest of the calculation is unchanged.

Table 13.42: Sample Data for Analysis of Vendor 7

Project	Vendor	Stratum	Indicator	Trk kWh (x)	Grs Hrs kWh (y)	Case Weight (w)	e	w (w-1) e^2
9	2	4	0	0	0	10	0	0
48	1	5	0	0	0	5	0	0
55	7	1	1	2,283	1,173	69	-1,430	9,707,856,646
74	7	3	1	31,791	35,392	15	-854	155,446,066
108	7	1	1	157	158	69	-21	2,158,557
113	7	1	1	349	350	69	-48	10,716,611
188	7	2	1	16,408	16,478	25	-2,230	3,056,831,758
204	7	4	1	44,967	72,812	10	21,543	38,315,538,143
209	7	4	1	45,544	46,204	10	-5,723	2,704,004,538
257	7	1	1	157	158	69	-21	2,158,557
263	7	1	1	6,273	5,435	69	-1,717	13,989,736,194
276	2	2	0	0	0	25	0	0
283	4	4	0	0	0	10	0	0
418	4	1	0	0	0	69	0	0
490	4	4	0	0	0	10	0	0
496	7	1	1	1,577	1,704	69	-94	41,897,470
522	7	5	1	164,464	166,935	5	-20,579	10,062,319,626
550	4	3	0	0	0	15	0	0
557	2	2	0	0	0	25	0	0
608	4	3	0	0	0	15	0	0
621	4	5	0	0	0	5	0	0
628	4	2	0	0	0	25	0	0
660	4	2	0	0	0	25	0	0
677	4	5	0	0	0	5	0	0
679	4	3	0	0	0	15	0	0

Table 13.43 shows the resulting gross realization rate and error bound.
Table 13.44 shows the estimated total gross savings for vendor 7. Finally, by combining the estimated gross savings shown in Table 13.41 and
Table 13.44, it was estimated that vendor 7 was responsible for 29.4% of the gross savings of all projects in the population.

Table 13.43: Realization Rate for Vendor 7

Vendor	7
Realization Rate	1.140
Standard Error	0.097
Error Bound	0.160
Low	0.980
High	1.300
Relative Precision	14.1%

Table 13.44: Total Gross Savings for Vendor 7

Total Tracking Savings	5,811,769	kWh
Estimated Gross Savings	6,626,315	kWh
Standard Error	566,227	kWh
Error Bound	931,443	kWh
Low	5,694,871	kWh
High	7,557,758	kWh
Relative Precision	14.1%	

Analysis of Potential Selection Bias

Recall that in Table 13.34 the response rate was reported as 79%. This present section will summarize a special analysis that was carried out to assess the potential impact of selection bias on the gross realization rate of this program. This analysis was carried out in the following three steps:

1. **Recruiting Case Weights** – create case weights that reflect all 63 projects that were included in the recruiting effort.
2. **Tracking Savings by Recruiting Outcome** – estimate the tracking kWh savings in the program in each of the recruiting categories: not found, refused, and scheduled.
3. **Realization Rate by Recruiting Category** – assess the realization rate in each of the three categories and calculate the overall realization rate of the program.

Table 13.45 shows the development of the recruiting case weights. In these calculations, the sample was considered to be all 63 sites that were contacted during the recruiting work. In each stratum, the case weight was calculated as the ratio between the number of projects in the population and the number of projects in the sample.

Table 13.45: Recruiting Case Weights

Stratum	Population	Sample	Case Weight
1	694	13	53.4
2	253	10	25.3
3	151	14	10.8
4	96	12	8.0
5	54	14	3.9
Total	1,248	63	

Now imagine that the recruiting work could be applied to all of the projects in the population and that each project in the population is put into one of the following three categories: those that would not be found, those that would refuse to participate in the study, and those that would agree to be scheduled. Then the case weights and the tracking kWh savings of each of the 63 projects in the recruiting sample were used to calculate the stratified ratio estimate of the kWh savings of all projects in the population in each of these categories. Table 13.46 shows the results.

Table 13.46: Tracking Savings by Recruiting Category

Stratum	Scheduled	Not Found	Refused	Total
1	1,533,215	288,599	148,102	1,969,916
2	3,950,854	0	0	3,950,854
3	3,139,203	248,143	1,000,057	4,387,403
4	3,700,268	0	775,887	4,476,155
5	3,916,728	557,261	860,998	5,334,986
Total	16,240,268	1,094,003	2,785,044	20,119,315

Finally, the gross realization rate in each category was assessed as shown in Table 13.47. In the scheduled category, the realization rate from Table 13.38 was used, i.e., 0.981. In the Not Found category, it was assumed that the trouble finding the program participant might be because they went out of business or were very seldom in their place of business. This suggests that their operating hours might be shorter than recorded in the tracking system, thereby leading to a lower realization rate than normal. A logical realization rate might be 0.5 or higher for this category. To be conservative 0.5 was assumed here.

In the Refused category, it was assumed that the realization rate was more likely to be similar to that of the Scheduled category. Perhaps program participants refused to cooperate with the onsite audit because they were unhappy with their savings (low realization rate), or perhaps because they were very busy (high operating hours and therefore high realization rate). But it seemed most likely that their refusal was unrelated to their savings. Therefore, the realization rate in the Refused category was taken to be equal to the realization rate in the Scheduled category, i.e., 0.981.

In each of the three categories, the gross kWh savings was estimated by multiplying the estimated tracking savings by the estimated realization rate. Then the total gross kWh savings was calculated across all three categories, 18,606,702, and the overall realization

rate was calculated as 0.925. By calculating the difference between this result and the realization rate previously calculated, shown in Table 13.38, it was concluded that the gross realization rate found from the sample of scheduled projects was likely to have a selection bias of about 0.056 or less.

It is important to recognize that the gross realization rates assumed for the Not Found and Refused categories were extremely subjective since there was very little information for these projects. In particular, there was no objective basis for assuming that the realization rate in the refused category was the same as in the Scheduled category. Sensitivity analysis was used to try different assumed realization rates for the Not Found and Refused categories. At worst, the realization rates could be assumed to be zero in these two categories. In this case the overall program realization rate was found to be only 0.779. In this worst possible scenario, the selection bias would have been 0.202, substantially greater than the error bound reported in Table 13.38.

Table 13.47: Realization Rate by Recruiting Category

	Scheduled	Not Found	Refused	Total
kWh Tracking Savings	16,240,268	1,094,003	2,785,044	20,119,315
Gross Realization Rate	0.981	0.500	0.981	0.925
Est. Gross Savings	15,668,643	638,708	2,299,350	18,606,702

Similar considerations were applied to the net-to-gross analysis, reported below. In this case it was felt that the net-to-gross ratio was unlikely to be substantially different in the three recruiting categories: Scheduled, Not Found and Refused.

Other Results

The net kWh savings, obtained after adjusting the adjusting for free ridership, was also estimated. Table 13.48 summarizes the results. Based on the site-specific net-to-gross analysis of the 50 sample projects, the program net realization rate was estimated to be 92.4% ± 9.4% at the 90% level of confidence. The achieved relative precision for this result was 10.1% at the 90% level of confidence. The corresponding error ratio was 0.41, very close to the 0.4 assumed in planning the study.

Table 13.48: Net Realization Rate

Realization Rate	0.924
Standard Error	0.057
Error Bound	0.094
Low	0.830
High	1.017
Relative Precision	10.1%
Estimated Error Ratio	0.409

The net realization rate for the sample was used to estimate the net savings of the program as illustrated in Table 13.49. From the net savings in Table 13.49 and the gross

savings in Table 13.41, the net-to-gross ratio was estimated to be 18,584,090 / 19,742,490 kWh = 94%.

Table 13.49: Net Program Savings

Total Tracking Savings	20,119,315	kWh
Estimated Net Savings	18,584,090	kWh
Standard Error	1,143,903	kWh
Error Bound	1,881,720	kWh
Low	16,702,370	kWh
High	20,465,810	kWh

Recall that the gross and net savings were measured in a series of steps – file review, technical adjustment for type of fixtures installed, adjustment for the quantity of fixtures installed, the reported hours of use, the reported free ridership, and the reported participant spillover. The analysis of these data is shown in Table 13.50. The table shows each ratio as well as the cumulative product of the ratios up through each factor. In this study, the principle adjustments came from the hours of use (ratio 4) and the free ridership analysis (ratio 5). The participant spillover (ratio 6) partially offset the free ridership. Note that the gross realization rate is the cumulative ratio 4 and the net realization rate is the cumulative ratio 5.

Table 13.50: Decomposition of the Realization Rate into Underlying Factors

	x-variable		*y-variable*	*Ratio*	*Cum Ratio*
1	Trk kWh	versus	Rev kWh	1.000	1.000
2	Rev kWh	versus	Grs Tech kWh	0.998	0.998
3	Grs Tech kWh	versus	Grs Quant kWh	1.003	1.001
4	Grs Quant kWh	versus	Grs Hrs kWh	0.980	0.981
5	Grs Hrs kWh	versus	Net kWh FR	0.941	0.924
6	Net kWh FR	versus	Net kWh PSO	1.018	0.940

These results were also reported in terms of program savings after accounting for each factor are shown in Table 13.51.

Table 13.51: Program Savings after Considering Each Factor

Factor	*kWh Savings*
Trk kWh	20,119,315
Rev kWh	20,120,853
Grs Tech kWh	20,076,578
Grs Quant kWh	20,138,764
Grs Hrs kWh	19,742,490
Net kWh FR	18,584,090
Net kWh PSO	18,909,596

Chapter 14: Evaluation and Cost-Effectiveness

Preface

Most of the chapters in this new Framework provide guidance on evaluation methodologies for conducting program-specific evaluations. This chapter, however, is not providing advice on how to conduct cost-effectiveness analysis. The California Public Utility Commission (CPUC) has directed that the *Standard Practice Manual (SPM): Economic Analysis of Demand-Side Management Programs*, referred to as the *Standard Practice Manual)*[351], provide the "how-to" for California cost-effectiveness analysis. In addition, the CPUC has required that a separate, more extensive study be conducted to reexamine the nature and scope of cost-effectiveness testing and the associated issues within the new market structures for California energy resources. For these reasons, this chapter is targeted to three audiences in the area where issues overlap between program evaluations, cost-effectiveness analysis, and their uses and interpretations. First, it will help evaluators to see how their evaluation designs will be used in cost-effectiveness analysis. Second, it will show program staff and administrators who calculate or use cost-effectiveness analysis how evaluation and cost-effectiveness work together. And third, it will help policy makers understand some of the key issues involved in using evaluation results to estimate cost-effectiveness, since these tests are often used to inform a policy decision on whether to continue to invest in a program.

In summary, evaluation results are vital inputs to the cost-benefit tests. Evaluation and cost-effectiveness intersect and interact in several areas, including the following:

- Evaluators must understand the type of data required in order to be able to apply the cost-benefit tests or to structure evaluation efforts that support these requirements;
- The design and delivery of programs can be improved if all parties that are involved in program design, administration, and implementation understand the factors that increase program cost-effectiveness and use these factors to help refine programs; and
- Program implementers should understand temporal and geographical differences in the value of load reductions, so that they can understand why a program that is cost-effective in one time period or area may not be cost-effective at other times or areas.

This chapter highlights issues at the intersection of evaluation and cost-effectiveness analysis and suggests that evaluations might be improved by recognizing these issues. This chapter is written with the following stakeholders in mind: evaluators, policy-makers, program designers, program planners, administrators, and those involved in conducting avoided cost studies and cost-effectiveness analysis.

[351] (* California State Governor's Office 2001).

Skills Needed

A broad array of professionals is involved in estimating and using estimates of program cost-effectiveness, requiring a broad set of skills across a number of stakeholders. While the issues of cost-effectiveness appear on the surface to be rather simple, comparing the cost of the program to program achievements, it is significantly more complicated than this simple comparison of two data points. Professionals who practice in the field of estimating program cost-effectiveness should have at least a sufficient understanding of program theory, economic theory, consumer behavior, energy impact cause and effect relationships, familiarity with energy technologies and related behaviors, an understanding of weather and weather patterns and their influence on energy use and technology efficiencies, and, of course, a strong understanding of the mathematical relationships, theories and tests associated with conducting cost-effectiveness estimations. (Professionals who develop the avoided cost-estimates used to value energy savings in cost-effectiveness analysis need a different set of skills.)

Cost-Benefit Analysis for Evaluators

For analyzing cost-effectiveness, the California Public Utilities Commission (CPUC) uses the standard cost-effectiveness methodologies articulated in the *California Standard Practices Manual (SPM): Economic Analysis of Demand-Side Management Programs*.[352] Two cost-effectiveness tests identified in the SPM are particularly important to the Commission in evaluating energy efficiency programs on an ongoing basis. The first is the Total Resource Cost (TRC) test – Societal Version. This test, as defined in the SPM, is intended to measure the overall cost-effectiveness of energy efficiency programs from a societal perspective, taking into account benefits and costs from more than just an individual perspective. The Commission primarily relies upon the results of this test in assessing program cost-effectiveness.

In addition to the TRC test, the Commission relies on the Participant Test (also identified in the SPM) to evaluate programs that are aimed at inducing individual customers to make energy efficiency decisions. The Participant Test measures the cost-effectiveness of a program from the perspective of energy consumers participating in the program. Proposals for programs designed to provide financial incentives directly to customers are required to include the results of the Participant Test as well as the TRC.[353]

All resource acquisition programs should have a TRC value greater than 1 to be considered cost-effective. In other words, the present value of the benefits counted under the test must exceed the present value of the costs, so that the benefit-cost ratio is greater than 1.0, or the net benefits (benefits minus costs) are positive.

---

[352] *Energy Efficiency Policy Manual Version* 2. (* CPUC 2003), page 15.

[353] There are other cost-effectiveness tests that are not mentioned in the *Energy Efficiency Policy Manual*. For example, the Utility Costs test focuses on the costs to the utility (and does not include customer costs) and may be used for allocating portfolio resources.

The benefits used in the TRC test are the present values of:[354]

- The avoided costs of electric and gas energy and capacity, which may be avoided by the participating customer or the utility, depending on future institutional arrangements for electric and gas supply;
- The costs of transmission and distribution investments expected to be avoided;
- The value of the environmental externalities related to burning fuel at the power plant (for electricity) or the end use (for natural gas); and the
- Avoided participant capital and operating costs, including reduced water bills and reduced maintenance costs (e.g., less frequent lamp replacements).

The costs used in the TRC are the present values of:

- Direct program implementation and overhead costs, excluding incentives paid to participants;
- The costs of energy, capacity, transmission and distribution, and externalities associated with any increases in energy usage caused by the program;[355] and the
- Increases in participant capital and operating costs.

Factors Influencing Program Cost-Effectiveness

Important points derived from the above benefit and cost streams that evaluators should understand include the following:

- Benefits are measured by avoided costs, not by energy or peak demand reductions. The true avoided costs of the load reductions will vary with time and location, but these variations may not be fully reflected in the avoided cost stream prescribed for valuing them.
- The TRC test is computed from the present values of benefits and costs. Ongoing costs and benefits are just as important as expenditures in the program year (although future cash flows are discounted).
- Longer-lived measures will have greater benefits than shorter-lived measures, all else being equal.
- From the perspective of the TRC test, a dollar of expenditure by a participant is just as important as a dollar of expenditure by the program implementer.

---

354 *Standard Practice Manual (SPM): Economic Analysis of Demand-Side Management Programs*. (* California State Governor's Office 2001).

355 Energy use may increase for any of several reasons. Fuel substitution increases the use of one fuel to reduce the use of another fuel. Conservation programs that reduce waste heat in occupied building space will also reduce air conditioning energy use, but will tend to increase space-heating energy use. Load management decreases use at some times, but may increase energy use at other times.

- At the implementation level, the objective is to achieve the maximum improvement in net benefits per dollar cost, rather than to maximize the ratio of benefits to costs. For example, it is better to spend $50 to save $100 (net savings is $50 and B/C is 2.0), than to spend $10 to save $30 (net savings is $20 and B/C is 3.0). Optimization of program design should be driven by the net benefit per application, not the benefit-cost ratio, to avoid cream skimming and creation of lost opportunities.[356]

Each of the above points has implications for evaluation and program design. As this is an evaluation framework document, some examples of how this could affect evaluation research design are mentioned:

- Because the value of achieved load reduction varies by time and location, at least for larger programs or groups of programs, it is important to calculate and present impact or measurement and verification (M&V) evaluation results by date, time and geographical location.

- The ongoing costs and benefits included in cost-effectiveness analysis create the need for evaluation to be concerned with the ongoing value of the costs and benefits. Hence, the need for persistence studies for measure retention, technical degradation, and persistence of savings. These have been an important part of the evaluation agenda through the prior measurement and evaluation (M&E) Protocols.

- In addition to focusing on technologies, it may be important to evaluate other streams of costs or benefits, such as the program's effect on ongoing maintenance costs and periodic training costs as are needed or reduced due to changes made through the program (in equipment or operation). However, in some cases these might be considered non-energy benefits and costs, so see Chapter 11 on Non-Energy Effects Evaluation regarding conducting evaluations to measure these types of benefits and costs. At the same time, if required activities are not undertaken and the consequences mean a shorter measure life or reduced persistence, this assessment needs to be included in the evaluation.

- In calculating the TRC, participant expenditures are just as important as program dollars. Hence, programs need to be designed not only to be cost conscious for the program but also to be cost conscious for participants (including such costs as training, staffing, maintenance, etc.). This is an area that might best be explored through process evaluation.

[356] If budgetary limitations preclude pursuit of all cost-effective efficiency opportunities, the ratio of net benefits to program expenditure is useful in allocating capital among independent efficiency investments (such as towards different types of customers, or on different end-uses or equipment), where deferred savings can be realized later. The B-C ratio is not appropriate among competing investments (on the same end use or equipment), where foregone investment opportunities are forever lost.

- Certain types of costs should be separated from program delivery costs to avoid distorting evaluation findings, as discussed in the next section.

Treatment of Special Categories of Costs

For many programs offered in California over the past few years, some incurred costs are associated with program delivery, but are not necessary for the delivery of program services itself. These costs should be separated from program delivery costs to avoid distorting evaluation findings by including costs that are not costs that would typically be a part of the program. These include, but are not limited to, participant-supported upgrades, program start-up costs, or evaluation costs.

Participant Supported Upgrades

In some program designs, participants may be able to specify equipment for the efficiency improvement that costs more, but offers other benefits that are not associated with improved energy efficiency: a low-flow showerhead with a massage setting, an efficient refrigerator with deluxe finish or customer preferred options, or lighting fixtures that reduce glare and increase productivity. In these situations, it is important to distinguish between the costs associated with the efficiency improvement at the existing level of other amenities (e.g., a white refrigerator replacing a white refrigerator) and the additional costs related to the upgrade (the incremental cost of the chrome finish). In these cases, the cost of the energy efficient upgrade is less than the total cost of the energy efficient technology as installed. The stainless steel ENERGY STAR® refrigerator will cost considerably more than the same technology without the stainless steel surface. If the cost of the stainless steel unit is considered as part of the cost of the energy efficient upgrade, then the cost-effectiveness assessment includes costs that are not related to the energy efficient improvement. When possible, these costs should be excluded, unless the program theory indicates that the stainless steel exterior is a program component that adds to the energy efficient improvement and is, therefore, a part of the program offering.

Participants may coordinate substantial remodeling with efficiency upgrades to lighting, HVAC and other systems. The costs of other activities, even if triggered by the efficiency measures and implemented through the same contractor, are not part of the measure cost. Where possible, the change in energy use should only be for those measures that are being replaced or retrofitted that directly relate to the program theory of how the technology is providing savings. On the other hand, if one were to include the energy savings from these other measures, then the costs for these additional measures should be calculated as well.

Start-Up and Future Improvement Costs

In evaluating and interpreting the cost-effectiveness of programs, it is important to distinguish between costs incurred in the past (particularly non-recurring “start-up costs”)

and costs that would be incurred for a similar program in the future. For example, non-repeating start-up costs such as market research, protocol development, regulatory costs associated with program approval and contracting, are different than normal program operating costs. Similarly, the incremental costs of gathering data for future improvements and future programs (e.g., market analysis), beyond those required for the operation and reporting of the existing program, are also costs that should not be considered normal program operational costs.

Charging these costs to the current year of program operation might make the program appear to be uneconomic, even if the program was producing benefits well in excess of normal direct program costs. All costs must be reported and included in historical summaries, but costs unrelated to continued operation should be excluded from the analyses of the cost under *steady-state condition*s or to determine the cost of *continuing* the program.

This concept is already included in key California cost-effectiveness policy documents. The CPUC recognizes that start-up costs are non-repeating, and should not bias analysis of continuing programs:

> *"The TRC should be calculated by treating programs as multi-year (rather than single-year) activities so that programs explicitly designed as integrated, multi-year strategies, which may have modest benefits (and/or high start-up costs) in early program years, could be evaluated considering the expected larger benefits (and/or lower costs) in later program years."*[357]

Evaluation Costs

As a general philosophy, evaluation costs can be considered as a standard cost of running an efficiency program with Public Goods Charge funding. At the same time, the level of evaluation costs could swing significantly from year to year depending upon the life cycle of the program, market circumstances, and purpose of the evaluation. This can create situations where use and interpretation of including these costs in a cost-effectiveness test could be problematic.

An extreme example could be found in a pilot program testing a new program concept where the evaluation costs could be as high, or higher, than the rest of the program costs. A continued program would be expected to have far lower evaluation costs than what might occur on a periodic schedule. Including the pilot's evaluation costs in the cost-benefit analysis that is used to decide whether to continue this program to full operation would make no sense and doom the pilot to failure.

In Chapter 4 on Evaluation Overview and Issues the concept of formative evaluations versus summative evaluations is discussed. To briefly recap: in summative evaluations, the focus is on documenting and measuring the effects of a program; in formative evaluations, the focus is on understanding why those effects occurred and

[357] *Energy Efficiency Policy Manual Version 2*. (* CPUC 2003).

identifying ways to improve a program. As a policy and evaluation tool, cost-effectiveness analysis could also conceptually be divided into formative and summative. If a test of cost-effectiveness is being used to inform a policy decision of whether to continue to invest in a program or program concept, it is best if the cost-effectiveness analysis take a formative perspective.

Given these concerns, good practice seems to indicate that program cost-effectiveness should be analyzed and reported both ways. Cost-effectiveness analysis needs to be conducted with all relevant program costs (to include apportioning administrative costs from the implementing organization as is relevant to the administrative needs of the program, regulatory costs, and evaluation costs), as provided by the *Standard Practice Manual*. Cost-effectiveness analysis could be reported without regulatory and evaluation costs. In the reporting of these cost-effectiveness tests, including a qualitative discussion about recurring future costs versus one-time or unusual costs would also add to the quality and usefulness of the reporting. This provides administrators and policy makers the best information for both summative and formative purposes.

Excluding evaluation or regulatory costs from program-specific cost-effectiveness tests does not in any way mean that these costs should be excluded in an assessment of the cost-effectiveness of the entire Public Goods Charge process for designing and delivering energy resources to the State of California. Rather, this concept is included here to indicate that evaluation costs that are above and beyond the costs that are needed to acquire the resources might be included in a cost-effectiveness assessment of the program portfolio, but not necessarily included in the assessment of individual programs.

Understanding Temporal and Geographical Variations in Avoided Costs

The value of a kilowatt-hour of electricity, or a therm of natural gas, varies over time and space.

- Avoided electric energy costs vary with load level. When demand is high, market prices for energy also tend to be high. As a result, market energy prices are typically higher during weekday business hours ("peak hours") than on weekends, and lower late at night ("off-peak hours"). Energy prices are higher in months with high cooling and heating loads and lower in milder months when the demand for energy is lower.
- Higher electric loads result in higher energy losses in the transmission and distribution systems, compounding the variation in energy prices at the generation level.
- Natural gas prices tend to vary from day to day (rather than hour to hour, as for electricity), with higher prices on the cooler days that require more heating, and hence more gas.

- Energy costs vary from one part of the state to another, depending on transmission constraints.[358]
- The costs of transmission and distribution (T&D) capacity to deliver energy to consumers are driven by peak loads. Those peaks may occur at different times (and even different seasons) for various T&D components.[359]

One important implication of the variation of avoided costs over time is that various load shapes of energy savings will have different benefits per kWh (or per therm). Reduction in electric loads for industrial motors that run all the time will save relatively little T&D capacity per kWh, and a large fraction of the energy will be saved at relatively low-priced hours. Reductions in commercial lighting loads will tend to be concentrated during business hours, with higher average energy prices, and more capacity savings per kWh. Reductions in commercial cooling loads in a summer peaking area will usually save energy primarily in the highest priced hours, when the T&D system is most heavily loaded, which creates high avoided costs per kWh and per kW.[360]

These variations cannot be entirely captured by dividing avoided costs into time periods by the calendar or clock. Within any particular period, such as summer weekday afternoons, loads, energy costs and losses will be higher on some days than others, depending primarily on weather conditions. Load reductions that tend to follow load (such as air conditioner efficiency) will have higher avoided costs than reductions that follow the clock. In addition, capacity availability can also influence cost. Generation facilities may be closed or in reduced capacity mode because of a range of conditions. Generation facilities may be shut down for scheduled or unscheduled maintenance causing shortages that increase cost. It is possible that the value of energy saved via a resource acquisition program can be significantly more valuable if it is saved during times of limited supply. These conditions mean that the value of energy saved from an energy program depends on the price of the energy at the time it is being saved.

Reductions affecting the peak of the average system load shape will tend to have higher avoided costs than reductions that are reducing energy use for all hours, since energy prices, line losses, and stress on the T&D system tend to be high when loads are high. Evaluators must be able to not just report how much energy is being saved, but be able to identify the time period over which the energy is saved and the amount of energy that is being saved within relevant price periods in order to identify the avoided costs associated with a program or group of programs.

---

358 Whether those differences are large and persistent enough to be reflected in avoided costs for program valuation remains to be seen.

359 While avoided electric T&D costs vary widely from one substation or feeder to the next, developing and using avoided T&D costs on that level of detail (potentially thousands of different values and peak periods across the state) is not likely to be feasible for most purposes. Whether aggregate regional estimates of avoided T&D costs would be more accurate than a statewide average, or vary enough to be worth reflecting in avoided costs for program valuation, remains to be seen.

360 Even for a single end use, different measures may have different load shapes and hence different avoided costs. For example, daylighting sensors on commercial lighting are likely to preferentially reduce lighting load at high-priced on-peak summer midday hours, while occupancy sensors may primarily reduce loads at night.

Recent changes in energy demand, changes in the regulatory environment, and changes within the energy supply and distribution environment have had, and are expected to continue to have, significant impacts on the price of energy in California. As a result, there is a significant amount of uncertainty associated with projecting the avoided cost associated with resources generated from energy efficiency programs. In the future, it may be more appropriate to use avoided cost projections that are reported with bounded confidence intervals and relative precision levels, so that the level of uncertainty associated with any given point estimate can be used to identify the expected and possible cost-effectiveness range associated with energy programs. (The methods described in Chapter 12, Uncertainty, can be used for chaining of results to estimate an overall relative precision.)

At the present time, the avoided cost calculation methods used in California do not incorporate an *economic condition hedge factor*. This would allow the avoided cost equation to include an estimate of the value of energy efficiency resources as a hedge against the economic effects of blackouts and curtailments on the California economy. When Public Goods Charge funds are used to support energy programs, there may be a need to incorporate a factor in the avoided cost estimate that incorporates the value associated with the effects of limited supplies on the California economy and the benefits resource acquisition programs can provide toward reducing the economic impacts of these conditions.

Providing Useful Avoided-Cost Estimation to Evaluators

The derivation of avoided costs is beyond the scope of this report. This section considers the form in which the estimated avoided costs should be provided to evaluators, to promote efficient and consistent use of those estimates. A number of features would contribute to making avoided costs easier for evaluators to apply, and to making the resulting program cost-effectiveness analyses more accurate. These comments do not presuppose the methodology used to produce the avoided costs.

First, the avoided costs provided to evaluators should be as simple and as complete as possible, so that evaluators need not be concerned with more separate cost components than necessary. For example, the avoided costs should include to the extent possible all externalities and (for electricity) load-specific losses to the end use.[361]

Second, avoided costs should be available for common technology-associated load shapes (e.g., commercial chilling, commercial lighting, and residential air conditioning) as well as for constant loads within each rating period. These avoided costs typically vary between geographical regions due to climate and due to differing capacity, demand, transmission and distribution constraints.

[361] If the avoided costs include avoided losses only to the customers' meter, the program evaluation will need to estimate losses on the customer's side of the meter, especially for customers metered at primary voltage.

Third, avoided costs should be available in various levels of detail, reflecting the range of program complexity and sophistication of potential administrators. At one extreme, for complex custom measures for large customers (such as large commercial new construction programs, or comprehensive industrial process retrofits), which are likely to be modeled and/or metered in great detail, avoided costs should be available as hourly demand and energy values, so that load effects by time and weather conditions can be matched with avoided costs. For less complicated applications, the avoided energy costs should be listed by rating period, weighted by the hourly loads of various end-uses and load types, while avoided demand costs should be listed, with the times and weather conditions driving the peak demands clearly identified.[362] These avoided costs can be provided on an annual basis, with the discount rate specified,[363] or as a present value for various lifetimes.

Fourth, avoided peak demand energy savings are likely to be based on avoided purchased energy during this time period. These savings should actually reflect the system benefits of reducing peak demand (e.g., less purchase of capacity, less spinning reserve, etc.).

Finally, for small programs using well-characterized measures, it would be helpful to have all the avoided costs for each common measure (e.g., residential lighting, small-commercial refrigeration) reduced to a single value of present value per annual MWh (or MMBtu), which includes load-weighted energy, demand, externalities and loss costs.

[362] For example, the peak conditions for a geographical area might be listed as "September weekday 2 PM, 95° F."

[363] In evaluating energy efficiency program proposals, the Commission requires everyone to use a pre-established discount rate of 8.15% (as noted in the Energy Efficiency Policy Manual) (* CPUC 2003).

Chapter 15: Overarching Evaluation Studies

In addition to the need for evaluating specific programs, or groups of related programs, there are significant benefits from conducting "overarching studies" to collect and analyze information at a statewide or regional level that can be used for program and portfolio evaluation planning as well as policy decision making. While some of the overarching information can be collected in evaluations of individual programs, large-scale studies may be needed that encompass a broader array of programs. These evaluation approaches are not addressed in this Framework. However, these overarching studies are important for understanding the market in which energy programs are placed and the types of measures that could be considered for inclusion in programs and the level of savings that can be achieved from programs and program-induced actions. The types of overarching studies include, but are not limited to:

1. Measure saturation studies,
2. Energy-savings potential studies (technical, economic, achievable-market),
3. Portfolio analyses (including "best practices" and "lessons learned" studies),
4. Market and market operations analysis (beyond program level),
5. Studies that update key parameters that influence multiple programs (e.g., measure lifetime, avoided costs), and
6. Development of improved methodologies for evaluating programs.

The CPUC, CADMAC and CALMAC[364] have funded these and other studies in the past, and the CPUC has been requiring these studies in 2003 and 2004 to help understand and plan for California's energy supply options today and for the future. The following paragraphs provide a summary review of these types of studies and indicate their importance in planning the needs of California's energy future and the amount of energy available via energy efficiency programs.

Measure Saturation Studies

It is important to establish a credible platform from which to determine which types of energy efficiency programs should be offered, which sectors should be targeted and what technologies should be included in California's programs. Measure saturation studies help to achieve this objective by collecting and analyzing data on the types of measures that are, for example, installed in residential and commercial buildings and identifying their current level of energy efficiency. Ideally, these types of studies should be conducted annually on all key energy efficiency measures. However, due to budgetary concerns and in some cases the pace of normal market change, it is more realistic to expect these studies to be conducted periodically over time. Or this effort can focus on

[364] CPUC – California Public Utilities Commission; CADMAC – California Demand-Side Management Measurement Advisory Coucil; CALMAC – California Measurement Advisory Council.

conducting annual or bi-annual studies in which the studies focus on just those technologies in which significant change is expected because of recent market changes or program accomplishments. Alternatively, saturation studies can be rotated through a series of key energy-intensive sectors, or energy-intensive end uses. Examples of past saturation studies in California include the following by: ADM Associates and TecMRKT Works,[365] Quantum Consulting,[366] RER,[367] RLW Analytics,[368] and SDG&E.[369]

The primary purpose of the measure saturation study is to baseline the status of the market in terms of the current penetration of energy consuming technologies or behaviors. This is done so analysts and policy makers can document changes in the market and identify those parts of the market in which energy savings potentials can be achieved through future programs and technology offerings.

Energy-Savings Potential Studies

In addition to having a good saturation baseline it is important to know where one should be headed with regards to markets and technology selection so that the appropriate energy efficiency programs can be identified and later designed and implemented. This is particularly important for resource acquisition programs. One tool for identifying the goals for a portfolio of programs is energy-savings potential studies. There are three main types of potential studies: technical potential studies, economic potential studies, and achievable potential studies.[370]

Technical potential is defined as the complete penetration of all measures analyzed in applications where they were deemed technically feasible from an engineering perspective. That is, the population of customers has facilities and technologies in which the energy efficient technologies or behaviors *can* be used.

Economic potential refers to that part of the technical potential that can *cost-effectively* be obtained when compared to supply-side alternatives. As technology and energy prices change, due to changing market conditions, including deregulation or other events, the levels of economic potential also change.

The *achievable potential* refers to the amount of savings that can occur in response to specific program designs and delivery approaches, including program funding and measure incentive levels. Achievable potential studies are sometimes referred to as market potential studies as they estimate what can be obtained within a specific set of

---

365 *Statewide Survey of Multi-Family Common Area Building Owners Market: Volume 1: Apartment Complexes*. (ADM Associates et al. 2000).

366 *Statewide LED Traffic Signal Saturation Study*. (Quantum Consulting Inc. 2001).

367 *Direct Assistance Market Saturation Study*. (Regional Economic Research 1991).

368 *Statewide Residential Lighting and Appliances Saturation Study*. (RLW Analytics 2000).

369 *Commercial and Industrial Lighting Retrofit Program: Base Equipment Saturation and Operating Hours by Building Type*. (San Diego Gas & Electric Company 1991).

370 *California's Secret Energy Surplus: The Potentials for Energy Efficiency.* (Rufo et al. 2002).

program design and implementation conditions presented to the target market. The achievable potential is what can be achieved through market interventions. It should be noted that the achievable or market potential is not fixed and is not easily estimated. The achievable or market potential is highly dependant on the programs offered, the technology or technology-mix presented, and especially the characteristics of the program relative to the approach used to reach the target markets, offer program services, and other program design and presentation factors. Over the years, policy makers have tended to rely on economic potential projections and the amount of energy that can be acquired before program costs make the acquisition no longer cost-effective. This type of assessment examines different program funding levels and program design approaches using market penetration assumptions to arrive at an estimate of what can be cost-effectively achieved. However, these types of estimates are often plagued by untested assumptions of what is required to obtain a given level of penetration within a specific market sector or market segment. As more process and market evaluations are conducted, policy makers are able to develop better, more accurate assumptions about what is needed from a program to achieve a given level of penetration.

The following diagram is presented in the *California's Secret Energy Surplus* report[371] and demonstrates the relationship between the market's technical, economic and achievable potential as well as the level of naturally occurring energy efficient change that occurs in the market.

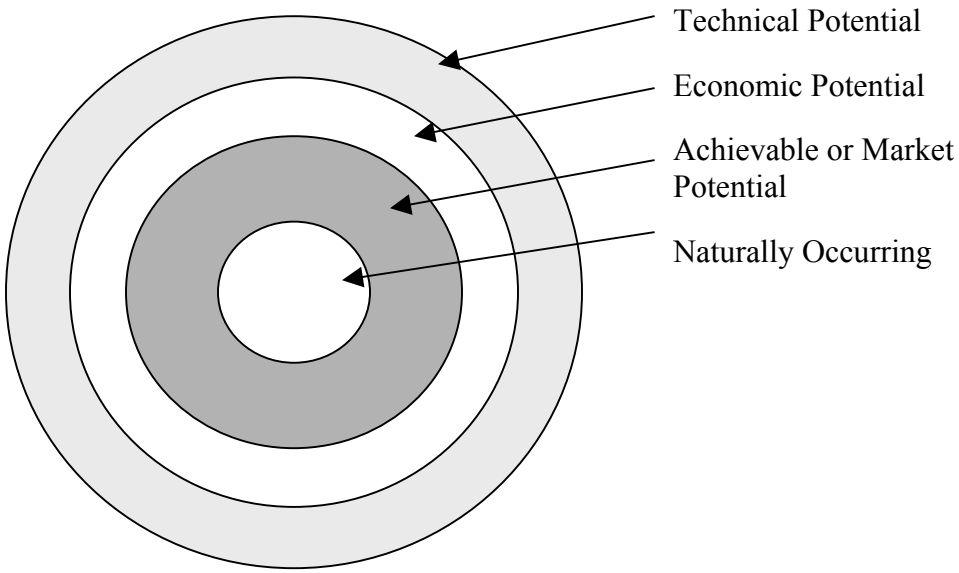

Figure 15.18: Conceptual Relationship Among Energy Efficiency Potential Definitions

These studies can be important decision tools as reflected in the 2003 CPUC rulemaking that require the CPUC to establish goals towards achieving the energy savings

[371] *California's Secret Energy Surplus: The Potentials for Energy Efficiency*. (Rufo and Coito 2002).

available.[372] Examples of past savings potential studies in California include work by Kema-Xenergy [373, 374, 375, 376a, 377], and Rufo and Coito.[378]

Persistence and Retention Studies

Persistence and retention studies are a class of evaluation studies that assess the long-term persistence of energy savings and effective useful life (EUL) of measures installed due to energy efficiency interventions. Persistence of energy savings is a combination of measure retention, measure installations retained versus those that have failed or that have been removed, and the incremental technical degradation (the difference between the technical degradation of the efficient equipment versus the degradation of standard equipment). The EUL, as defined by the prior measurement and evaluation (M&E) Protocols[379], is "an estimate of the median number of years that the measures installed under the program are still in place and operable."

The prior M&E Protocols had a required schedule for periodic program-specific retention and EUL assessment evaluations to be performed. Given this, there have been a large number of these studies conducted in California. (See the CALMAC searchable database at www.calmac.org.) The purpose of these studies was to estimate measure retention and assess whether ex-ante EUL appeared reasonable to retain (within the statistical confidence of the estimates or if it appeared to be the most reasonable assumption given unreasonable EUL ex-post estimates) or whether analysis provided the information to estimate an ex-post EUL that differed from the ex-ante EUL. Examples include studies done by Quantum Consulting (residential refrigerators),[380] Megdal & Associates (residential weatherization), [381, 382] ASW Engineering (commercial gas equipment, primarily cooking)[383], RLW (non-residential new construction)[384], Dohrmann

---

372 Assigned Commissioner's Ruling Proposing Direction and Scope for Further Rulemaking. (CPUC 2003).

373 *Conservation Potential Study, Vol 1., Results and Methods*. (Kema-Xenergy 1992).

374 *SCE Energy Efficiency Potential Study*. (Kema-Xenergy 1992).

375 *Commercial Sector Energy Efficiency Potential Study*. (Kema-Xenergy 2002).

376 *California Statewide Residential Energy Efficiency Potential Study*. (Kema-Xenergy 2003).

377 *California Statewide Commercial Sector Natural Gas Energy Efficiency Potential Study*. (Kema-Xenergy 2003).

378 *California's Secret Energy Surplus: The Potentials for Energy Efficiency*,. (Rufo and Coito 2002).

379 *Protocols, and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs* (* CADMAC 1999).

380 *Fourth Year Retention Study for PG&E's 1996 & 1997 Residential AEI Program Refrigeration Technology* (Quantum Consulting Inc. et al. 2001).

381 *Final Report: Statewide Study of the retention of Measures Installed Under the Direct Assistance Program.* (Megdal & Associates et al. 1998).

382 *Final Report: Measure retention Study -- 1994 & 1995 Residential Weatherization Programs (RWRI)*. (Megdal & Associates et al. 1999)

383 *Measure Retention Study of the 1996 Commercial Energy Efficiency Incentive (EEI) Program*. (ASW Engineering Management Consultants Inc. et al. 2001).

384 *SCE Non-Residential New Construction Persistence*. (RLW Analytics 1998).

(commercial, industrial, and agricultural equipment)[385], Richardson (residential new construction)[386], and Bordner (various).[387] An alternative method of estimating measure lifetime could be a study outside of a program that used a field study of age at replacement, as seen in a study done by Robison et al.[388]

Several overarching technical degradation studies were sponsored by CADMAC in the late 1990s as being a more cost-effective way to conduct these than through program-specific or even technology-specific studies. An article discussing the issues surrounding technical degradation and early survival analysis can be found in O'Drain and Caulfield 1998.[389] Most of these studies were conducted in the late 1990s (e.g., Proctor Engineering Group[390, 391, 392, 393, 394], and SBW Consulting[395]).

Survival analysis is the process of analyzing empirical failure/removal data in order to model a measure's survival function. A variety of types of survival analysis was used in the California retention studies to date. Besides finding relatively high retention rates in most cases, a consistent finding in these prior studies is that a longer period of time is needed for conducting these studies so that technology failure and removal rates can be better documented and used to make more accurate savings persistence and persistence impact models. This suggests that future evaluation policies should significantly alter the timing of retention and persistence studies.

As mentioned earlier, the prior M&E Protocols required program-specific retentions studies at specific times. Given the experience gained through these studies, a more cost-effective alternative might be to conduct retention studies by sector, across numerous programs and program years as overarching or consolidated evaluation studies. Though the retention analyzed could still be program-specific, conducting the data collection and analysis across a sector or across multiple sectors should provide a more cost-efficient alternative. Additionally, for larger commercial and industrial customers that can have

---

385 "A Longitudinal Study of Non-Residential DSM Measure Retention." (Dohrmann et al. 1999).

386 "Measure Retention in Residential New Construction," (* Richardson et al. 2000).

387 "The Application of Survival Analysis to Demand-Side Management Evaluation," (* Bordner et al. 1994).

388 "Measure Lifetime Derived from a Field Study of Age at Replacement," (* Robinson et al. 1996).

389 "Assessing Persistence: Experiences Documenting Savings Persistence Under the California Protocols," (* O'Drain et al. 1994).

390 *Summary Report of Persistence Studies: Assessments of Technical Degradation Factors, Final Report* (Proctor Engineering Group 1999).

391 *Persistence #3A: An Assessment of Technical Degradation Factors: Commercial Air Conditioners and Energy Management Systems, Final Report* (Proctor Engineering Group 1999).

392 *Statewide Measure Performance Study #2: An Assessment of Technical Degradation Factors, Final Report.* (Proctor Engineering Group 1998).

393 *Negative Technical Degradation Factors Supplement to Persistence Studies, Final Report.* (Proctor Engineering Group 1998).

394 *Statewide Measure Performance Study: An Assessment of Technical Degradation Rates, Final Report.* (Proctor Engineering Group 1996).

395 *Pacific Gas & Electric Company PY94 Nonresidential New Construction Retention Study.* (SBW Consulting Inc. et al. 1999).

participants in multiple programs across multiple years, this consolidated sector or cross-sector approach may be less intrusive to the participants.

Portfolio Analyses

This Framework is defining portfolio analyses as a class of evaluation studies that work with, evaluate, and assess the portfolio of Public Goods Charge programs; energy efficiency efforts across PGC and procurement, or energy efficiency, demand-response, and renewable energy sources.

The most common type of portfolio analysis, and the one the energy efficiency community is most familiar with in California, is one that aggregates energy and demand savings (a summative evaluation) with an effort to ensure no double counting of effects. This is done to estimate the PGC-funded total energy and demand savings. This estimate can be used to compare to the CPUC's energy efficiency goals as a way of measuring progress and can be used to inform power-planning efforts. The analysis could also be conducted for key market segments (e.g., residential, commercial, and industrial), or key end uses (e.g., lighting, motors, HVAC, etc.) or for the programs offered by a single administrator or group of administrators (e.g., PG&E's portfolio or third-party administrator X's portfolio) or for a given area (e.g. Orange County portfolio), or a portfolio of programs by size (e.g. programs that spend less than X dollars per year or acquire less than X kWh per year) or as is primarily used in this Framework, the Public Goods Charge portfolio, or perhaps in the future, the resource procurement portfolio. This type of evaluation could assess overall portfolio level cost-effectiveness, including all costs for all program types (resource, information, demonstrations, resource centers, infrastructure development). An example of this type of portfolio evaluation is the *Summary of 2001 Programs* completed by Global Energy Partners.[396]

A summative portfolio evaluation can be relatively simple or quite complex. For example, the analysis of the portfolio of programs could be a simple aggregation of the program and evaluation data collected on each program (e.g., costs, savings, number of measures, number of participants, etc). Nonetheless, techniques need to be incorporated to assess instances of double counting and minimize these occurrences. The methods needed to ensure there is no (or a limited amount of) double counting gets significantly more complex if there are both market effects estimates and resource acquisition estimates as part of the program effects being addressed. A similar but opposite concern that would add complexity to this type of evaluation would be when additional market transformation evaluations are conducted to provide input into a summative Portfolio evaluation. These efforts would be undertaken in order to measure market effects created by the aggregation of programs that were not a component of any other program's or set of programs' market transformation program evaluation (i.e., adding in savings missed from study aggregations that "fell through" the accounting cracks).

[396] *California Summary Study of 2001 Energy Efficiency Programs*, (Global Energy Partners 2003).

The above analysis is primarily a "bottom-up" assessment, adding up the program results and then working at minimizing overlap and filling gaps in that analysis. An alternative summative portfolio evaluation could be performed that would be a "top-down" approach. This could be done through evaluations of statewide energy usage, trends, changes, and the causes of these changes, one of which is the program interventions. Another portfolio evaluation could include matching, comparing, and assessing both approaches together in order to triangulate to an estimate that might help ensure the estimate is not biased. (Of course, this requires having studies that produced both types of estimates.)

An economic impact evaluation of program-induced efficiency investments is a type of non-energy effects evaluation. (See the Chapter 11 on Non-Energy Effects Evaluation for this Framework's decision roadmap concerning assessing these effects in California.) These types of studies are often conducted at a sector or portfolio level and, therefore, may be viewed as a type of portfolio evaluation. Economic impact evaluation of portfolios of programs are being done in New York[397] and Wisconsin,[398] and have been done for Austin Electric.[399]

Another type of related portfolio analysis that could be both summative and formative is analyzing the portfolio from a systems analysis approach and examining synergies between programs. Synergies could be either positive or negative. The analysis could be done to document synergies (summative) or as an evaluation to assess how positive synergies can be increased and better leveraged. This type of work has begun in New York.[400]

There are also several types of formative portfolio evaluations that could be undertaken. One of which would be to evaluate the portfolio within the financial perspective of the portfolio. In other words, to evaluate the portfolio in such ways as to obtain a portfolio risk assessment. This would include examining the tradeoffs between resource acquisition and market transformation; high risk and low risk programs/target markets; and pilot, equity, technology development support versus established acquisition programs. Finding the proper mix of investments to maximize savings at an appropriate risk level for PGC-funded or procurement-funded investments would not be a small task. Though the concept has been discussed,[401] we know of no actual formative portfolio evaluation of this type that has been conducted in the energy efficiency field.

---

397 "Linking Market-Based Energy Efficiency Programs to Economic Growth, Sustainable Development and Climate Change Objectives." (Smith et al. 1998).

398 "Quantifying Economic and Environmental Benefits." (Sumi et al. 2002).

399 "The Development of a Local Energy Efficiency Economic Impact Model for Use in Integrated Resource Planning." (Megdal et al. 1992).

400 "A Theory-Based Systems Approach for Evaluating Energy Efficiency Investments Portfolios." (DeCotis et al. 2001); "Systems-Based Portfolio Evaluation: Diagnostic Benefits and Methodological Challenges." (Eisenberg 2002).

401 "Portfolio Approach to Designing and Evaluating Buildings Energy Efficiency Programs." (DeCotis et al. 2000).

A more straightforward formative portfolio evaluation would involve combining potential studies, market studies, process evaluations, and market analyses with an evaluation to help guide portfolio investment decisions on what types of efforts to continue or fund in the future. As part of this, a gap analysis as a formative portfolio evaluation could be conducted to inform what types of programs, sectors, delivery mechanisms, or technologies to request for the next funding cycle.

Best Practice Studies

Another type of overarching study is the assessment of "best practices" within the energy program design and implementation arena. These studies document, for example, the best technology specific or customer type specific delivery approaches, such as the best lighting delivery approach for the small commercial sector, or the best marketing or customer contact approach, or the best measurement and evaluation techniques for certain types of technologies, programs or markets. These reports not only identify the key lessons learned in designing and implementing energy programs, but also provide guidelines for using the best practices. These studies help speed the rate of diffusion of the best practice in energy program design by identifying "best practices" and informing others about the practices and the results from the use of the practice.

Market Analyses

Market Structure and Operations Studies

Energy programs are typically designed for small sections of California rather than an area that is normally considered a technology marketing area. For example, the California public benefits funded programs have significantly more programs that target a small section of the state (utility market area, bay area, hard-to-reach area, language-targeted neighborhoods, Orange County, etc.) than programs that are offered statewide. However, few providers of energy consuming equipment set market boundaries for their corporate marketing efforts by these types of boundaries. Rather, most technology manufacturing and distribution companies organize their market areas around considerations for total cost, available profit and status of the competition within an associated distribution network. When energy programs attempt to affect a market by changing the way product choice and use decisions are made, most programs are essentially trying to change the behaviors within a section of a larger market.

If portfolio designers and policy makers are to understand the ability of energy programs to affect a market, they need to know about the characteristics of the markets they are trying to effect. Market structure and operation studies might be needed, for example, to help plan what types of programs are needed in particular markets. These studies typically go beyond single program-based market assessment initiatives and focus on the structure, function and operations of a larger market in order to help assess the ability of a set of programs to influence a smaller portion of a larger market. This is especially

important for market transformation efforts that are trying to change the normal operations of a market to be more energy efficient. Individual programs that focus on small sections of a much larger market will not have the same potential as a portfolio of programs targeting that market or a larger market. Without information on how markets are structured and operate, it is difficult to know if a portfolio of programs has the capability to make a lasting change in that market.

As a result, there is a need to conduct market assessments to support a portfolio decision in addition to market assessments that an individual program might need to design their program, or that an individual program evaluation might conduct to identify the effects of an individual program on a given market. The portfolio manager may decide that certain types of market analyses should be conducted at the portfolio level.

Market Share Tracking Studies

In addition to market structure and operations studies there is also a need to conduct overarching studies that track market shares of energy efficient technologies. These studies allow policy makers and portfolio managers to assess the ability of a set of programs to achieve additional market share. If the market share for a given energy efficiency technology is already high, then additional program efforts may not have a significant effect on influencing changes to that share. Likewise, if the market share for an energy efficient technology is low, and there is evidence that energy programs can be designed to increase penetration cost-effectively, such that policy makers may want to encourage programs to address this area. These studies use technology tracking from a portfolio perspective in that they examine market share as well as the consequences associated with market changes and program interventions. This data can allow additional portfolio analyses to assess where programs may be beneficial to push a specific technology in order to acquire cost-effective resources.

Updating of Key Parameters

In the design and evaluation of programs, energy-savings and cost-effectiveness calculations are based on key parameters that need to be updated periodically. These key parameters are the type of data that cut across several programs addressing multiple markets with multiple technologies rather than parameters that only apply to one or two programs. These studies would better and more cost-efficiently be addressed using overarching evaluation studies rather than a set of program-specific evaluations. For example, measure costs and cost differentials, avoided costs, load shape and peak demand information can be used with data from energy savings impact evaluations to estimate and understand a wider range of effects on participants. Elements within the Database for Energy Effiency Resources (DEER)[402] are critical portfolio planning parameters that need to be reviewed and updated over time so that ex-ante energy savings projections can be more accurate. As energy price and supply conditions change, it is

[402] (Xenergy, ADM Associates et al. 2001).

important to change the way avoided costs are calculated and the time periods for which these costs apply so that cost-effectiveness assessments can reflect current and expected time and load sensitive price conditions. Similarly, as energy technology prices change with the movement of the market there is a need to periodically identify the cost difference between the standard technology and the energy efficient product choices so that program incentives and program designs can be tailored to match market conditions. Likewise, as these overarching studies are conducted it is appropriate to update the deemed savings estimates and the DEER to allow the use of better energy savings estimates consistent with each technology.

Evaluation Methodology Development

As in any research field, evaluation methodologies need to be periodically reviewed and updated due to advancements in evaluation theory, practice, and measurement technologies. A review of evaluation approaches and methods and applicability to specific kinds of programs and research goals should be conducted periodically as the evaluation field evolves. This Framework represents one of these types of studies that examines evaluation methods and approaches in order to help assure that the best evaluation approaches are used to understand the energy supply potential and implementation effects of California programs.

Examples of these types of studies that might be considered for the future include:

- A vintage study on impact evaluations to learn what circumstances determine how often an impact evaluation needs to be conducted;
- A benefit-cost analysis of standardization approaches and the development and use of Designated Unit of Measurement (DUM);
- A study to assess the best, least biased methods to econometrically derive net-to-gross estimates;
- A study to develop a standardized survey method and analysis approach for establishing net-to-gross estimates, perhaps calibrated to one or more econometric approaches for use by smaller programs; and
- Studies to assess and support better linkages in the translation of evaluation outputs into procurement/supply analysis inputs.

Managing and Funding Overarching Studies

Overarching studies are different than program-specific evaluations and need to be budgeted and implemented outside of the program-specific evaluation framework. Because these studies complement the ability of the energy efficiency or energy procurement statewide portfolio of programs, they need to be organized from this higher-level portfolio or statewide energy supply and acquisition perspective. These studies

need to be planned and directed by an organization that has an overview perspective of the programs, technologies and services provided statewide, and an understanding of the potential effects of a wide range of programs. For these reasons, an organization like the CPUC or CALMAC can be considered for managing and funding overarching studies that benefit the ability of energy programs to achieve statewide energy resources.