

HANDBOOK OF EVALUATION OF UTILITY DSM PROGRAMS

ERIC HIRST AND JOHN REED
Editors

BENSON BRONFMAN, GEORGE FITZPATRICK, ELIZABETH HICKS,
ERIC HIRST, MARC HOFFMAN, KENNETH KEATING,
HARVEY MICHAELS, STEVEN NADEL, JANE PETERS, JOHN REED,
WILLIAM SAXONIS, ANDREW SCHÖN, AND DANIEL VIOLETTE
Authors

DECEMBER 1991

Sponsored by
New York State Energy
Research and Development Authority
David R. Wolcott, Project Manager
and
Office of Conservation and Renewable Energy
U.S. Department of Energy
Diane Pirkey, Project Manager

OAK RIDGE NATIONAL LABORATORY
Oak Ridge, Tennessee 37831
managed by
MARTIN MARIETTA ENERGY SYSTEMS, INC.
under contract No. DE-AC05-84-OR-21400
for
U.S. DEPARTMENT OF ENERGY

MASTER

EB

TABLE OF CONTENTS

PREFACE	vii
LIST OF ACRONYMS	ix
1. INTRODUCTION	1
BACKGROUND	1
WHAT IS EVALUATION?	2
TYPES OF EVALUATION	2
THE EVALUATION PROCESS	3
EVALUATION SKILLS	6
EMERGING TRENDS	6
HANDBOOK CONTENTS	11
CONCLUSIONS	12
2. PLANNING EVALUATIONS	13
INTRODUCTION	13
EVALUATION GOALS	13
BASELINE INFORMATION	18
PLANNING FOR INFORMATION-SYSTEM REQUIREMENTS	18
STAFFING AND BUDGETING	19
COORDINATION WITHIN THE UTILITY	21
SELECTING AN APPROPRIATE EVALUATION APPROACH	22
RESEARCH DESIGN	23
CONCLUSIONS	27
3. MONITORING THE IMPACTS OF DSM PROGRAMS	29
INTRODUCTION	29
OBJECTIVES OF MONITORING	29
WHAT DATA TO COLLECT	30
TYPES OF INFORMATION FOR IMPACT EVALUATIONS	32
SPECIFYING THE INFORMATION TO BE COLLECTED	41
DESIGNING THE SAMPLE	44
ENSURING DATA QUALITY	47
CONCLUSIONS	49
4. ANALYZING DATA	51
INTRODUCTION	51
STATISTICAL COMPARISON APPROACHES	52
USING PAIRED DATA TO ENHANCE PRECISION	55
MULTIVARIATE-SAVINGS-MODEL SPECIFICATIONS	60
SELF-SELECTION BIAS AND FREE RIDERS	63

SELECTED SPECIAL ISSUES	67
CONCLUSIONS	71
5. REPORTING AND USING EVALUATION RESULTS	73
INTRODUCTION	73
REPORTING EVALUATION RESULTS	73
USING EVALUATION RESULTS	81
CONCLUSIONS	88
6. PERSISTENCE OF ENERGY SAVINGS	89
INTRODUCTION	89
UNDERSTANDING THE ISSUES	89
METHODOLOGIES AND ISSUES IN PRIOR RESEARCH	93
CURRENT AND PLANNED RESEARCH	98
PROGRAM IMPLICATIONS AND CONCLUSIONS	98
7. USE OF LOAD RESEARCH IN EVALUATION	101
INTRODUCTION	101
LOAD-RESEARCH PROGRAM ACTIVITIES	102
THE COST OF LOAD RESEARCH	112
ORGANIZING LOAD-RESEARCH PROGRAMS	115
APPLICATIONS OF LOAD RESEARCH TO EVALUATIONS	115
JOINT UTILITY END-USE MONITORING	118
CONCLUSIONS	118
8. FREE RIDERS AND OTHER FACTORS THAT AFFECT NET PROGRAM IMPACTS	119
INTRODUCTION	119
FREE-RIDER MEASUREMENT	120
SURVEYS TO MEASURE FREE RIDERSHIP	121
MARKETPLACE DATA TO ESTIMATE FREE RIDERS	127
COMPARISON GROUPS TO ESTIMATE FREE RIDERS	129
STATISTICAL MODELS TO ESTIMATE FREE RIDERS	130
FREE DRIVERS	132
DISCUSSION AND CONCLUSIONS	133
9. PROCESS EVALUATION OF DSM PROGRAMS	135
BACKGROUND	135
KEY ISSUES	136
DATA COLLECTION AND ANALYSIS	141
REPORTING AND USING RESULTS	145
LINKS BETWEEN PROCESS AND IMPACT EVALUATIONS	145
CONCLUSIONS	147
ACKNOWLEDGMENTS	149

REFERENCES	151
ANNOTATED BIBLIOGRAPHY	165
INTRODUCTION	165
GENERAL EVALUATION METHODS	165
STATISTICAL AND ANALYTICAL METHODS	168
ENERGY EVALUATIONS	171
AUTHOR BIOGRAPHIES	175

PREFACE

Diane Pirkey
U.S. Department of Energy

and

David R. Wolcott
New York State Energy Research and Development Authority

We are pleased to have the opportunity to produce this handbook on the evaluation of utility demand-side management (DSM) programs. This volume is the result of a research project that was cosponsored by the Integrated Resource Planning programs of the New York State Energy Research and Development Authority and the U.S. Department of Energy.

Program evaluation has become a central issue in the world of utility integrated resource planning. The DSM programs that utilities were operating to meet federal requirements or to improve customer relations are now becoming big business. DSM is being considered an important resource in a utility's portfolio of options. In the last five years, the amount of money that utilities have invested in DSM has grown exponentially in most regulatory jurisdictions. Market analysts are now talking about DSM being a \$30 billion industry by the end of the decade. That is enough money to warrant serious efforts to determine how effectively utilities are carrying out the mission of implementing energy efficiency in their customers' facilities.

If the large volume of DSM-program investments was not enough to highlight the importance of evaluation, then the introduction of regulatory incentives has really focused the spotlight. Many regulators now understand the importance of aligning a utility's profitability with integrated resource planning goals as the means of obtaining the enthusiastic support of utility management for DSM. Various mechanisms have evolved to allow a utility to recover its costs and earn a bonus for successfully and aggressively acquiring DSM resources. The most popular mechanisms pay the utility either a "bounty" per unit of energy saved or a "shared savings" percentage of the total resource savings generated by a DSM program. In both cases, the incentive is based on the amount of energy saved, which can only be determined by some evaluation method. In this context, evaluation has gained the attention of all the participants in utility rate cases, from regulatory accountants to ratepayer advocates. Evaluation has come of age!

This handbook was developed through a process that involved many of those people who represent the diverse constituencies of DSM-program evaluation. For example, a team was

organized in New York, composed of regulatory staff, environmental advocates, utility personnel, and state energy policy planners. The group selected the handbook topics, suggested chapter authors, and produced a successful conference that brought the chapter authors together with almost 200 DSM practitioners. The resulting "chemistry," involving many different points of view, had a lot to do with the success of this handbook.

We have come to recognize the many technical disciplines that must be employed to evaluate DSM programs. An analysis might start out based on the principles of utility load research to find out what happened, but a combination of engineering and statistical methods must be used to "triangulate" an estimate of what would have happened without the program. The difference, of course, is that elusive but prized result of evaluation: what happened as the direct result of the DSM program. Technical performance of DSM measures is not the sole determinant of the answer, either. We also recognize the importance of such behavioral attributes of DSM as persistence and free ridership. Finally, DSM evaluation is meaningless without attention to planning an approach, communicating results to relevant decision-makers, and focusing as much on the process as the impacts of the program. These topics are all covered in this handbook.

We were honored that each of the authors of chapters in this handbook was our number one pick as the most authoritative person on each of these particular topics. This is a world-class group of authors, made more so by our editors, Eric Hirst and John Reed. We are indebted to both of them for having the vision to see the need for this volume, the technical competence to "pull off" a credible result, and the perseverance and discipline to produce such a professional product. We are sure you will agree that this handbook makes a significant and original contribution to the field of DSM-program evaluation.

LIST OF ACRONYMS

ACEEE	American Council for an Energy Efficient Economy
BPA	Bonneville Power Administration
CI	Confidence interval
C&I	Commercial and industrial
CMP	Central Maine Power Company
DOE	U.S. Department of Energy
DSM	Demand-side management
EDA	Exploratory data analysis
EIA	Energy Information Administration
EPRI	Electric Power Research Institute
ESCO	Energy service company
HSEM	Hybrid statistical/engineering model
HVAC	Heating, ventilating, and air conditioning
NEES	New England Electric System
PC	Personal computer
PG&E	Pacific Gas & Electric Company
PRISM	Princeton Scorekeeping Method
PSC	Public service commission
PUC	Public utility commission
R&D	Research and development
RFP	Request for proposals
SAE	Statistically-adjusted engineering
SD	Standard deviation
SH	Space heating
SIC	Standard Industrial Classification
WAP	Weatherization Assistance Program
WH	Water heating

INTRODUCTION

Eric Hirst and John Reed

BACKGROUND

The 1990 annual report from Consolidated Edison (1991), which serves three million customers in New York, states that the Company's goal

... is to provide reliable electric service at a competitive cost, in an environmentally compatible manner, while providing a fair return for our stockholders. Energy efficiency and diversity are keys to achieving this goal. During 1990 we dramatically expanded our programs to help customers use energy more efficiently and thereby limit growth in electricity usage. ... The company expects to spend \$140 million on these programs over the next two years. ... Our aim is to reduce usage of electricity by about 15 percent from what it otherwise would be by the year 2008.

Utilities and state public utility commissions (PUCs) in other parts of the country are making similar commitments to using DSM programs. Utilities in Maine, Massachusetts, Connecticut, New York, Wisconsin, Washington, Oregon, and California use DSM programs as alternatives to building new power plants. For example, the New York utilities will spend \$4 billion on DSM programs during the 1990s (Bradford 1991). Niagara Mohawk, Consolidated Edison, and Long Island Lighting, the three largest electric utilities in New York, plan to reduce projected electricity consumption by the year 2000 by 5%, 8%, and 9%, respectively [New York Public Service Commission (PSC) 1990]. And these utilities plan to cut summer peak demands by 10%, 13%, and 11%, respectively, that year.

As these DSM programs grow in size, cost, and expectations, measuring their effects and cost-effectiveness is increasingly important. Providing credible answers to certain key questions is vital: Are DSM programs providing the energy savings and load reductions expected of them? Are these programs cost-effective for utility customers? What can be done to increase program participation, lower costs, and boost energy and demand reductions?

Evaluation, the subject of this handbook, is the process that provides answers to these and other questions about the operation and effectiveness of DSM programs. The purpose of this handbook is to provide semitechnical guidance on how to plan and conduct evaluations and to point out the pitfalls to avoid. It is written for midlevel utility staff who manage and evaluate DSM programs. It also should be valuable to users of evaluation results. In particular, staff at PUCs, environmental groups, and other organizations participate in the planning and review of utility DSM programs; this handbook should help them become well informed reviewers and consumers of evaluation plans and evaluation results.

WHAT IS EVALUATION?

Evaluation is the systematic measurement of the operation and performance of programs (Hirst 1990a). Evaluations depend on objective measurements rather than anecdotal evidence and personal impressions. Evaluations use social-science research methods and technical data to produce valid and reliable results. And evaluations are intended to affect future decisions; they are not academic exercises (Kushler 1989).

Evaluations provide information for program managers and staff that they can use to modify and improve program operations. More broadly, evaluations provide information needed by utility executives and regulatory agencies to assess these programs. The results of evaluations

- Document the energy savings, load reductions, and cost-effectiveness of DSM programs
- Show how to improve programs by increasing participation rates, raising energy savings, or cutting costs
- Suggest ways to improve the design of future programs
- Support DSM budgets before the utility's budget committee
- Provide data to improve the company's load forecasts and resource planning

Utilities should measure the performance of DSM programs with the same competence and diligence with which they monitor the performance of power plants. During the past several decades, utilities have developed detailed information on construction costs and time; operation, maintenance, and fuel costs; heat rate, availability factor, and capacity factor; the duration and causes of each outage; and fuel consumption, plant output, and emissions for each of their power plants. Because the history of DSM programs is so short, comparable data generally do not exist for energy-efficiency and load-management programs on program-participation rates, energy savings (GWh), load reductions (MW), and program costs (Hirst and Sabo 1991).

TYPES OF EVALUATION

Evaluators typically speak of two types of evaluation. *Process* evaluations examine program operations to identify how well the program is implemented and to suggest ways to improve program delivery (Fig. 1). Such evaluations focus on program goals, history, and activities and often are based on interviews with utility program staff, program managers, participants, and trade allies.

Because process evaluations usually occur during program implementation, they can provide prompt feedback to help improve program design and implementation. In addition, process

evaluations provide important insights that can be used to interpret findings from impact evaluations.

Impact evaluations examine the effects of the program (Fig. 1). They provide quantitative documentation of program benefits and costs. Impact evaluations measure program participation, participant acceptance of the recommended DSM measures and practices, performance of the DSM technologies promoted by the program, program energy and load reductions, and program costs.

Impact evaluations compare what happened to program participants with what would have happened to participants if the program had not existed. These evaluations deal with two aspects of energy savings and load reductions. *Total* savings are the changes in annual electricity use and peak demand experienced by participants in the utility's program. *Net* savings are the portion of the total savings that can be directly attributed to the utility program (Fig. 2).

Net savings are the difference between total savings and the savings that participants would have achieved had the program not existed. Nonprogram savings reflect customer responses to changes in electricity and fossil-fuel prices, changes in economic activity or personal income, introduction of new electricity-using technologies, and other nonprogram factors. While it is possible to measure directly total savings, nonprogram savings cannot be measured because the program exists. Estimation of nonprogram savings requires a comparison group to approximate what program participants would have done without the program. Nonparticipating customers that are eligible for the program are often used as a comparison; sometimes, preparticipation data for participants is used as a comparison. Using data from nonparticipants and preprogram data from participants can often give the most reliable estimates of net effects.

Energy savings and load reductions are determined primarily by analyzing monthly electricity bills and load-research data. These electricity-use data are often supplemented with data on weather, occupant and equipment characteristics, facility characteristics, and program records. Depending on the purpose of the programs, evaluations focus on energy savings (GWh), peak-load reductions (MW), or changes in load shape.

THE EVALUATION PROCESS

Table 1 lists the activities associated with an evaluation. First, the purposes of the DSM program and the objectives of the evaluation need to be clearly identified. The goals of the evaluation depend on the goals of the program. In particular, the evaluation should focus on topics that will inform future decisions about the program. Evaluation goals, in turn, affect the design of the evaluation. If, for example, the program is just beginning, the evaluation will probably focus on program process because an impact evaluation would be premature. If, on the other hand, a pilot program is ready to be implemented systemwide and the expanded program is expected to contribute substantial energy or capacity resources, then an impact evaluation is appropriate.

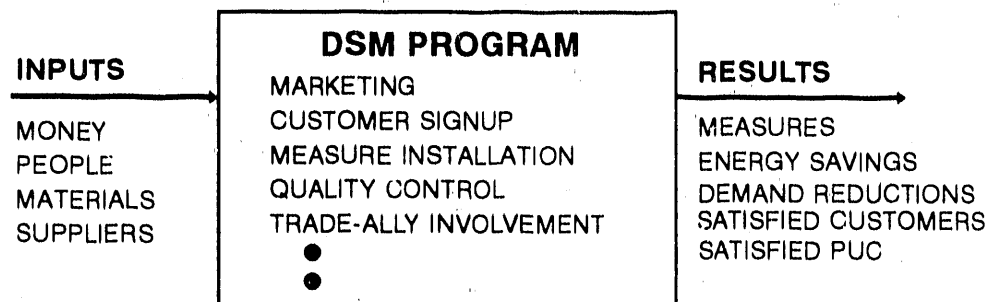


Fig. 1. Relationships among a DSM program, its inputs, and outputs. Process evaluation deals with the activities inside the box. Impact evaluation deals with the relationships between program inputs and results.

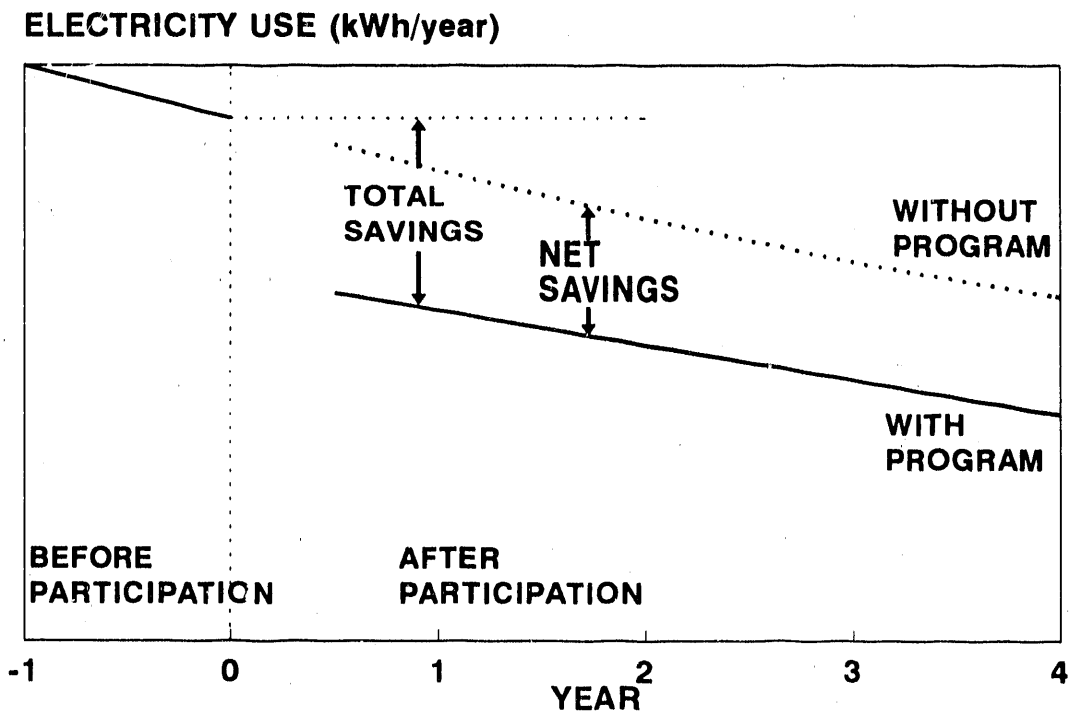


Fig. 2. Schematic of electricity use per participant for commercial buildings with and without a utility program aimed at improving heating, ventilating, and air conditioning (HVAC) systems. In this example, preparticipation electricity use is declining in response to increasing electricity prices.

Table 1. Sequence of activities in evaluating DSM programs

Plan evaluation

- Identify program goals and future decisions
- Determine purposes of evaluation and questions to address
- Identify evaluation resources (budget, staff, time, data, and analytical methods)

Implement evaluation

- Define populations to sample
- Collect data (program records, weather data, monthly electricity bills, customer surveys, interviews with program staff, and end-use load monitoring)
- Analyze data

Present evaluation results

- Review results with program implementation and planning staffs
 - Report results to utility executives and to the PUC
 - Prepare final report and briefings
 - Act on findings
-

The level of effort (dollars and staff) devoted to the evaluation needs to be determined at this stage. The evaluation should reflect the importance of the program itself. Evaluations need to be carefully planned so they address the major issues of the particular DSM program. Evaluations also should examine alternative explanations for what occurred (called threats to validity in the evaluation literature). Such threats include history (another event, such as an increase in electricity prices, that was responsible for the observed changes), selectivity bias (program participants and nonparticipants may differ in ways that affect their electricity use), and instrumentation (errors in the methods used to measure changes in electricity use). Finally, the evaluation plan should identify the data that will be collected, by whom, how, and when.

Implementing an evaluation includes the selection of samples of customers, utility employees, and vendors to interview and the development and testing of the questions to ask them. Other data sources, such as utility billing records, local weather data, and program records are identified and collected at this stage. The objectives of the program and its evaluation will suggest what types of data are most important. If the program is aimed at reducing demand during system peak (e.g., a direct load-control program), then load-research data may be required. If the program is intended to improve overall energy efficiency (e.g., through rebates for efficient appliances and lighting), then monthly billing data are suitable. These data then need to be merged, checked for errors, and analyzed.

Finally, evaluation results are prepared. This involves first a review of the results with program staff and management. Then results are presented to utility executives, and reports are prepared for internal and external distribution.

EVALUATION SKILLS

The evaluation group in a utility should include professionals with diverse educational backgrounds and experience. Appropriate skills include market research, economics, psychology, sociology, statistics, engineering, and business. The group should possess a mix of personal, communication, technical, and analytical skills to deal with the diversity of issues that arise in planning and conducting evaluations.

Technical skills are required to examine and use alternative data sources, including energy audits that identify cost-effective DSM options, and special meters and wiring used to measure end-use load-shapes. Analytical skills are needed to address a variety of potential problems that can confound interpretation of evaluation data. These problems include accuracy of self-reports in surveys, nonresponse to surveys, self-selection into the program,¹ and differences between early and later participants. Statistical skills are needed to define sample frames and suitable sampling methods and to define appropriate sample sizes. Computing skills are needed to manage and ensure the quality of the large and diverse data sets created for evaluations. Personal and communication skills are needed because collection of evaluation data and application of evaluation results depend on the cooperation of people in other departments within the utility.

A recent study of energy savings stimulated by retrofits in commercial buildings illustrates the complications that can occur in interpreting evaluation data and results and, therefore, the importance of capable evaluators (Exhibit 1).

EMERGING TRENDS

Surveys conducted by the Electric Power Research Institute (EPRI) show that U.S. electric utilities are conducting more than 1300 DSM programs. In 1990, these programs cut annual electricity use by 1.3% and cut summer peak demand by 3.7% (Faruqui et al. 1990). The energy and load impacts of utility programs are likely to increase dramatically during the next two decades (Hirst 1991). Several factors that affected utility DSM programs in the 1980s will change dramatically during the 1990s to increase the size, scope, cost, and effectiveness of DSM programs, including:

- Growing interest in integrated resource planning, which involves explicit consideration of DSM programs as cost-effective alternatives to some new power plants
- Increasing public concern about the environmental effects of electricity production and transmission, especially global warming and acid rain

¹Because participation in most programs is voluntary, there may be important differences between those who choose to participate and those who choose not to participate.

Exhibit 1. Engineering audits, billing data, and end-use data yield different interpretations

Puget Energy Services (1991) used engineering analyses of the measures actually installed, monthly billing data, and end-use load data to estimate electricity savings for a few commercial buildings in the Pacific Northwest. These buildings had been retrofitted as part of a conservation program run by the Bonneville Power Administration (BPA). Although, on average, the three sets of estimates agreed closely, the differences for individual buildings were quite large (Fig. 3). For only five of the eleven buildings with end-use load-research data did the engineering predictions fall within 20% of the "actual" savings. Three of the buildings showed *negative* savings, based on analysis of monthly electricity bills.

These differences in estimates of energy savings are less a question of finding the correct answer than of defining the appropriate question to ask. For example, monthly billing data and end-use data measure different properties. Billing data measure changes in electricity use for the building as a whole, whereas the load-research data measure changes in electricity use for particular end uses. Thus, changes in equipment and operating practices for end uses not covered by this retrofit program show up in the billing data but not in the load-research data or the engineering analyses.

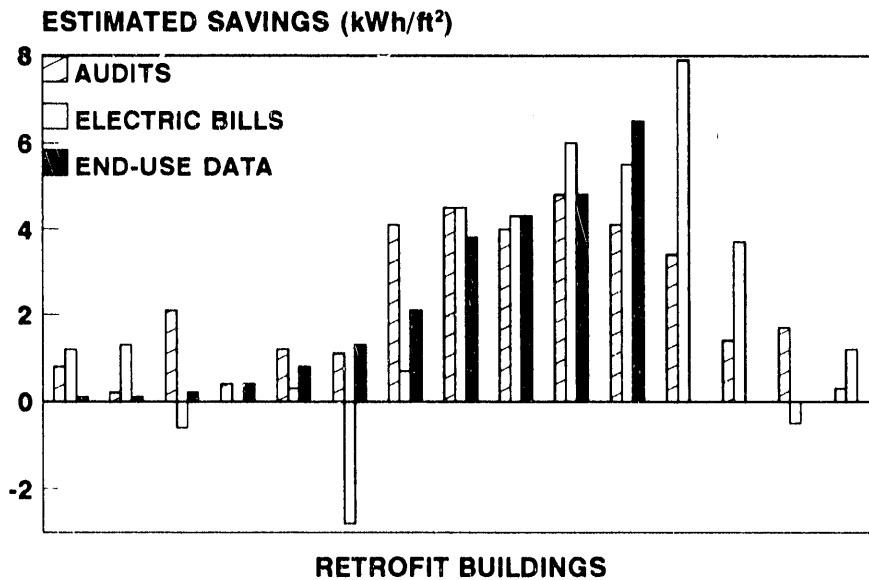


Fig. 3. Estimates of electricity savings in 15 retrofit commercial buildings in the Pacific Northwest (Puget Energy Services 1991).

- Provision of financial incentives to utility shareholders for implementing cost-effective DSM programs
- Growing recognition of the powerful role that utilities can play in overcoming market barriers to the adoption of cost-effective energy-efficiency opportunities

Perhaps the most important of these changing trends (and the one most likely to affect evaluations dramatically) is the provision of financial incentives. Such regulatory mechanisms already exist in Massachusetts, Rhode Island, New Hampshire, New York, and California and will soon be adopted in several other states. All seven utilities in New York State have incentive systems in place (Gallagher 1991).

The most popular incentives use shared-savings mechanisms, in which the utility keeps part of the net benefit provided by its DSM programs (Fig. 4). The net benefit is the difference between the total benefits and program costs. Total benefits are typically defined as the amount of energy saved by the program multiplied by the avoided energy cost plus the amount of demand reduction multiplied by the avoided capacity cost. The critical element in computing net benefits is estimating energy and demand reductions, the province of program evaluators.

Consider a hypothetical commercial lighting program as an example of the ambiguities in a carefully conducted evaluation. The program is aimed at medium-size office buildings and includes general information and onsite lighting audits. These activities identify suitable lighting measures and encourage the customer to apply for the 50% rebate offered by the utility. The rebate helps defray the costs of energy-efficient lamps, ballasts, fixtures, and controls.

The utility ran a comprehensive evaluation that included three elements, using approaches discussed by Violette et al. (1991) and XENERGY (1990):

- Analysis of two years of electricity billing data, one year before participation and one year after, for samples of participants and eligible nonparticipants
- Thirty days of time-of-use metering, pre- and postretrofit, of a sample of lighting circuits for a sample of participants only (no comparison group)
- Engineering analysis of the energy and load reductions caused by the measures actually installed by participants (again, no comparison group)

The utility used multiple methods to estimate program savings because each method is imperfect. If the utility knows how the results of these disparate methods will be used to determine program effects, such triangulation can build confidence in the estimates ultimately used. Not surprisingly, these approaches and their associated analytical procedures gave different estimates of net energy savings (Table 2); for simplicity, the load-reduction effects of the program are ignored in this example.

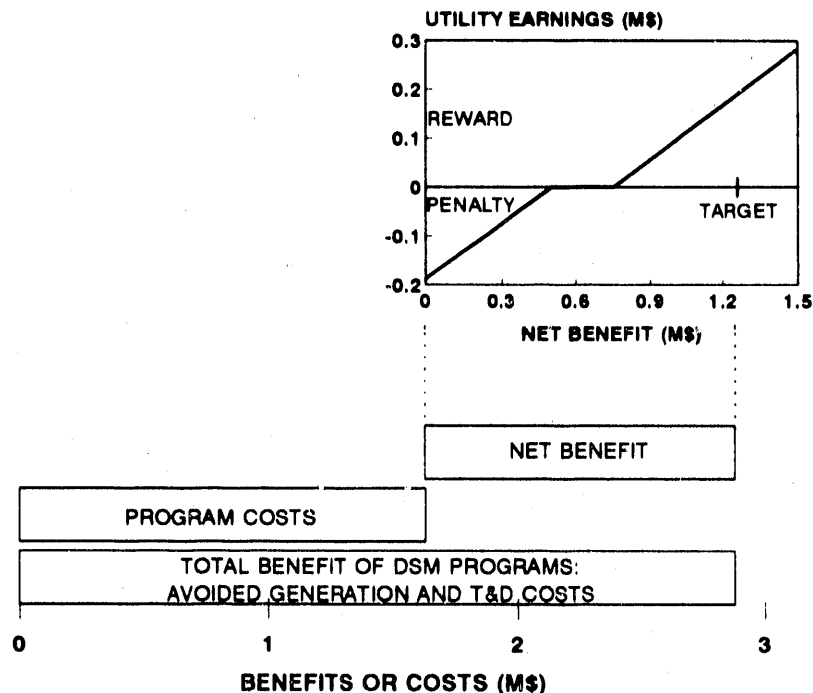


Fig. 4. Schematic showing the mechanics of a shared-savings mechanism to reward utility shareholders for implementing cost-effective DSM programs. In this example, the program is expected to cost \$1.6 million, the total benefits are expected to be \$2.9 million, and the net benefit is then \$1.3 million. Shareholders earn \$0.19 million if the utility achieves its net-benefit target of \$1.3 million.

Comparison of pre- and postretrofit electricity use, based on monthly billing data, for participants showed a reduction of 9600 kWh/year per rebate participant. Nonparticipants showed a small increase in electricity use during the same two-year period. (The local economy was growing during this period, which led to higher occupancy levels and longer hours of operation for office buildings. These changes in building use explain the increase in electricity use for nonparticipants.) And the customers who received a lighting audit but did not apply for the rebate (one-third of the total number of customers that received audits) cut their consumption slightly. Table 2 shows the roughly 50% difference in estimated annual electricity savings, based on analysis of billing data. These differences depend on whether participant savings are adjusted for nonparticipant changes in electricity use and for the savings achieved by audit-only participants.

The data from ad hoc metering showed savings about the same as those from analysis of the billing data for the participants only. Complications arose in scaling up the metering results to a full year. The metering covered roughly 60 days during a three-month period, with the middle month devoted to installation of new lighting measures. The amount of electricity used for lighting varies with season, and is larger in the winter than in the summer. On the

Table 2. Effects of evaluation results on estimates of DSM-program net benefits and utility incentive

Evaluation method	Results per rebate customer			
	Energy savings (kWh/year)	Gross benefit ^a (\$)	Net benefit ^b (\$)	Utility incentive ^c (\$)
Bill analysis				
Rebate participants only	9,600	2,300	680	0
With comparison group	12,800	3,100	1,450	260
Plus audit only ^d	15,200	3,650	2,020	480
Ad hoc metering	10,040	2,410	780	10
Engineering analysis	14,400	3,460	1,830	410

^aGross benefit = energy savings x \$0.24/kWh.

^bNet benefit = gross benefit - \$1625.

^cUtility incentive = 0.38 x (net benefit - \$753) if net benefit > \$753; otherwise incentive = 0.

^dSavings/rebate-participant = 12,800 + 0.5 x (0.15 kWh/ft²)(32,000 ft²), which reflects the 2:1 ratio of rebate/audit-only participants.

other hand, the indirect electricity savings associated with reduced air-conditioning (AC) loads is greater in the summer than in the winter.

Similar problems arose in estimating the savings with engineering calculations. These calculations were based on the change in connected load multiplied by the number of hours of use per year. The assumed change in connected load did not, however, properly account for the fact that many of the new lamps replaced ones that were burned out; thus the estimated reduction in load (and therefore in electricity use) was too high.

The percentage differences in electricity savings among these approaches are magnified when estimating net benefits (Table 2). This magnification occurs because net benefit is the difference between total benefit (directly proportional to energy savings) and program cost. In this example, the cost averaged \$1625 per rebate participant. With an avoided cost of 6¢/kWh and a measure lifetime of four years, the net benefit is roughly half the total benefit. Thus, a 10% error in estimating total benefit leads to a 20% error in net benefit.

As shown in Fig. 4, the utility incentive is typically a share of the net benefit, adjusted for a minimum threshold level. In this example, the target net benefit per participant was based on a planned savings of 11,000 kWh and a cost of \$1625, which yields a net benefit of \$1255. If the threshold is 60% of the target value, then the utility will receive an incentive only if

the net benefit exceeds \$753 per rebate participant. Again, following Fig. 4, the incentive fraction is 38% of the net benefit above the 60% threshold level. Thus, the structure of the incentive mechanism adds more leveraging in going from net benefits to utility incentive.

While the use of evaluation results in contested regulatory proceedings may become a particularly difficult issue for evaluators, many other challenges are likely to face utility evaluators (Table 3). For example, as the budgets and staffing for evaluations increase, it will become more important to coordinate evaluations with other activities in the utility, including market research, load research, load forecasting, DSM-program planning, and integrated resource planning.

Table 3. Challenges likely to face evaluators of utility DSM programs in the 1990s

Role of evaluation in contested PUC hearings

Integration of evaluation with other customer-analysis activities

Measurement of the long-term durability of energy savings

Definition of key DSM-program terms

Development of standard approaches for evaluation

Evaluation of the performance of DSM technologies

Measurement of the savings from DSM-bidding programs

Evaluation of DSM programs that aim to transform the market

HANDBOOK CONTENTS

Chapters 2 through 5 are organized around the evaluation process shown in Table 1. Elizabeth Hicks describes the steps associated with evaluation planning in Chapter 2. Harvey Michaels, Marc Hoffman, and Andrew Schön explain the various types of data that are collected for evaluations in Chapter 3. Daniel Violette presents alternative engineering, statistical, and combined techniques to analyze evaluation data in Chapter 4. And Steven Nadel explains how the voluminous evaluation data and analyses can be presented to different groups in Chapter 5.

Chapters 6 through 9 discuss several critical topics that pervade evaluations of DSM programs. Kenneth Keating presents data and suggests methods to measure the long-term energy and load reductions caused by utility DSM programs in Chapter 6. George Fitzpatrick discusses end-use load-research data (data that are particularly detailed, valuable,

and costly to obtain) in Chapter 7. William Saxonis discusses several factors that affect differences between total and net savings, especially free riders (customers who participate in a program, but would have adopted the program recommendations anyway) in Chapter 8. Finally, Benson Bronfman and Jane Peters provide a comprehensive description of process evaluations in Chapter 9, including a discussion of how process and impact evaluations should be integrated.

Clearly, a single volume – even one whose authors are as skilled, knowledgeable, and experienced as the writers of subsequent chapters – cannot cover all the information that utility staff need to conduct competent, thorough, credible, and cost-effective evaluations. The annotated bibliography and the reference lists suggest other publications that provide additional detail and insight into the topics covered here. In particular, the proceedings from the evaluation panel of the biennial conference on Energy Efficiency in Buildings (e.g., Keating and Hicks 1990) and the biennial conference on energy-program evaluation (Energy Program Evaluation Conference 1991) provide a wealth of material on the methods for and results of evaluation of energy programs.

Although the topics covered in this volume are extensive, limits on money, time, and weight (after all, we want readers to be able to lift this handbook!) forced us to ignore several important topics. Some of these topics, candidates for inclusion in later volumes, are evaluation of DSM technologies; comparison of actual vs predicted savings; low-cost evaluation methods; transferability of evaluation results from one utility to another; evaluation of rate-design programs; evaluation issues unique to different market segments and program types; multiutility evaluations; and use of evaluation results for load forecasting, resource planning, and DSM-program planning (Table 3).

CONCLUSIONS

The number and cost of DSM programs is increasing rapidly. Perhaps more important, the expectations for these programs – energy savings, reduced peak demands, and environmental quality – are growing also.

Careful evaluations can transform guesses, estimates, numbers, and data into useful information on the costs, performance, and operations of utility DSM programs. Program evaluation is integral to responsible management and is as important for DSM programs as for power plants.

Evaluations play a vital role in determining whether the ambitious goals utilities set for their programs are achieved. Without competent and credible evaluations, DSM programs cannot provide low-cost, environmentally benign, publicly acceptable alternatives to power plants. Evaluations are needed to legitimize and document the status of DSM programs as energy and capacity resources, equivalent in output and reliability to power plants.

PLANNING EVALUATIONS

Elizabeth Hicks

INTRODUCTION

This chapter discusses the critical steps in planning an evaluation (Exhibit 2). Perhaps the key is identifying evaluation goals. Of necessity, evaluation goals must be closely linked to program goals. The various stakeholders in an evaluation may have different goals and may differently perceive the importance of shared goals. Obtaining agreement on evaluation goals among the stakeholders is one way of building commitment to evaluation and the use of evaluation results.

Marshaling resources is an important part of any evaluation plan. One aspect is identifying baseline information requirements and locating the available information. Another part of the resource puzzle is mobilizing information-system resources. Determining budget and staffing is a significant element in any plan. The skills of the available staff will determine whether evaluations are done in-house or are contracted. Successful evaluations usually require the coordination of resources among many departments within the utility, which in turn requires a good deal of planning and political skill.

Selecting an evaluation approach is an important step. Which evaluation approach is chosen depends on a variety of factors, including the size of the program, the life cycle stage of the program, the available resources, the goals of the program, the goals of the evaluation, how the evaluation will be used, regulatory interest, the possibility of collaborative efforts, and the characteristics of the utility (including future capacity needs and the extent to which evaluation results may influence rates and incentives). Finally, developing a research design is a critical part of the evaluation-planning function. The research design identifies when measurements are made, which measurements are taken, and how the measurements are analyzed. A good research design will help to unambiguously answer the question of whether a program has been effective by ruling out alternative explanations for any observed effects.

EVALUATION GOALS

Evaluation needs should be considered at the initial stages of program design. Program design will influence evaluation requirements, the type of evaluation conducted, and the extent of evaluation activities. And evaluation design may influence the nature and timing of program activity (e.g., the type of data collected and the timing of data collection).

Exhibit 2. Evaluation planner's checklist

Managers and evaluators might refer to this checklist as they plan and conduct evaluations of utility DSM programs:

- The program has clear goals that utility executives, managers, program-implementation staff, and other interested parties understand and accept.
- The evaluation goals are also clear and explicit, and parallel the program goals.
- The interests of nonutility stakeholders are identified and, where possible, included in the evaluation plan.
- The evaluation plan addresses future program decisions. The information needed and the time when that information is required to support decisions are known.
- Utility management and staff are committed to the evaluation as reflected by the availability of time, money, and people to conduct the evaluation.
- Data requirements are known. Methods to obtain and manage information are established.
- Corporate information-systems managers are involved in identifying and meeting evaluation data needs.
- The evaluation plan and results are coordinated with other departments, including program implementation, program planning, load research, market research, load forecasting, resource planning, and information systems.
- The evaluation approach and cost are consistent with the size, expected energy and load effects, and importance of the program.
- The research design is clearly stated and is appropriate for the evaluation issues being addressed.
- The sampling plan is clearly stated and is expected to support the desired levels of accuracy.

Evaluation goals should be set in parallel with and should reflect program goals, and they will be partly determined by how the evaluation will be used, the maturity of the program being evaluated, and other factors (Exhibit 3).

Program managers, program planners, utility executives, regulators, and evaluators are stakeholders in the evaluation process, and will all have different goals for an evaluation. Regulators may want to minimize costs to consumers, the utility president may be interested

Exhibit 3. Relating evaluation goals to particular tasks

As part of a three-year, multimillion dollar evaluation of the U.S. Department of Energy (DOE) Weatherization Assistance Program (WAP), Beschen and Brown (1991) prepared a detailed evaluation plan. This plan began with enumeration of the purposes for this major evaluation. These evaluation objectives led, in turn, to establishment of several studies to address one or more of the objectives (Table 4). For example, the impact evaluation of single-family homes weatherized in 1989 will examine energy and nonenergy impacts, cost-effectiveness, factors that account for program success, and the remaining potential for energy savings.

Table 4. Relationship of evaluation goals to five studies in evaluation of the DOE low-income Weatherization Assistance Program

	Principal submarket studies			Related studies	
	Fuel oil study	Single-family study	High-density multifamily study	WAP network characterization	Eligible client profile
Estimate energy savings	●	●	●		
Assess nonenergy impacts	●	●	●		
Assess cost-effectiveness	●	●	●		
Analyze contributing factors	●	●	●	○	○
Characterize eligible population	○	○	○		●
Describe WAP network capabilities	○	○	○	●	○
Define promising opportunities	○	●	●	●	●

○ = Minor focus ● = Major focus

in maximizing revenues or making the company appear responsive to environmental concerns, a vice president may be interested in increasing control over program resources, and a program manager may be interested in demonstrating effectiveness as a program manager. All parties with a stake in the evaluation should be involved in setting the goals of the evaluation. This involvement may be accomplished by joint goal setting, collaborative processes, or regulatory proceedings. Changes in goals or in the scope of proposed evaluations need to be communicated in a timely fashion by the evaluators to the other parties (Hicks 1990).

Evaluations, particularly process evaluations, are inherently political. A clear understanding of the issues to be examined in a particular evaluation will likely reduce potential conflicts and will contribute to constructive comments as well as to the use of the evaluation results (McRae 1990). If evaluation results are to be accepted and used, all interested parties must be involved in setting the evaluation goals.

At the utility level, this policy implies that evaluation, planning, and implementation staff work jointly to establish evaluation goals, requiring a mutual understanding of program and evaluation tasks and processes. Each evaluation must be tailored to the individual program (Rossi and Freeman 1982). Outside parties who have been involved in program design as part of a collaborative planning processes and/or parties who may review programs and program results, such as regulators, should also be involved in developing evaluation goals. If evaluation results will be used to determine a financial incentive for the utility, all interested parties should jointly develop guidelines for this procedure and agree on the samples, data, and analytic procedures that will be used to determine the incentive.

How the evaluation results are to be used should be factored into the planning of the evaluation and the setting of goals (Chapter 9). Evaluations can provide feedback on program performance, improve planning estimates, and provide cost-benefit analyses of DSM programs. An emerging trend is to use impact-evaluation results to establish utility DSM-incentive payments. Implementers often have specific questions (e.g., what will be the effects of different rebate levels on participation rates or cost-effectiveness?) that need to be examined in the evaluation.

Utilities operate different types of programs, research and development (R&D) programs, pilot programs, full-scale innovative programs, and mature system-wide programs. Evaluation goals will vary with the type of program.

In an R&D program, a small number of customers is treated, and the results are measured carefully. Generally, the objective of this type of program is to develop better estimates of impacts of a technology or a delivery mechanism and to identify potential implementation problems. An example of such a program might be a space-heating (SH)-control experiment conducted in 50 homes with end-use metering. The purpose of the experiment might be to see how SH could be controlled without causing significant discomfort to customers. This type of experiment can give a good indication of the potential gross demand savings, but it would not be useful for determining free-ridership levels. Detailed monitoring of specific technologies, testing of delivery mechanisms, and innovative rate experiments are other types of R&D programs.

A pilot program is a moderately-scaled version of a program with several hundred to several thousand participants, which later may be run at full scale. The purpose of a pilot program is to improve the organizational efficiency of a program and to identify organizational and technological problems that could prove costly if the program is run at full scale. Process-evaluation results may be used to fine-tune the full-scale program design. Preliminary impact evaluations are performed with traditional methods, such as billing analysis, although the program may become a full-scale program before an impact evaluation is completed.

A full-scale innovative program is one that is open to all customers of a given class and that the utility is still refining. The refinements may include adding or deleting measures, changing the incentives, or providing more services to the customer. Either process- or impact-evaluation techniques might be used to evaluate this type of program. And these programs present an interesting challenge because the results from an evaluation can quickly become dated if the program changes. To illustrate, if the incentive structure for a program is changed, earlier free-rider estimates will no longer be valid.

Steady-state programs are those that have been operating for several years and have reached the point that program changes are minor and several impact evaluations have been completed with consistent results. Evaluations for this type of program may be focused on fine-tuning the program's operation and/or on examining the persistence of savings. Evaluations for this type of program may not require large amounts of resources and may largely be based on records maintained by program implementers and the billing department. Very few utility programs currently are at this stage although a few examples exist, such as Seattle City Light's multi-family retrofit program (Okumo 1990).

Many evaluation studies, regardless of cost or sophistication, are not used by practitioners (Franklin and Thrasher 1976). Part of the reason may be the lack of management commitment to evaluation. Obtaining management commitment is usually easier if the evaluator can provide evaluation results in a timely fashion and if the results are relevant to management decision making. In such a situation, the evaluator must anticipate decision-maker information requirements and the time frames in which decisions must be made. The evaluator must communicate with decision makers about their needs and may need to help them analyze their information requirements. The timeliness issue may shape the nature of the evaluation because it may require the use of results from other studies, it may limit efforts to a process evaluation, or it may limit impact measurements to a small sample. If management is committed to evaluation, the staff, time, and money needed to perform good evaluations will be made available, and evaluation findings will be used.

A major consideration in setting evaluation goals is determining the characteristics of the utility system. The evaluator needs to keep in mind such things as whether a utility is summer or winter peaking, whether it is energy or capacity constrained, and whether avoided costs vary significantly with time. These factors will guide how much emphasis to place on measuring energy or demand impacts and whether time-differentiated impacts are needed.

For example, two utilities may offer a similar commercial lighting program. Utility A is capacity constrained and has very large time-of-day and seasonal differentials in its avoided costs. Utility B does not need capacity for 15 years; most of its baseload generation is coal and it has a very small differential in avoided costs between on and off peak. Utility A may focus its evaluation on when the savings occur during the day and season while utility B may only need to determine total kWh impacts with little regard for load-shape effects.

BASELINE INFORMATION

A crucial element in planning for evaluations is taking stock of the internal and external sources of information about a utility's customers. Basic customer information is particularly useful in planning and evaluating DSM programs. At the very least it is useful to have load data by customer class or by building type as well as information on appliance characteristics and saturations and commercial and industrial (C&I) building stock (Fig. 5). This information, if not kept in the departments that perform DSM planning or evaluation, might be found in the load-research, market-research, or load-forecasting departments. Other possible sources include neighboring utilities, EPRI, the Energy Information Administration (EIA), the Bureau of the Census, and state agencies, especially state energy offices. If this information is not available, the evaluation group may have to collect it, and the cost of this data collection will need to be factored into budget planning.

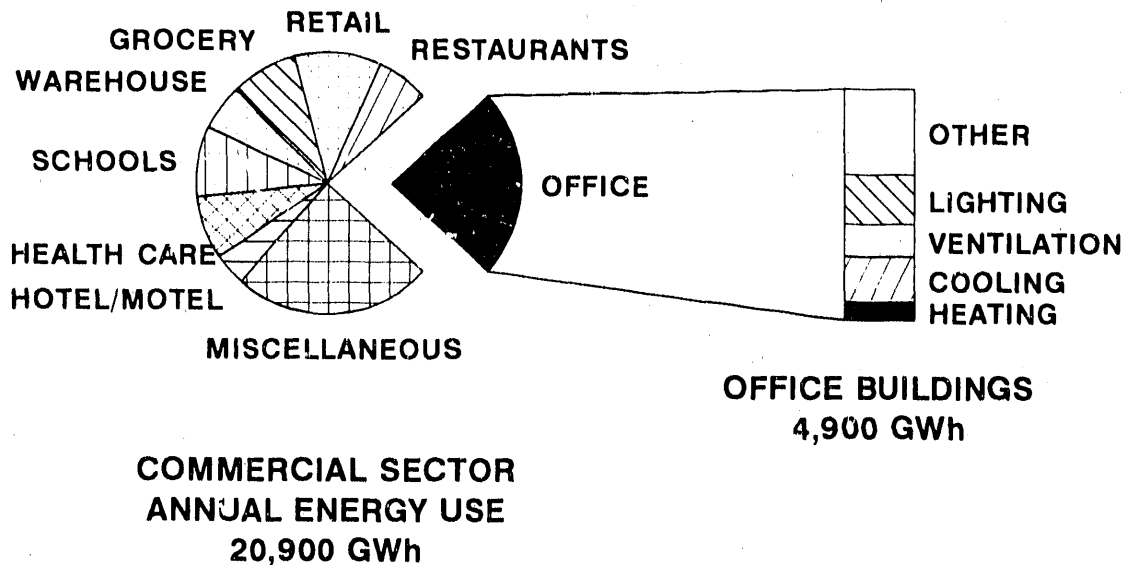


Fig. 5. Data and estimates on annual electricity use by building type and end use for the commercial sector in the Georgia Power service area. This information is important for DSM technology and program screening because it helps to identify the market segments with the largest potentials.

PLANNING FOR INFORMATION-SYSTEM REQUIREMENTS

Information systems that contain data on program activities and installations associated with a particular program are important to any large-scale evaluation effort. Depending on the utility and the nature of the program (R&D, pilot, or full-scale), information-tracking systems may be program specific or part of the larger corporate information-management systems.

The program evaluator needs to be very careful about creating a program-specific standalone information system that is independent of other corporate systems because evaluation data often need to be integrated with other corporate databases.

Generally, data from program databases and the corporate information system are needed to make initial estimates of program savings. Ideally, these systems contain information on such items as rate class, building type, Standard Industrial Classification (SIC) code, measures installed, hours of facility operation, and costs of measures, information that is essential to process and impact evaluations.

Evaluation requirements should be considered when program databases are being developed. Often this foresight is not practiced, and evaluation information is collected later through different systems. This difference can lead to inconsistencies that are difficult to reconcile during the evaluation. Evaluators should work closely with program implementers and information-systems people to ensure that their data-collection needs are met. Evaluators should also perform validity and quality-control checks throughout the life of a program (especially in the early stages) to assure the quality of the data. Evaluators should discuss their data requirements with the corporate information manager before the program or the evaluation gets underway rather than after it is started.

STAFFING AND BUDGETING

Once general evaluation planning has started, the utility will need to consider both budgeting and staffing. Clearly, the evaluation effort should be commensurate with the scope of the program, the importance of the program, and the use to which the evaluation results will be put. The cost of an evaluation can range from 1 to 10 percent of DSM spending. Staffing may vary greatly, depending on corporate structure and how consultants are used.

Staffing

An early consideration is where evaluation fits in the corporate structure. Utilities handle this differently, and there is no right answer. Typical structures include:

- Evaluation and implementation can be kept totally separate (e.g., under a different vice president) as is currently the case at Central Maine Power (CMP)
- Evaluation and implementation can be placed under the same vice president with different directors or managers, as is currently the case with New England Electric System (NEES), Pacific Gas & Electric (PG&E), and Northeast Utilities
- Evaluation can be combined with program planning with or without the same vice president as implementation
- Evaluation can be combined with load research or market research
- Evaluation can be combined with load forecasting

Each of these structures has its strengths and weaknesses. Separating evaluation from implementation avoids perceptions of conflict of interest and is effective provided frequent and good communication occurs between the groups. Having evaluation and implementation under the same vice president can ensure that conflicts are resolved quickly but may lead to perceptions of a conflict of interest. This arrangement also increases the likelihood that evaluation results will be used in revising and planning programs. Combining evaluation with planning has the advantage of sharing personnel and skills but may lead to situations at startup in which personnel are more focused on planning and refining programs than on evaluating programs (Franklin and Thrasher 1976). Combining evaluation with load research can lead to strong data collection. A combination with market research may provide economies of scale in company-wide market research as well as provide evaluators with good access to customer information; but it may tend toward a focus on market issues rather than technological or persistence issues. It is hoped that a utility will structure evaluation activities to achieve the strongest possible balance among competing interests.

Another consideration concerning staffing is the amount and type of evaluation work to be done by in-house personnel and by consultants. This decision will determine the type of personnel that the in-house evaluation group will need. For reasons of credibility and objectivity, utilities generally prefer to have consultants perform process evaluations. This is not to say that a distinctly separate in-house evaluation group could not undertake process evaluations. Indeed, when evaluation and implementation had different vice presidents at NEES, some process evaluations were done in-house. Even in this situation, though, a utility might use a consulting firm to conduct survey work unless the company has a strong market-research department capable of doing its own surveys and having its own phone bank for surveys. Most utilities will probably opt to contract for process-evaluation work.

Impact evaluations can involve analysis of customer electricity bills, metering studies, and engineering studies. These can be done by in-house staff or by consultants. Billing and other statistical analysis of customer kWh data typically involves sample design, extraction of customer information from implementation databases, extraction of billing data from billing files, customer surveys, and the analysis of bills. Even if a consultant is used, the utility staff will probably have to extract the data from the implementation and billing databases, a time-consuming step. The decision to complete the other tasks with in-house personnel should be based on the qualifications and availability of utility staff. A utility with strong load-research capabilities and a strong metering department may find it easier and less expensive to do its own metering. Significant advantages accrue to having utility staff who understand metering devices and are available to maintain metering equipment. In some cases, a utility may want a mix of direct employees and contract workers: a utility might use its own staff to place end-use meters on residential water heaters and to place whole-building meters at SH-customers' sites, and might use a contractor to install end-use meters on HVAC systems. Engineering analysis could be done by utility staff or by a contractor.

In-house personnel requirements are determined by the decision on how much work will be done by consultants. If most evaluation work is being done by consultants, the utility need not have specialists in every evaluation area. In this situation, utility staff should probably

be good project managers and have experience managing contracts. As discussed in Chapter 1, appropriate backgrounds include market research, economics, psychology, sociology, policy analysis, statistics, and engineering. If the utility chooses to do much impact-evaluation work in-house, expertise in certain areas is essential. These areas include econometrics or statistics for statistical and billing analysis, engineering analysis, and electrical and electronics engineering for load-research and metering studies.

Evaluation staff need to have good analytical and communication skills. The ability to write effectively is important for presenting evaluation results in a usable format. Computer skills, particularly knowledge of spreadsheet and statistical packages, are essential. Good interpersonal skills are needed to be able to interact with people from other departments.

Budgeting

Budgets will depend on the number of programs being evaluated, the available staff, the amount of work being done by consultants, and whether some work is being done by other departments. Often, in the first year of a program, only a process evaluation will be conducted because it is too early to use impact-evaluation techniques.

Table 5 provides crude guidelines for budget-estimation purposes. Clearly, the utility will need to follow its own procedures in determining a budget. Metering costs are not included, and they will vary greatly depending on the type and duration of metering and whether metering is done by utility staff or by a consultant. The load-research and metering staffs at the utility can provide good estimates of metering costs; (see also Table 22 in Chapter 7).

Table 5. Typical evaluation costs

Item	Person-hours	Approximate cost @ \$90/hour (1991-\$)	
Process evaluation			
Residential program	300 – 800	27,000	– 72,000
C&I program	600 – 1,500	54,000	– 135,000
Impact evaluation, billing analysis			
Residential program	400 – 700	36,000	– 63,000
C&I program	600 – 1,800	54,000	– 162,000

COORDINATION WITHIN THE UTILITY

Regardless of the structure of the utility, evaluators must be prepared to work closely with the following departments: program implementation (field staff), load research, information systems, metering, program planners, load forecasting, market research, and resource

planning. Coordination with program implementers is vital for producing usable and reliable evaluation results. Evaluators should be involved with program planning, and implementers should be drawn into the evaluation process early and provide continuing feedback. The load-research and metering departments are critical if the company plans to do any metering on its own. Selection of metering equipment and protocols needs to be consistent with that for any existing load-research system. Such issues as additional workload for union employees may need to be addressed with the metering department. Interaction with the information-systems department is very important. If the information needed by evaluators is collected at the outset, subsequent tasks will be easier. The load-forecasting and market-research departments may have baseline data that can be used by evaluators. In turn, evaluators can provide these departments with additional information on customers.

SELECTING AN APPROPRIATE EVALUATION APPROACH

The selection of an evaluation approach for a given DSM program depends upon the size and the life-cycle stage of the program, the available resources, the goals of the program, the goals of the evaluation, how the evaluation will be used, regulatory interest, the possibility of joint efforts with other utilities, and whether multiple evaluation approaches might be appropriate.

Process evaluations usually examine participation, nonparticipation, operational efficiency, contractor-utility interactions, usefulness of the corporate and implementation databases, effectiveness of marketing, customer satisfaction, implementation effectiveness, and free ridership. The methods used in process evaluations usually consist of customer surveys, interviews with appropriate staff and contractors, and field visits to locations where measures are installed. However, the foci of the process evaluations tend to vary. Evaluations of new programs often focus on customer satisfaction with the program and suggestions for program improvements. If free ridership or program delivery are of concern, these issues may be examined in more detail (e.g., by asking retailers their impressions about changes in consumer buying habits or by comparing sales figures for specific models of appliances before and after a program has started). The scope of the process evaluation for a given program must be based on a utility's assessments of its needs, concerns of regulators and other parties, and the resources available for the evaluation.

The choice of impact-evaluation method may be influenced by the use to which the results of the impact evaluation will be put (will it be used for regulatory incentives?), the stage of the program (is this a new program and how many participants does it have?), the costs and size of the program, the utility operating characteristics, and regulatory interest. More rigorous approaches and multiple methods are appropriate when the utility receives money (e.g., DSM incentives for shareholders or wheeling of conservation) based on the evaluation results; where large resources are being expended on a given program; and, in R&D programs, where the evaluation results will be used to develop a larger-scale program.

As an example, consider two utilities with weatherization programs aimed at reducing electricity use for SH. The first utility has an immediate capacity need. Its peak is driven

by heating, it receives incentives based on evaluation results, and it is spending about 1% of its annual revenues on this program. The second utility does not need new capacity until the year 2005. It is summer peaking, it just receives cost recovery on its conservation programs, and it is spending 0.1% of its revenues on this program. Both utilities might perform process evaluations, and they also might estimate kWh savings through billing analysis. The first utility also might use end-use metering to evaluate the capacity benefits from the program and to verify both kW and kWh savings because incentives are involved. The second utility may not use end-use metering and use those evaluation dollars elsewhere because it has no immediate capacity requirements, is summer peaking, and is offering no incentives.

Evaluators should remember that, for some programs (e.g., a low-flow showerhead give-away program), accurate estimates can be obtained with simple techniques (e.g., engineering estimates) and that the use of more-sophisticated and -costly techniques may not buy corresponding increases in accuracy (EPRI 1991a).

Certain types of programs lend themselves to combined efforts with other utilities. An example might be a program to develop baseline savings estimates for new construction practices. Four Massachusetts utilities used this approach (XENERGY 1991; New England Power Service Company 1991). A combination program of metering and statistical estimation of load-shape impacts has been proposed for New York state (XENERGY 1990). Several joint evaluations have been conducted in Wisconsin (Prah 1988). Also, several utilities in New England are performing a joint evaluation of an appliance-efficiency program.

RESEARCH DESIGN

The savings from a DSM program are estimated by comparing energy or demand prior to a program with energy or demand after a program has been implemented. Figure 6 is a conceptual representation of this process. One approach is to compare the posttreatment consumption in year three (Point C) with preimplementation consumption (Point A). The difference (A - C) represents the effects of the program. However, this difference also includes the effects of all the other changes that occurred in the intervening three years. If this customer were a commercial building, this difference might include changes in occupancy in the building; increased energy-use intensity caused by the introduction of new equipment, such as laser printers; increased or decreased activity in the building because of changes in the economy; etc. If it were a residential dwelling, a variety of factors might affect energy or demand, including changes in the composition of the household (e.g., the addition of a baby or a young adult leaving home); the introduction of new appliances; the replacement of older appliances that were not part of the program with new more-energy-efficient appliances; or changes in lifestyle, such as the maturation of young people.

Thus, to accurately estimate the savings of a DSM program, one needs to measure the difference between the electricity that would have been consumed over time without the program and what is actually consumed with the program. Figure 6 shows that without the

ELECTRICITY USE PER PARTICIPANT (kWh or kW)

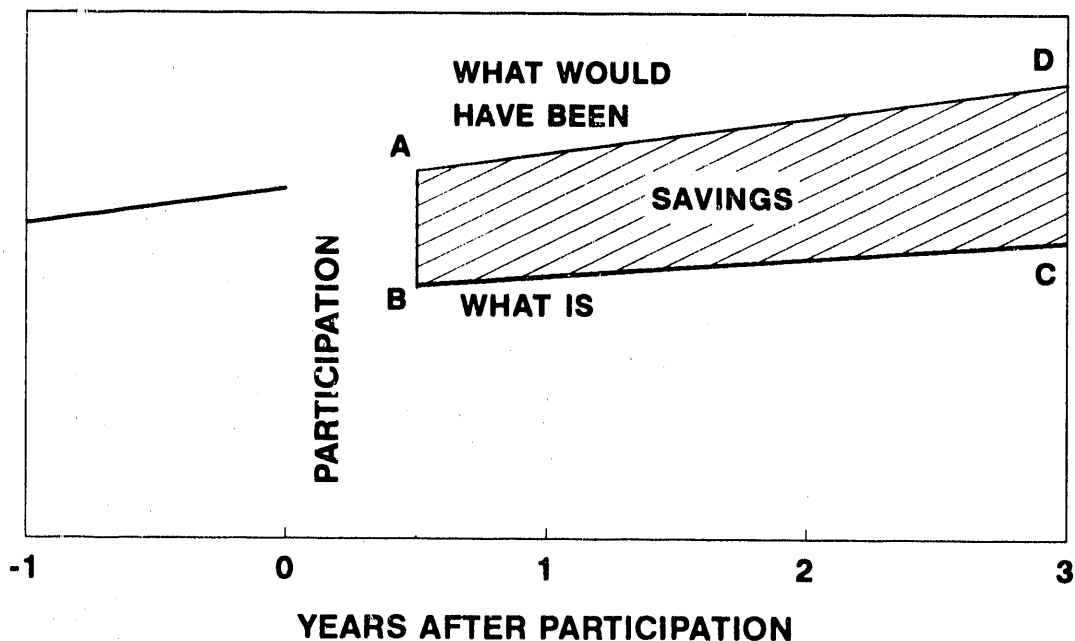


Fig. 6. Schematic showing electricity use before participation in a DSM program, after participation in the program (BC), and an estimate of what electricity use would have been without the program (AD). The difference A - B is the immediate postparticipation change in electricity use. The difference D - C is the long-term change in electricity use.

program, demand or consumption increased during the three-year period, the line with A and D as endpoints. Even with the DSM measures, energy or demand increased some, as represented by the line BC. Thus, the savings are greater than would be represented by the quantity A - C and would perhaps be better represented by the quantity, D - C. The cumulative savings are represented by the area, ABCD.

The purpose of research design is to minimize the impacts of extraneous effects of the type described above. These extraneous factors are rival hypotheses or "threats to internal validity." Threats to internal validity arise when the differences that exist can be explained in part or in full by other phenomena. Cook and Campbell (1979) list seven potential threats to internal validity (Table 6) that, if not controlled by the design, can bias results.

In addition to internal threats to validity, there are external threats to validity. External threats are those factors that make generalization of findings difficult. An example would be findings from a sample of customers that is not representative of the larger population of interest. Another example is a reactive or interactive effect of testing (e.g., where pretests of marketing materials during a pilot program lead to a decreased sensitivity to the materials during a full-scale program). Yet another form of external threats to validity is the interactive effect between selection biases and the experimental variable. In this situation, customers who are innovators may be more inclined to participate in a technological

Table 6. Examples of threats to internal validity

Threat	Example
History	Another event such as an economic crisis, a war, or an "oil shock" causes a change in electricity consumption.
Maturation	Natural changes, such as an addition to a family, component aging, or changes in lifestyle, occur which cause changes in energy consumption.
Regression to the mean	Buildings or households chosen for an initial measurement from the extremes of a distribution tend to have measurements nearer the middle of the distribution on subsequent measurements. If high energy users are targeted for the program and measured in the initial interval, high energy use may not be sustained in subsequent measurements.
Selection	Selection is the situation in which participants are systematically different than the remainder of the population. Selection bias is often a problem with voluntary programs.
Testing	People adjust their behavior or adapt to changed circumstances on the basis of programmatic activities. People may adjust their thermostats in response to AC load-control tests. People "take back" some of the savings from a weatherization program.
Mortality	Participants drop out for some reason such as relocation or dissatisfaction with the program.
Instrumentation	Changes in how the impact variables are measured account for the changes. A question on a survey is changed, which causes people to answer the question differently because they assume a different context, or metering is changed from whole-building metering to end-use metering between the pre- and postperiods.

program than would the general population. Thus, if projections of reductions in energy usage from people volunteering for an advanced-heat-pump program includes a disproportionate number of innovators, the total amount of savings may be overestimated.

Campbell and Stanley (1966) argued that the "classical control group" design minimizes these threats. In this design, pre- and posttreatment usages by a group of participants and a group of nonparticipants are compared. The key to the classical design is the *random* assignment of customers to the participant and nonparticipant groups. Randomly assigning customers

to both groups reduces the likelihood that these groups will be biased in some way or that the results will not be generalizable. Also, in the strictest sense, participants and nonparticipants should not know to which treatment group they are assigned and nonparticipants should be isolated from treatment-related information received by participants. With the classical design, the change in energy usage is estimated as follows:

$$\text{Change in usage} = (\text{participant pretreatment usage} - \text{participant posttreatment usage}) - (\text{nonparticipant pretreatment usage} - \text{nonparticipant posttreatment usage})$$

Unfortunately, practical circumstances almost always preclude the use of the classical design. Examples of such circumstances might include a PUC order that a program must be offered to all customers or an order that the same incentives be available to all customers. Each of these orders would preclude the random assignment of customers. Yet another case would be a rebate program in which it is impossible to treat some customers without influencing others.

In these circumstances, the evaluator must look to alternative designs to minimize, but not necessarily eliminate, threats to validity. Table 7 presents some of these options. For these designs, the nonparticipant group is called a comparison group because nonparticipants are not randomly assigned (as a "control group" would be) and thus are not strictly equivalent.

Table 7. Research-design options

Design option	Pretreatment period	Participation	Posttreatment period
Classical design (random assignment)			
Participants	0	X	0
Nonparticipants	0	--	0
Nonequivalent comparison group			
Participants	0	X	0
Nonparticipants	0	--	0
Cross-sectional comparison group			
Participants	--	X	0
Nonparticipants	--	--	0
Time-series comparison participants	0	X	0

0 = observation; X = participation; -- = no data

However, this does not mean that participants and nonparticipants may not have been randomly selected.

The nonequivalent comparison group design tries to make up for the absence of random assignment of participants and nonparticipants to treatment groups by constructing a comparison group that is similar to the participant group. This fashioning of the comparison group may be done by matching the characteristics of nonparticipants with selected characteristics of participants or by drawing a random sample weighted for a characteristic (or group of characteristics) from the population represented by the participants. An example of this approach might be to compare those who volunteered for a program with a weighted sample of customers based on total kWh sales during a pretreatment period. If the two groups are comparable, then threats to internal validity may be minimized, but external validity threats (i.e., the generalizability to the larger population) may remain.

In a design with a cross-section comparison group, the energy usages of two groups (participants and nonparticipants) are compared for the posttreatment period only. If the selection of participants is random or the two groups are well matched, then many of the threats to internal validity are minimized, although there still may be problems with generalization. Often, this design is used in situations where an evaluation is initiated after a program has begun or where the utility does not have historical billing data for a group of customers or where no pretreatment period is available (as in new construction). If evaluation planning is part of the program planning, problems of obtaining pretreatment data will be minimized.

Another design, a time-series comparison group uses the participants as their own comparison. For example, in a load-control program, electricity use on noncontrol days with similar weather is often used as the baseline against which to compare electricity use on control days. Figure 14 in Chapter 7 illustrates this approach for a residential air-conditioner control program. For such programs, this time-series comparison is a strong design because the short time span and the use of weather conditions to select noncontrol days preclude most threats to validity. This type of comparison group might apply to other programs as well. If the participants are randomly selected, then threats to internal validity, with the possible exception of history and perhaps instrumentation, are minimized. Many of these problems can be dealt with by the use of normalization techniques or the use of multivariate analysis. If participants are not randomly selected, then many of the threats to validity remain.

It should be clear from the preceding discussion of research designs that sample design is a critical issue. The lack of a good sample design can pose threats to both internal and external validity. This subject will be discussed in some detail in the next chapter.

CONCLUSIONS

Planning an evaluation involves an array of activities. Many of these activities revolve around identifying the goals of the evaluation and obtaining agreement on the goals from

stakeholders. Agreement on goals is essential for obtaining commitment to the evaluation and for ensuring that evaluation results are used. A clear set of evaluation goals are also essential to ensure that the appropriate data are collected and available for analysis when needed.

The choice of evaluation approach depends on the type of program, size of program, customer segment, and ultimate use of the evaluation's results. Finally, the choice of a research design is an important part of the evaluation plan. A good research design will guide the where, when, and how of data collection and it will help ensure that the results can withstand the challenges of alternative explanations of the findings.

MONITORING THE IMPACTS OF DSM PROGRAMS

Harvey Michaels
Marc Hoffman
Andrew Schön

INTRODUCTION

This chapter addresses monitoring and data collection. Its objective is to describe the major options available to the DSM-program evaluator for data collection, data specification, and sample design and to provide an assessment of the strengths, weaknesses, resource requirements, and implementation approaches for the various data-gathering techniques.

Monitoring is the process of collecting the data needed for the analysis of energy savings, including identifying the data requirements to establish energy savings, normalizing energy use for nonprogram factors, estimating free-rider impacts, and determining persistence of savings. Monitoring includes data specification, survey design, sample selection, metering installation, data collection, quality control, and reporting.

The selection of data-collection method, data specification, and sample design are key decisions that must be made jointly by those responsible for planning, monitoring, and analysis. These decisions determine the cost of the evaluation and the quality of the analyses that can be performed. This chapter discusses the objectives of monitoring, the strengths and weaknesses of various sources of monitoring data, how to specify the data to be collected, sample selection, and data-quality issues.

OBJECTIVES OF MONITORING

The primary objective of monitoring is to provide systematic, defensible, and objective measurements for a DSM-program evaluation. Monitoring, therefore, should be carried out in a manner that achieves a quality research standard, as the following standards identified for educational evaluations (Joint Committee on Standards for Educational Evaluation 1981, partial and edited; Patton 1982a):

The purposes and procedures of the evaluation should be monitored and described in enough detail so that they can be identified and assessed.

The sources of information should be described in enough detail so that the adequacy of the information can be assessed.

The information-gathering instruments and procedures should be chosen or developed and then implemented in ways that will assure that the interpretation arrived at is valid and reliable for the given use.

The data ... should be reviewed and corrected so that the results will not be flawed.

The evaluation procedures should provide safeguards to protect ... against distortion by the personal feelings and biases of any party to the evaluation.

WHAT DATA TO COLLECT

The purposes of data collection are to establish the quantity of energy and peak-demand savings and to support comparisons, projections, and program design. Almost all impact evaluations seek to quantify kW and kWh saved. Data collected to support these calculations can include billing data and metering data as well as inputs to engineering or statistical models. Data of this type is often the highest priority because of its relationship to utility rates, profits, and/or cost recovery.

As was pointed out in the previous chapter, external validity is important for generalizing conclusions. To perform such comparisons and projections, additional data (such as customer location, market-segment information, customer preferences, weather, and economic variables) are needed. If these data are collected, then impacts may be normalized, making them transferable, which enhances the likelihood that meaningful comparisons can be made with similar programs in other service territories. The collection of this type of data also makes it possible to develop predictive models, which can be used to estimate program performance for different service territories, customer segments, and economic conditions. Being able to posit causal relationships opens the possibility of projecting impacts for programs with different designs and of reassessing program impacts in response to changes in avoided costs, the economy, or utility or regulatory priorities. It also supports program redesign to improve future performance.

The previous chapter pointed out that the demand or energy savings are the differences between what would have been consumed without the program and what is actually consumed with the program. The data requirements for DSM-impact evaluation are

- The usage and demand of the facility prior to measure installation and changes in operating characteristics of the end-use or facility
- The daily or annual operating profile of the energy-consuming systems affected by the measure
- The operation and maintenance (if applicable) of the measure and the life of the measure (persistence), including estimates of the decline in efficiency of the measure and/or its removal

- The data to estimate what would have been had no program been offered
- The net change in power or energy requirements resulting from the measure

Because measurements and estimates contribute to the savings calculation, measurement error in any component will be carried through to the assessment of the impact. Making measurements that are rigorous for one or two elements will not result in an accurate estimate of impacts if the remaining parameters are not measured or are measured in a limited way. Accuracy of the total evaluation is limited by the accuracy of the weakest measurement. Therefore, the ideal approach is to develop a balanced data-collection strategy, where variance for each measurement is about equal.

The requirements for data collection are determined primarily by the analysis method chosen (Table 8). There are three general types of analysis methods and associated data:

- Direct measurements, which are used to calculate changes in energy use by comparing measurements made at different times. Direct measurements include customer billing, whole-building metering, end-use metering, and frequency meters.
- Engineering modeling, which uses physical models ranging from simple tracking models to complex physical simulations of buildings to analyze energy use. Data used in these models include weather, customer surveys, facility and equipment inventories, and operating patterns.
- Statistical modeling, which uses statistical techniques to evaluate changes. Data that are often used in these types of models include billing data, market-segment information, demographic characteristics, and economic variables, such as electricity price.

These techniques are often used in combination. Billing data are often combined with customer data for analysis. For example, a regression of monthly energy use against weather data can be used to detect the presence of heating, cooling, and nonweather-sensitive uses as well as to determine individual use levels. Conditional-demand analysis uses direct measurement and statistical regression (Chapter 4). Hybrid statistical/engineering method (HSEM) refers to the use of engineering and direct-measurement data to drive engineering based statistical regressions.

Based on the objectives and analytic methods, the evaluator can define procedures and specifications for data collection. These specifications should include the source, the type (e.g., definition of units and tolerances), sample sizes, frequency of data collection, and quality-control requirements.

Table 8. Data options by analysis method

Data source	Analysis method			Combinations		
	Direct measurement	Engineering model	Statistical model	Billing analysis	Conditional demand	HSEM
End-use metering	■					
Whole-building metering	■				□	□
Billing data	■			□	□	□
Facility inspection		□				□
Mail survey			■		□	
Telephone survey			■		□	
Onsite survey			■		□	
Weather data		□		□	□	□

■ Data source sufficient for analysis method.

□ Additional data source required.

TYPES OF INFORMATION FOR IMPACT EVALUATIONS

The variety of data sources includes facility energy-use measurements, facility inspections, customer-supplied data, research to establish typical values, and combinations of data sources. Determining which source to use depends on the evaluation objective, analytic method, costs of data collection and analysis, and accuracy. There is no best or worst source; each has strengths and weaknesses (Table 9).

Facility Energy-Use Measurement

Facility energy-use measurements can come from customer electric and fuel bills, hourly consumption from whole-building meters, and end-use metering. These are direct measurements and should be accurate within the tolerances of the metering device. The evaluator needs to remember that these are measurements of present use and not changes in use; they have no explanatory power in and of themselves.

End-use metering is expensive but has the advantage of providing a direct measurement of load for the device or circuit of interest. With advances in solid-state technology and experience, the costs of end-use metering are dropping. Recent costs for the metering device, installation, data-acquisition system, and maintenance range from \$5,000 to \$20,000

(in 1991 dollars) per building per year; see also Table 22 in Chapter 7 and Exhibit 4. Several projects conducted in the past five years have sampled three to eight loads in each of 25 to 100 buildings.

Table 9. Characteristics of data sources

Data source	Strength	Weakness	Cost
1. Facility energy-use measurement	Accurate to bounds of the meter	Measures present use not change parameter; no explanatory power	Varies, see below
• End-use metering	Highest unit accuracy of the target system	Most expensive per unit; budget limitations result in high sampling error	\$5,000 to \$20,000 per building
• Whole-building metering	Measures short-internal whole-building loads that can indicate target-system loads	Lower unit accuracy than end-use metering	\$500 to \$3,000 per building
• Customer billing data	Census eliminates sampling error	Difficult to detect change in treated system	\$2 to \$10 per building
2. Facility inspection	Collects causal technical factors best	Time limitations at site cause many values to be estimated	\$20,000 to \$40,000 fixed costs plus \$100 to \$600 per facility
3. Information supplied by customer	Collects customer market data and attitudes best	Error rates high on technical data	See below
• Mail survey	Low cost permits large sample	High error rate and risk of nonresponse bias	\$10,000 to \$20,000 fixed costs plus \$15 to \$40 per survey
• Telephone survey	Higher response rate and lower error rate than mail surveys	Higher cost than mail surveys	\$15,000 to \$30,000 fixed costs plus \$25 to \$60 per survey
• In-person interview	Lowest bias and error on customer questions	Higher cost	\$15,000 to \$30,000 fixed costs plus \$150 to \$300 per site

Exhibit 4. The costs of multichannel, automated, end-use monitoring

Components of this cost are highly variable, but recent ranges for a full-scale data-acquisition system with a central personal computer (PC), remote meters/recorders, and communication system are as follows.

- Meter and wiring costs depend on the number of channels, how the monitored data is communicated to the building recorder, and the difficulty of isolating the monitoring point from the rest of the electrical system. Costs for a building recorder and meter transponders are \$500 to \$2,000 per facility while specification and installation add an additional \$500 (residential) to \$3,000 (complex commercial) per facility.
- Data-acquisition system set-up costs range from \$20,000 to \$100,000. In addition, the annual costs per facility for data acquisition and management are \$500 to \$1,000.
- In total, end-use-metering fixed costs range from \$20,000 to \$100,000 for the first year plus \$1,500 to \$6,000 per facility annually.
- Analysis of the data bears additional costs.

Barring technical problems with the wiring, the meter, or the data-acquisition system, the electronic end-use meter provides a high-resolution measurement (frequent recording of values), which is often the standard for accuracy to which other measurements are compared. End-use metering is the only measurement technology that combines high resolution, accuracy, and isolation of the energy-consuming system. As a selective research tool, it can be applied to particular questions that other data-collection forms fail to answer satisfactorily. For example, end-user monitoring may be the only technology that can be used to accurately resolve the effect of lighting efficiency on cooling-energy use, an important issue in lighting programs.

Because of their cost, end-use meters tend to be used with small samples, which means that extrapolations to the total population produce a fairly large error. Several studies, notably the evaluation plans of Northeast Utilities and Empire State Energy Research Corp., have found that other methods may produce more-precise population estimates because they use larger sample sizes at lower costs (Townsend and Wright 1990; XENERGY 1990). Another possible problem with end-use metering is that the customer may be aware of the monitoring and may change behavior (the Hawthorne effect).

In addition to the multichannel, automated, end-use metering described above, other types of end-use monitoring exist, many of which have lower costs. These methods include instantaneous metering, portable/temporary electronic metering, and nonelectronic portable/temporary metering.

Instantaneously recorded amperage, voltage, wattage, power factor, frequency, etc. on a single appliance can be used to establish instantaneous load conditions, such as the kW for air conditioners, lighting fixtures, motors, heating elements, etc. A variety of handheld and portable meters are available to measure these values at costs ranging from less than \$100 to \$5,000. These devices are useful for providing field data to validate engineering estimates. In the case of a pure unswitched lighting circuit, this type of monitoring device might be used to evaluate net change in load by determining connected load before and after high-efficiency lighting is installed. In this example, care needs to be taken to account for burned-out bulbs and fixtures. Generally these meters are not useful for gathering data where the interest is in daily, weekly, or seasonal variations.

Portable/temporary meters can function over a longer period, ranging from several days to several months. Usually these systems take measurements, store the measurements in a recorder, and subsequently upload the data to a PC for analysis. These systems cost between \$500 and \$700 per recorder channel and come with meter/recorder hardware as well as PC software for analysis, graphics, and reporting. With portable/temporary meters, daily, weekly, or seasonal variations can be captured without communication lines.

Portable nonelectronic meters are similar to electronic meters except that the data are collected on circular/strip charts, cumulative value registers, or runtime meters. Utilities often overlook the fact that the simplest and cheapest form of end-use metering can be a standard watt-hour meter, which can be connected to an end-use load. This works as long as the measurement of interest is cumulative kWh and some provision can be made for reading the meters at the desired frequency.

Whole-building load-meters are used to record building energy use at short intervals, such as each 15 or 30 minutes but sometimes as frequently as each minute. These meters are less expensive than end-use meters and are often installed without the knowledge or consent of the customer. The cost of the meters is similar to that of end-use meters, but installation (usually \$200 to \$500), the data-acquisition system, and annual meter management are less expensive. Because many utilities already use such meters for load research and routine metering of large customers, the fixed costs may be reduced. Depending on which of these circumstances apply, total load metering can cost \$500 to \$3,000 per facility.

While providing an accurate measurement of total load, whole-building meters are at a disadvantage compared with end-use metering for measuring subsystem loads. They do not provide a direct measure of the treated system. If the treated system is a large fraction of the total load, the change in load from the treatment is large, and the frequency of measurement is high, accurate estimates of end-use loads and changes in end-use loads can be extracted. For example, HVAC loads, which are responsive to weather, are often a statistically significant component of the load and can therefore be estimated with whole-building meters.

The lower cost of whole-building meters permits larger samples and may increase population accuracy for a fixed evaluation budget. Whole-building meters are being assessed as a primary data source for HSEM. This form of analysis, which also requires a facility

inspection, is being investigated for comparative bias and accuracy to end-use metering and other methods (XENERGY 1990).

A powerful extension of the whole-building metering concept is the appliance signature or nonintrusive meter, which records variations in real and reactive load components. As appliances turn on and off, the signature of these components is identified and recorded electronically. This concept has been tested in residential applications, and it explains up to 30% of the load (EPRI 1989).

Standard customer billing data are available for all participants and nonparticipants. Those data can provide such information as monthly or bimonthly kWh consumption, monthly peak demand for large commercial accounts, and (occasionally) time-of-use measurement. Because these data are collected by utility billing operations and are generally maintained on customer-information databases, the acquisition costs are very low. The ability to use a census gives the analysis of this data the potential for eliminating extrapolation error. The disadvantage is that the limited information on each customer makes the detection of changes in consumption for individual systems difficult. The energy consumption of loads that are large and fluctuate with the season can be differentiated. The Princeton Scorekeeping Method (PRISM; see Chapter 4) uses billing and weather data along with a statistical heuristic to produce end-use-specific estimates of energy use.

Facility Inspection

Facility inspections collect physically observable data, such as nameplate data and information about observable operating conditions, through energy audits and site surveys (Fig. 7). These data permit independent verification of customer- or implementor-supplied information and provide greater detail and consistency in the collection of observable parameters. Facility inspection data often are used for the detailed inputs to engineering models.

The reliability and accuracy of predicted energy consumption and hourly load estimates from engineering models is limited by the difficulties of collecting the data that serve as model inputs. Many of the data requirements for engineering models are related to the behavior of the building occupants. Accurately describing occupant behavior requires extensive observations, so it is usually estimated based on a brief facility inspection and reported behaviors. For example, the occupancy profiles and operating schedules of equipment may be observed during a facility inspection, but data about off-peak occupancy and operating schedules may be based on the customer's statements and perceptions.

The amount of time available for inspection may limit the number of measurements at a site with attendant reductions in the overall quality of facility-inspection data. For example, in inspections of large or multibuilding facilities, building attributes such as envelope thermal characteristics, equipment efficiencies, lighting intensities, ventilation rates, etc. are generally sampled or estimated to reduce costs. Then the results are extrapolated to the whole facility.

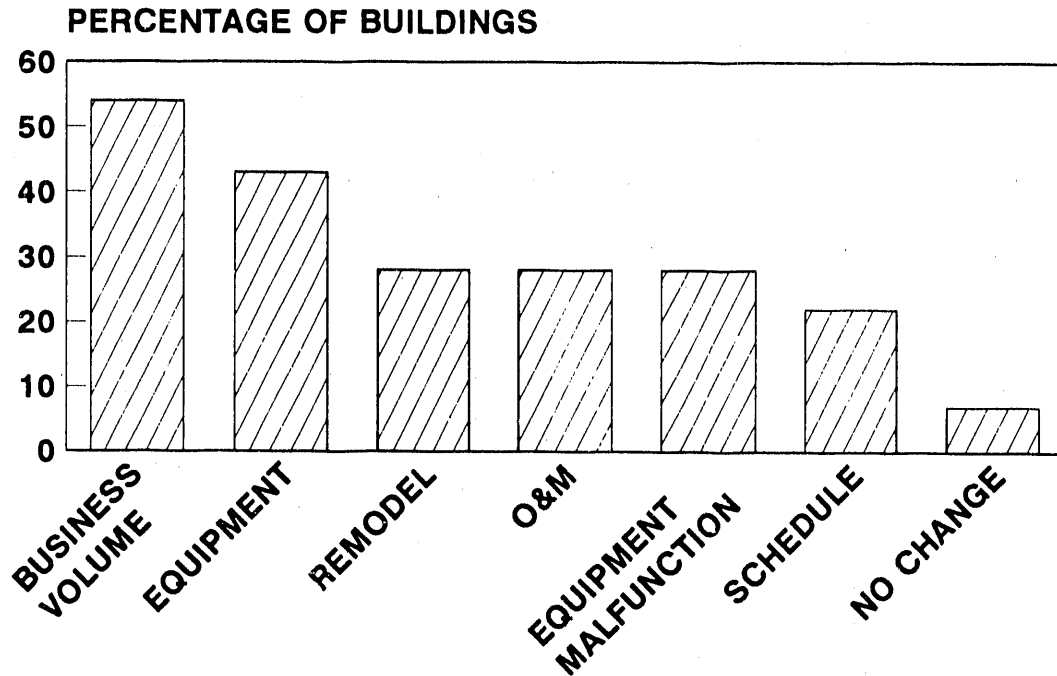


Fig. 7. BPA conducted onsite inspections of 46 buildings one to six years after the facilities had participated in its Institutional Buildings Program. These inspections found many changes that probably affect postretrofit electricity use.

A brief facility inspection does permit collection of most of the technical data required for modeling a building. By limiting observation, homes and small commercial facilities can be inspected in as little as 30 minutes, medium to large commercial facilities in 1 to 4 hours, and very large commercial or industrial facilities in a day or less. By way of contrast, a complete engineering inspection that includes measurement and/or observation of all significant engineering factors can take two to five times as long.

The total cost of an engineering facility inspection includes scheduling and travel time, data preparation, data entry, and quality-control procedures. Typical fixed costs are \$20,000 to \$40,000 to prepare the survey instrument and protocols, recruit and train inspectors, field test the instrument and survey process, and develop quality-control procedures. It is difficult to complete more than two surveys per day even if the average inspection time is only an hour. Because more than half of the typical inspector's time is associated with travel, schedule limitations, waiting, and canceled appointments, considerable savings are associated with inspections conducted during visits to the customer for other purposes. Unit costs are \$100 to \$300 if the inspection is part of an otherwise scheduled visit or \$250 to \$600 if conducted separately.

Customer-Supplied Information

Information is often supplied by the customer in response to a survey rather than measured or observed. Information to determine the customer's market segment, equipment-operating patterns, actions, and attitudes can be collected as part of the application for program participation or by response to mail, telephone, or in-person interviews.

Using customer-supplied data has advantages. Data from this source are inexpensive compared with data from other sources. And this is the only method for gathering direct input on customer attitudes, opinions, and preferences.

However, customer-supplied data has some disadvantages. Customers cannot be relied upon to supply detailed technical data. Customers may erroneously report data about the physical and operating characteristics of a facility or appliance. Customers may not know the answers. They may not be motivated to determine the answer. Or, they may give false or misleading answers. It is not uncommon for customers to confuse a hydronic heating system with the domestic hot water system (they are both water-heating systems). A customer is unlikely to know the loading on motors. Customers cannot be expected to respond to burdensome questions, such as asking a business owner to report the nameplate sizing of AC units or the number of four-lamp fixtures that currently have two lamps. Reports may be biased by the desire to report a "tight ship," for example, by indicating that the lights are off when the facility is unoccupied. Appraising customers' willingness to purchase equipment or determine future operating characteristics of equipment based on questionnaires often fails because customers' attitudes do not predict their behavior (Converse and Presser 1986).

The three basic methods of soliciting customer information are the mail survey, the telephone interview, and the onsite interview. Mail surveys are inexpensive, costing \$15 to \$40 per completed survey, including a computer-generated report. The size of the sampled population is the greatest determinant of costs. Survey design and protocol development are typically \$10,000 to \$20,000. Mail surveys often have low response rates and high error rates. Response rates for mail surveys can be 5 to 10% in commercial surveys and 20 to 40% for residential surveys. High error rates result from a lack of incentive to complete the survey, carelessness, ambiguity, and lack of supervision of the respondent. The problems can be mitigated by improved packaging of the questionnaire, better timing, providing a clear and nonjudgmental set of questions, and providing an incentive to respond (Dillman 1978; Train 1988).

Telephone surveys cost more than mail surveys. Survey design and protocol development tend to be \$5,000 to \$10,000 higher than for mail surveys. Conducting the survey costs \$25 to \$60 per completed questionnaire. Telephone surveys generally have higher response rates and lower error rates than mail surveys. However, with the increased use of telemarketing by businesses, the public is becoming increasingly wary of unsolicited calls, and researchers are reporting difficulties in getting an adequate sample. Telephone response rates can be improved by sending a letter in advance of the call explaining the purpose of the call in general terms.

Respondents to a telephone survey cannot be expected to provide specific physical data about equipment and often can provide only a very general indication of the types of appliances they have. For example, residential customers may not be able to distinguish between heat pumps and strip heaters unless survey questions are carefully worded. Telephone surveys can be worded so that they provide an indication of types of equipment, satisfaction with performance, type of facility, operating hours, and attitudes (Dillman 1978).

An in-person interview allows for more depth with the obvious trade-off against cost. If conducted in isolation, interviews cost \$150 to \$300 per interview. If facility inspections are conducted for physical data, the marginal cost of an interview is usually small, perhaps an additional 30 minutes of site time. The costs associated with survey design and protocol development are similar to those for telephone surveys. Unstructured interviews can be very useful in eliciting an understanding of a customer's or project manager's experiences, especially in the initial stages of a project. However, unstructured interviews are difficult to analyze. The analysis of unstructured interviews requires the use of content-analysis techniques, and the results are difficult to generalize. Open-ended questions in structured interviews have many of the same problems. Close-ended questions have the virtue of asking the same question and providing the same response set to each respondent. This makes the responses easier to analyze and interpret, but if the questions are not well designed, the information obtained can be limited.

Research to Establish Typical Values

Bench testing of equipment or borrowing data from other utilities are two ways that utilities can establish typical values. Many impact characteristics can be established by bench testing based on data provided by manufacturers or independent laboratories. A laboratory measurement of unit savings in combination with reported information on installation, operating patterns, and prior conditions can be used as inputs to engineering models.

This approach is not without risks. The conditions under which a manufacturer's measurements are made may not be typical of conditions in the field. Further, conditions in the field may interact with the equipment to change the operating characteristics. The operating patterns used in tests may be based on extreme patterns or those that are assumed to be typical but have not been validated.

Facility operating patterns, baseline characteristics, customer maintenance behaviors, and persistence may be available from a similar utility or may be predicted by examining nonenergy products with similar economics for the customer. A literature review provides an important starting point for predicting typical values.

Collecting Local-Weather, Equipment, and Customer-Characteristics Data

Economic, demographic, and weather data are almost always essential for establishing a baseline. Sources for these types of information include the population census; the housing census; special censuses, such as those for C&I establishments; economic statistics; and EIA's

Commercial Buildings Energy Consumption Surveys. Private databases on construction, space inventories, and equipment sales also may be useful. A utility may maintain records of new construction notifications or permits (Weiss 1972).

Combinations of Data Sources

Evaluations often use combinations of data sources to measure impacts (as discussed in the section on Selected Special Issues in Chapter 4). This practice has significant implications for data-collection procedures. Triangulation is the term used to describe the parallel use of multiple data sources. A common approach is to use billing analysis for the population or a large sample, end-use metering for a small sample, and a field survey and engineering models for a medium-sized sample. This may result in three different impact outcomes. A triangulation strategy uses combinations of data to improve the overall estimate. For example, the metered sample can be used to determine the bias in the engineering approach. It can also, using Bayesian mathematics, be used to reduce the variance in the billing analysis.

Samples used in triangulation can be independent of each other although there may be some overlap. Some flexibility in the timeframes of the samples may also occur. Nonetheless, some coordination is required, including

- Developing an internal communication system to reduce confusion within the utility and with the customers
- Identifying customers appearing in more than one sample
- Developing a protocol for customers who will be contacted more than once
- Developing combined systems for sampling, monitoring management, and quality control to take advantage of economies of scale
- Designing evaluation databases to simplify data comparisons and analyses

The use of models to explain the energy savings may require data leveraging. In this situation, facility inspections or survey data are complemented with billing data, whole-building metering data, or end-use data. These data are then used to develop an explanatory regression model. Examples include billing analysis, conditional demand analysis, and HSEM (Schön and Hamilton 1990).

To perform these types of analysis, various data sets must be linked. This linking is a more stringent requirement than for triangulation methods, where measurement and survey samples may be independent. Linked data requires protocols that ensure a match among facility, end uses, and metered data. This matching can be difficult if some end-uses are included in a master meter and some in a customer's meter, the boundaries for floorspace and energy systems do not match (for example, a store within a mall), multiple meters are used for a facility, or data for multiple fuels are required.

When floorspace and meters do not match, the correction procedure varies depending on whether it is the customer's business boundary, the meter boundary, or the building boundary that is of interest. Which boundary to use is an issue for the analysis method and the method chosen to extrapolate the results. Depending on the choice, the data-collection process may require adding meters or surveying space to create a match. This additional effort increases the costs or time to investigate the match, entails additional data collection, and requires expanded quality control. The best assurance of a reasonable match is to estimate the building energy use from the survey data provided. If the estimated use and the measured use are not within a reasonable tolerance, the survey can be set aside for investigation.

Evaluators may perform a succession of analyses and commit more resources to improve sample designs while minimizing overall costs of an evaluation. For example, an evaluator might do a billing analysis to identify residences that have and use AC; a facility inspection to identify residences with central air conditioners and room air conditioners and to identify the structural characteristics of the residence; an energy audit, a survey, and whole-house monitoring for a large sample of the customers with central AC; and ultimately end-use metering on a selected subsample of the whole-house-metered group. The sample design may be improved at each step.

A good data-management plan is required to support the ongoing data-collection effort as well as the interim analytic steps. Near real-time input and quality control of data are required as well as a database structure that can easily be prepared for analysis while the surveys are still being conducted.

SPECIFYING THE INFORMATION TO BE COLLECTED

Specifying Facility Energy-Use Data

End-use metering requires the specification of the circuits within a facility that are to be metered. The end use must be carefully defined. For example, ventilation is difficult to separate from other elements of heating and AC. The pumps and motors associated with distribution of heated or chilled water may be defined as a heating, cooling, or miscellaneous end use.

End uses rarely are perfectly aligned with electrical circuits. Lighting circuits often include outlets with nonlighting loads. In this case, the evaluator may collect data to account for the plug loads, sample lighting-only circuits, or isolate lighting by rewiring. Obviously, the last method is the most expensive and least practical.

Not all end uses of a single kind in a building may be treated. For example, electronic ballasts may not have been installed in every lighting fixture. It may be difficult to identify and isolate fixtures where the ballasts were installed. The evaluator may have to measure treated as well as untreated lighting. While this captures the savings, it also introduces measurement error.

Measures may have secondary impacts. For example, a reduction in the energy used for lighting reduces cooling requirements in the summer and increases heating requirements in the winter. Both the primary and secondary effects of a measure must be considered. In the lighting example, a well designed data-collection strategy includes measurements to determine HVAC effects. This measurement may be difficult if both lighting and HVAC measures are installed in a facility with a single meter because a single meter can only measure the combined energy use.

Whole-building-metering procedures are well established in most utilities. When whole-building metering is used for evaluation, however, the ambiguities associated with accounts, customers, and buildings need to be addressed in the protocol for meter placement. Congruence is required among the metered space, the end use, and the area for which survey data are collected.

Billing analysis has a number of data requirements, especially when billing data are matched with weather data. Any useful customer-segmentation information on the billing record (such as SIC code, rate class, or owner-renter designation) should be retained. If multiple years of billing data are available, then it is preferable to collect and use all of these data to identify trends. Estimated data in the billing record should be identified as such because it does not make sense to do regression analysis on data that have been constructed. Meter-reading dates are especially important for weather normalization. If weather is not a factor, recording data by month is sufficient. The measured peak kW should be gathered if it is different from billed demand.

Data Specification for Facility Inspections and Customer Surveys

The method of analysis directly affects survey data requirements (Table 10). A bill analysis may not require survey data, while end-use metering may need a prior determination of electrical-system layout. Engineering models are driven by data on physical equipment; such data are obtained through facility surveys. Inputs to statistical models, such as market segment, demographic, or economic data, are usually customer reported. Mail or telephone surveys may suffice, although accuracy and response rate may be increased by an in-person interview. A hybrid model needs physical data and interview information.

Specifying Data for Transferability

Generalizability does not just happen. Data collection must be designed to support the transfer of results to other programs and other utilities where the characteristics of the program and the target audiences may be different; see the discussion of external validity in Chapter 2.

Opportunities for transfer are enhanced when care is taken to ensure that there is a statistically adequate representation of respondents or facilities by space, time, market segments, weather, and economic variables. In selecting segments and the sizes for experimental programs, the population mix of the regions to which the results are to be

Table 10. Data requirements by data-collection method and evaluation approach

Data requirements	Data-collection method			Evaluation approach			
	Mail or phone survey	Onsite survey	Secondary sources	End-use metering	Engineering model	Statistical model	Hybrid model
End-use equipment							
Inventory checklist	<input type="checkbox"/>	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Size, age, and efficiency	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Electrical-system layout		<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>			
Building structure/segment							
Building type	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Floor space	<input type="checkbox"/>	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Envelope (R-values, etc.)		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>
HVAC zones		<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>		<input type="checkbox"/>
Facility operations							
Business hours	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Occupancy	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>		<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Equipment schedules	<input type="checkbox"/>			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	
Weather data							
			<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
Demographic data							
	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>			<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

Required
 Optional
 Reliable data
 Data available but unreliable

transferred, as well as the population mix where the program is to be implemented, should be considered. Where statistically adequate independent samples are not possible, models and data-leveraging techniques may be used to extrapolate from the surveyed segments to market segments with no sample or an inadequate sample.

If the goal is to develop a market-acceptance model, samples should be heterogeneous with respect to ownership status, age, income, attitudes toward utilities and conservation, and economic climate. Variations should also be incorporated into incentives and marketing strategies (e.g., bill stuffers, telephone contacts, etc.). For a variety of practical reasons, a single utility may not be able to offer the required range of incentives and marketing strategies, but a group of utilities may be able to do so. A group of utilities can then pool data to increase variation in technical and market parameters, and can use different incentive and marketing strategies in different service areas. This strategy will work only if great care is taken to gather the same information with the same methods and the same procedures in each area.

DESIGNING THE SAMPLE

Sampling is often the least understood and most poorly executed part of an evaluation. The basic steps in developing a sample design are as follows.

- The population of customers to be represented by the sample, the sample frame, must be defined. Examples of typical frames are all residential customers, small commercial customers with demands less than 40 kW, all residential customers with central AC, or all C&I customers with more than 15 kW of connected fluorescent-lighting load.
- If necessary, the sampling frame is segmented into homogeneous groups on the basis of some characteristic(s), such as kW, kWh, and/or SIC code. SIC code is often used, but many utilities have found it unhelpful because of the wide variation in the size of firms represented within an SIC code.
- The size of the impact variable (e.g., Δ kWh or Δ kW) is estimated for the sample frame as a whole or for the segments. An estimate of savings may be based on engineering estimates or savings reported from other localities.
- Finally, for the population or for the segments, the variance of the impact variable (i.e., savings) must be estimated as well as the variance associated with one's ability to measure the variable. This measurement is usually accomplished by examining a customer database that can be used as a proxy for measured savings (Exhibit 5). Commonly, these proxies are the estimated savings from applications or audits or the variance in customer kWh or kW. Measurement variance and bias are often discounted as unmeasurable or insignificant or as a nonquantified limitation to the sample design. This assumption may not be a good one.

The population size (which is known), the estimated average impact, and the variance of the impact (which is estimated) provide the basis for the calculation of sample size requirements to meet the desired sampling error, specified as a precision (such as $\pm 10\%$) with a specified level of confidence (such as 90%).

Trading Off Sources of Error

Sources of error are manifold. If precise measurement of savings were possible, then sampling (as compared with conducting a census) would be the only cause of error. However, measurements are also subject to error. Much of the data collected in an evaluation are treated as observed data when in fact they contain measurement error. For example, operating hours for a commercial building are observed during an onsite survey but may be reported for off-peak periods. These off-peak data are treated as measurements but contain an unknown amount of error. Reported evening operating hours in commercial buildings are often lower than true levels.

Exhibit 5. Calculating sample sizes for an evaluation

Utility A plans to run a weatherization program that is similar to that of Utility B. Utility B had the following results when it conducted its weatherization program.

Table 11. Mean consumption or savings (kWh/year) and standard deviation (SD)

	Mean consumption or savings (kWh/year)	Standard deviation
Preprogram electricity use	26,500	8,480
Postprogram electricity use	21,480	8,130
Savings	5,020	4,270

Utility A wants to know how large a sample it needs to determine the savings for its program. Because the programs are nearly identical, Utility A assumes that the savings from its program will be similar to the savings from Utility B's program. It therefore decides that it can use the information from Utility B's program to determine the sample size for its program based on the following formula:

$$n = (t^2s^2N)/(E^2N + t^2s^2) ,$$

where n = sample size; t = the t value associated with a sample size and level of confidence taken from a t-table in a statistics textbook; s = estimated SD; N = the population size; and E = required level of accuracy.

Initially, the evaluator assumes that the sample size should be based on the postprogram consumption (mean = 21,480 kWh). She then realizes that the postprogram consumption was not of interest but rather the savings and the SD for the savings. She also realizes that management will ask questions about the sample size, so she constructs the following table to help management determine how to best use evaluation resources.

She assumes three levels of confidence (0.1, 0.05, and 0.01), three levels of accuracy (10%, 5%, 1%), and three levels of population (10,000, 20,000, 100,000). The results are displayed in the following table that shows sample sizes as a function of confidence level and accuracy:

Exhibit 5. (Continued)

Table 12. Sample sizes as a function of confidence level and accuracy

Confidence	Accuracy		
	10%	5%	1%
Population = 10,000			
0.1	117	453	5,349
0.05	192	725	6,544
0.01	376	1,352	7,911
Population = 20,000			
0.1	118	464	
0.05	194	753	
0.01	384	1,450	
Population = 100,000			
0.1	119	472	
0.05	195	775	
0.01	389	1,539	

The computation for the first cell is:

$$n = 1.282^2 4270^2 10000 / (502^2 10000 + 1.282^2 4270^2)$$

Several things are clear from this table.

- The size of the population has little effect on sample size.
- The required sample size for 1% accuracy is so large as to be impractical to implement.
- The trade-offs between accuracy and confidence level are about equal.

Other errors occur during data handling. Data are often mistakenly entered or misread. While good quality-control procedures can reduce data errors, they will not eliminate them.

Measurement accuracy and sample size can be traded off to minimize sources of error within a fixed evaluation budget. For example, end-use metering is accurate but expensive, and samples are usually kept small to reduce costs. The consequence is a large sampling error. Within the same budget, sampling error might be reduced by using a large sample in conjunction with load meters, field inspections, and hybrid statistical engineering models that have a lower per unit cost. Progressively larger samples can be afforded when field inspection is used with engineering estimation, mail surveys are used with statistical

modeling, and bill analysis is used with a census. These larger samples have progressively lower sampling errors but progressively larger unit measurement errors.

Defining Precision Requirements Sensibly

Defining precision requirements is a critical element in determining sample sizes. In making this determination, one must consider the importance of the measurement relative to evaluation objectives. For example, an estimate of lighting electricity savings may require measurement of peak kW savings and operating hours. The peak impact of a high-efficiency lamp can be determined with a fair amount of certainty. The midday operating pattern may be reasonably predictable, but the evening operating pattern may be much more variable. For a utility with an evening peak, the precise measurement of evening operating hours would be more important than it would be for a utility with an afternoon peak. A larger sample size may be required for the evening-peaking utility.

In a recent study, building data were collected for a load-research sample. As part of the sample-planning process, statistical, engineering, and hybrid models were examined for accuracy and cost in relation to metering options. The study also assessed the relative importance of individual measurements to the impact variables. Survey data from a large sample were determined to be lowest in cost but were found to have a relatively large uncertainty. Hybrid approaches applied to a subsample of critical building types and measures substantially reduced the potential bias. Where uncertainty and importance remained, end-use metering appeared to be the best option (XENERGY 1990).

ENSURING DATA QUALITY

Support, Training, and Management of Survey Takers

When possible, experienced staff should be used for field and telephone surveys. When this is not possible, evaluators should recruit people who have a basic understanding of energy and who know how to inventory and evaluate building envelope and energy systems. In addition, survey takers should be familiar with the programs, processes, and applications referred to in the survey. Equally important are people with social skills.

Support materials for survey takers should include, but not be limited to, a program description and introduction, line-by-line instructions for the survey instrument, technical support materials on equipment and devices, a protocol for dealing with issues that may arise in the field with particular attention to sampling, a set of procedures for scheduling conflicts, and account-contact guidelines.

Training survey takers is one of the most important aspects of a quality-control program. Even if the survey takers have extensive energy-survey experience, a training session is essential to instruct them on the specifics of the particular project and to brief them on the issues that they may encounter in the field. The training session, in conjunction with the survey training manual, helps to ensure that each survey taker interprets the questions in the

same fashion, thereby providing the required consistency. Survey takers should practice on each other and perhaps on a group of volunteers before going into the field.

A supervisor should observe survey takers from time to time to be sure that they continue to ask questions appropriately and follow procedures. Feedback about survey-taker performance should also be solicited from a randomly selected group of those being visited in the field. Questionnaires should be edited immediately after they are returned, and any patterns of omission or error should be called to the survey taker's attention. Finally, supervisors should review information on the face sheets of surveys (e.g., the amount of time spent conducting the interview) to see if these data give indications of interviews being too short or too long or the timing of interviews being inconsistent with estimated travel requirements.

Data Quality Control

Despite training, errors will occur in reporting and in data entry. Several methods of prevention are available. Double entry of data minimizes input errors. Range checks can catch reporting and input errors. More sophisticated validity checks, which compare data with a set of predetermined parameters (such as watts per square foot, occupants per square foot, watts per fixture, etc.) can be programmed into software. Finally, consistency checks can verify predetermined relationships between survey questions. For example, a centrifugal chiller system in a 10,000-ft² building is unusual, suggesting the need for further investigation.

Survey data errors can be identified and reduced by offering a computer-generated energy analysis to the customer. This is a particularly useful practice in mail surveys. This inducement encourages the customer to exercise care and provides an additional error check by virtue of the customer's review of computed results based on the data he/she provided. In addition to reduced error rates, this offer may increase residential mail survey response rates as much as 45%, reducing self-selection bias.

Quality Control of Experimental Procedures

Logs should be maintained for all customers being recruited, and the status of all customers participating in a survey. Tracking the status serves as a record to ensure a survey's statistical value, as a management tool to evaluate the effectiveness of survey resources, and as an aid to the utility in dealing with customer issues.

The data-collection manager should conduct ongoing analytic activities to identify potential problems with a survey. Quality-control procedures include testing variance assumptions, correcting survey problems, and adjusting analytic methods.

The research plan is designed to measure impacts to a desired precision and confidence level. The estimate of sample size required to meet this level includes variance assumptions for the sample population. Because of the uncertainties about the actual variance, sample sizes may be too high or too low. An ongoing analysis of the variance of key variables during data collection might identify the need to adjust sample size. The variance data can

be examined at any time. However, the evaluator should remember that, while samples are randomly drawn, they are usually reorganized on the basis of geographic proximity to facilitate recruitment, and proximity reduces heterogeneity. This concern needs to be kept in mind when estimating variances.

Every survey should be pretested, and the results of the pretest carefully analyzed. A minimum of three or four pretests is usually required. Despite pretesting, problems may arise in the field. The respondent may have a problem recalling the wattage of incandescent lamps replaced by compact fluorescent, remembering whether four-lamp fluorescent fixtures had any lamps removed prior to installing a reflector, or reporting the operating pattern of lights during evening cleaning. Such problems should be identified as they occur, and special methods of acquiring the information (such as proxy questions, special inspections, and spot metering) should be developed to correct bias during subsequent analysis.

To protect the experimental design, the original data-collection method and survey should continue. Changing the wording of questions or technique may alter the survey, and care should be taken to minimize this. For example, if the responses to early surveys asking for the night lighting schedule are biased, the question should be retained so that the bias correction can be applied uniformly. Care should be taken to collect the same information for all respondents, even those who completed the survey prior to the discovery of a problem.

Whole procedures may need to be evaluated. An engineering HVAC model may require input of ventilation rates in cubic feet per minute. This information is often not available in the field and cannot be reported by the building management. Reliable data may be available only for floorspace cooled and tonnage of equipment. In such a case, it may be necessary to use spot metering and hybrid modeling to meet the desired evaluation goal.

CONCLUSIONS

Monitoring is not conducted in a vacuum, it is intertwined with evaluation planning and analysis. The selection of a data-collection method, data specification, and sample design are key decisions that determine the cost and the quality of analyses that can be performed. Whenever possible, data should be collected to permit transfer of results, make prediction possible, and allow for the assessment of causality.

Not all data sources suit all evaluation objectives and analytic methods equally well. Selecting data sources is a pivotal specification that can dramatically impact budget requirements. No one data source is best; each has strengths and weaknesses. Sampling and sample sizes are also important problems. Measurement accuracy and sample size can be traded off to optimize precision and cost.

Data-quality issues abound. Data quality starts with careful collection of data by the use of capable survey takers, providing quality survey support materials, survey-taker training, and survey-taker management. Data-validation checks are another way of ensuring data quality.

Finally, monitoring and analysis while data are being gathered can significantly reduce data-quality problems.

ANALYZING DATA

Daniel Violette

INTRODUCTION

This chapter discusses data-analysis methods for estimating the impact of DSM programs on kW or kWh consumption. For any given program, several candidate analysis techniques will exist, each with its own strengths and weaknesses. The selection of an appropriate method will depend upon:

- Specific program design
- Available prior information on program impacts
- Program's priority in the utility's DSM portfolio
- Implementation stage (whether it is a pilot, new, or mature program)
- Objectives of the evaluation
- The technologies and customer segments targeted by the program

Evaluations can be based predominately on engineering methods, billing-data analyses, metered data, or combinations of these methods. While most data analyses can be classified as falling into one of these categories, the use of multiple methods is a trend. These approaches go beyond conducting parallel engineering and billing-data analyses. Instead, they often mean leveraging and integrating data across more than one approach. This technique could involve using estimates from an engineering study as inputs to statistical models, or it could involve using relationships derived from a statistical analysis of consumption data within an engineering model.

An impact evaluation has two purposes, to measure the *change* in participant electricity use and to identify the factors that explain the observed change, especially the effect of the utility DSM program. The challenge for evaluators is to determine how this intervention in the market for energy services influenced the energy use of participants and, often, the energy use of nonparticipating customers as well.

Because the purpose of an impact evaluation is to measure changes in energy use, a baseline must be identified from which a change can be measured or estimated. Determining this baseline is a critical step. Monthly billing data (or meter data) can measure *current* energy

consumption exactly; however, knowing the postparticipation consumption does not provide an estimate of the change in energy use caused by the program. To determine impacts, an estimate of the amount of the energy that would have been used in the absence of the program is needed; see Fig. 6 in Chapter 2. This is accomplished by establishing a baseline to serve as a proxy for the consumption that would have occurred without the DSM program.

Many of the estimation problems associated with impact evaluation are related to the selection of an appropriate baseline. Three such problems (self-selection, free riders, and free drivers) can be viewed as biases in the baseline (see Chapter 8 of this volume; EPRI 1991b; Violette, Ozog, and Wear 1991). Self-selection bias occurs when program participation is voluntary. In this circumstance, systematic differences may occur between nonparticipants, who either chose not to participate or were unaware of the program, and participants. Free riders and free drivers are important when the actions undertaken by the comparison group may not accurately reflect the actions that would have been undertaken by participants without the program.

STATISTICAL COMPARISON APPROACHES

This section outlines several simple-comparison approaches, shown in Table 7 of Chapter 2, plus methods that use other data available at the utility.

Three Comparison Options

The time-series approach compares the participants' pre- and postparticipation energy usages. The energy use of the participants prior to their participation in the program is used as the baseline energy use. Energy savings is estimated by:

$$\text{Savings} = Q_b - Q_a ,$$

where Q_b is the quantity of energy used before participating in the program and Q_a is the quantity used after participating. The before and after time periods may be annual periods or heating and cooling seasons.

In this case, the comparison group is comprised of the same set of participating customers. Among the comparison approaches, this method has the advantage of using a comparison group that is nearly identical in its dwelling and demographic characteristics because they would not be expected to change substantially over a 2- to 3-year period. However, this approach cannot capture other factors that may change with time, in particular electricity prices or weather. In addition, this comparison approach, by not using the change in energy use for nonparticipants, is not able to estimate natural conservation.

The cross-sectional comparison of participants with nonparticipants examines postprogram energy use across two customer groups:

$$\text{Savings} = Q_{np} - Q_p ,$$

where Q_{np} is the quantity of energy consumed by nonparticipants and Q_p is the quantity consumed by participants. In this case, the baseline is the nonparticipant energy use. This approach assumes that, in aggregate, the nonparticipant comparison group is identical to the participant group in all respects except for program participation. The advantage of this approach over the time-series method is that, because one time period is used, no biases result from factors that changed over time (e.g., weather conditions or prices). However, a representative comparison group may be difficult to obtain.

Combined time-series/cross-sectional approaches combine the advantages of both approaches and, most importantly, also control for natural conservation. These approaches compare *changes* in energy use over two time periods for two groups:

$$\text{Savings} = [Q_{bp} - Q_{ap}] - [Q_{bnp} - Q_{anp}] ,$$

where Q_{bp} and Q_{ap} are the quantities of energy consumed by participants before and after participating in the program, respectively. Q_{bnp} and Q_{anp} are the quantities of energy consumed by nonparticipants for the same time periods.

A comparison approach that uses percentage changes in energy use is often used:

$$\text{Percent savings} = [(Q_{bp} - Q_{ap})/Q_{bp}] - [(Q_{bnp} - Q_{anp})/Q_{bnp}] .$$

This percentage approach is generally preferred to the absolute-change approach because percentages help to normalize the estimated natural change in consumption for differences in the average size of customers between the participant and comparison groups.

The time-series/cross-sectional approaches are the best of the comparison approaches. Correcting the program savings estimates by the savings that occurred among nonparticipants helps to control for factors that change over time and influence energy use. Also, to the extent that the propensity for natural conservation is the same for participants and nonparticipants, the approach also controls for natural conservation (Rogers 1989; BPA 1990).

The drawback to this approach is that it assumes that factors such as weather and energy prices affect both participants and nonparticipants equally. This may not be the case. However, even the most sophisticated multivariate, simultaneous-equations approaches discussed later in this chapter are nothing more than variants of this simple-comparison concept. The sophisticated approaches simply attempt to develop better proxies for baseline energy use.

Comparison Approaches Augmented by Other In-House Data

All the comparison approaches discussed above are limited in their ability to control for those factors (aside from participation) that affect energy consumption, such as weather,

electricity prices, appliance stocks, and dwelling characteristics. However, utilities generally have information on weather and electricity prices in-house, so it is possible to control for these factors without undertaking additional customer surveys. A simple regression framework can accomplish this control where program participation, weather variables, and energy prices are used as explanatory variables. One widely used approach to control for weather is PRISM (Fels 1986; Fels and Reynolds 1991; Brown et al. 1991).

Besides weather, other confounding variables can be incorporated into the comparison techniques without having to rely on customer surveys. One such variable is electricity price. If the evaluation spans a 2- to 3-year period, the impact of changes in electricity price on consumption can be important. Because most utilities have this information within the billing file, obtaining it requires minimal effort. Other potential variables include house type (for residential programs), SIC codes (for commercial programs), or other customer and premise data contained in the utility's master customer-account files.

Statistical Precision

Designs and sample sizes that estimate *levels* of consumption with a high degree of precision must be distinguished from studies designed to estimate DSM savings, (i.e., a *change* in energy use). The large differences in precision between estimates of levels and estimates of changes can be illustrated with a simple example.

Assume a DSM program in which end-use metering is conducted on 100 participants and on 100 nonparticipants. The average consumption for participants was 25 kWh per day with a SD of 8 kWh per day. (Estimates of kWh savings per day are often used because they eliminate the need to control for different numbers of days in billing cycles, months, or heating seasons.) The mean use for nonparticipants was 30 kWh per day with a SD of 10 kWh per day. The savings estimate is the difference between these two mean consumption figures: 5 kWh per day or 16.7% of the mean nonparticipants' consumption.

From these data, the precision with which the mean consumption levels are estimated can be calculated for both groups. The following formula is used to calculate the absolute error of the estimates of means:

$$\text{Absolute Error} = t_{cr} \sqrt{\frac{s^2}{n}}, \quad (4-1)$$

where t_{cr} is the critical t -value associated with the confidence level (in this case, 1.98 for a 95% confidence level), s is the SD of the estimate (8 and 10 kWh, respectively), and n is the sample size (in this case, $n = 100$).

The end-use metering resulted in an absolute error of 1.98 for the estimated mean consumption for nonparticipants and 1.6 for participants. The relative precision is defined as:

$$\frac{\pm \text{Absolute Error}}{\text{Estimated Mean}}$$

The relative precision is $\pm 6.5\%$ for both participants and nonparticipants. Given these very precise estimates of mean consumption levels, the next issue is to determine the accuracy of the savings estimate based on the difference between these means. This accuracy is calculated by determining the absolute error of the savings estimate (i.e., the 5-kWh difference between the two means):

$$\text{Absolute Error}_{\text{diff}} = t_{cr} \sqrt{\frac{s_p^2}{n_p} + \frac{s_{np}^2}{n_{np}}},$$

where s_p and s_{np} are the SDs of participants and nonparticipants, respectively, and n_p and n_{np} are the sample size of participants and nonparticipants, respectively.

A 95% confidence level gives an absolute error of 2.5 kWh. The relative precision is found by dividing this absolute error by the 5-kWh/day savings, giving a $\pm 50\%$ relative precision. This large absolute error for the estimated savings contrasts sharply with the $\pm 6.5\%$ precision around the mean-consumption estimate for each group. The example illustrates that research designed to produce precise estimates of levels of consumption may not produce the hoped-for precision in DSM-impact estimates (i.e., the change in consumption).

USING PAIRED DATA TO ENHANCE PRECISION

Paired data occur when the sample observations contain two or more data elements. Examples of paired data can include (1) a prior estimate of savings and the floor area of the same building taken as a pair or (2) an engineering estimate of energy savings paired with a field estimate of energy savings based on billing data. In both instances, two observations for the same customer form a pair. Such paired observations can, in certain circumstances, provide estimates with smaller variances (Walpole and Myers 1985).

Combining Population and Sample Information

Certain data that might be available for the population of participants can be combined with sample data to increase the precision of estimated savings. The procedure illustrated here consists of normalizing energy use over a *known* variable to reduce variance in the estimates. One potential cause of variation in energy savings across buildings may be the size of the building. Large savings are more likely in large buildings. If this cause of variation can be eliminated, then the variance around the estimated savings can be reduced. For example, kWh savings per square foot may have a smaller variance and, therefore, can be estimated with greater precision. However, this procedure requires estimates of floor area for every customer in the sample, and if estimates of total program savings are to be made, these data are needed for all participants. Because energy audits are conducted as part of many DSM programs, such information is often available for all participants.

To illustrate, consider a program that has 250 participants, with meters installed to measure before and after electricity consumption for five of them. Two cases are presented: Case 1 (Column 1 of Table 13) uses sample data only and Case 2 (all three columns of Table 13) uses information from the energy audits.

Table 13. Estimates of electricity savings without and with additional information on building floor area

	Estimated savings, kWh (metered data) (1)	Floor area (ft ²) (2)	Estimated savings (kWh/hundred ft ²) (3)
Customer 1	25,000	15,000	166.7
Customer 2	45,000	25,000	180.0
Customer 3	20,000	12,000	166.7
Customer 4	50,000	31,000	161.3
Customer 5	15,000	9,500	157.9
Mean	31,000	18,500	166.5

With only the estimates in Column 1, the mean of the estimated savings for this sample is 31,000 kWh with a standard deviation of 15,600 kWh. A 90% confidence interval (CI) around this estimate is ± 2.13 times the SD divided by the square root of the sample size, or $31,000 \pm 14,860$ kWh. Thus, the level of precision for this estimate is $\pm 48\%$.

The second and third columns of Table 13 illustrate the use of additional information available on the population of program participants from audit records, in this case the floor area of each building. Now it may be possible to estimate program energy savings more precisely. This example shows that some of the variation in energy savings is explained by the size of the building. The mean savings per 100 square feet is 166.5 kWh with a SD of 7.5. The precision around this estimate is $\pm 4.3\%$, significantly less than the $\pm 48\%$ from the unnormalized sample.²

This example illustrates the advantage that can be gained by using information available on the population of participants to help explain the variation in savings. Often, data like these are available from audits conducted prior to program participation, or they can be collected as part of program implementation. Because program savings were, in this example, correlated with the building square footage, the gain in precision was significant.

²This example does not necessarily assume that the floor area for each building is measured precisely. Measurement error in these estimates will contribute to a larger SD in the estimated savings per square foot. As long as the floor-area data are positively correlated with energy use, the precision of estimated savings will be increased.

Difference and Ratio Estimates

Difference and ratio estimates can increase the precision of estimates of energy savings from conservation measures. These techniques have been applied in many areas of statistical analysis (Arkin 1974). In fact, one approach used in some evaluations, Statistically Adjusted Engineering (SAE) estimates, is a variant of the ratio estimator.

Difference- and ratio-estimation approaches can be viewed as procedures for "auditing" the accuracy of the prior engineering estimates. This auditing is accomplished by using field measurements of a sample from this population. Because many DSM programs, particularly C&I programs, develop detailed engineering estimates as part of program implementation, difference and ratio estimators have the potential for widespread use in program evaluation.

The difference estimator *modifies* the prior population engineering estimates by calculating an estimated mean difference between field and prior-engineering estimates using a sample of customers. For example, assume a commercial DSM program with 750 participants. The engineering estimates made as part of the program found that the average savings per participant for the total population would be 34,000 kWh. The program metered a sample of five participants to obtain field estimates of savings (Table 14).

Table 14. Electricity savings (kWh) from meter data and engineering estimates

	Estimated savings (meter data)	Estimated savings (engineering)	Difference (meter-engineering)
Customer 1	25,000	29,000	-4,000
Customer 2	45,000	44,000	+1,000
Customer 3	20,000	26,000	-6,000
Customer 4	50,000	48,000	+2,000
Customer 5	15,000	13,000	+2,000
Total	155,000	160,000	-5,000
Mean	31,000	32,000	-1,000
SD	13,928	--	3,742

Based on these data, the average engineering estimate is 1,000 kWh more than the metered estimate. (Note that the mean engineering estimate for the population is 34,000 kWh, but the mean for the 5 customers in the sample is only 32,000 kWh.) Therefore, an adjustment factor based on this estimated difference is applied to the mean engineering estimate for the population. With an estimated difference of -1,000 kWh, the resulting estimate of the population mean is $34,000 + (-1,000) = 33,000$ kWh.

If the conventional estimation approach is employed, with only the mean value for the metered sample of customers, the resulting estimate of the sample mean is 31,000 kWh. The precision of this estimate is found by:

$$SD_{conv} = \sqrt{\left[\frac{\Sigma(X_i - Mean_x)^2}{(n - 1)} \right]}$$

For the data in the example, the SD for the conventional approach is 13,928 kWh. With a 90% confidence level, the relative precision is 43%.

To measure the sampling precision of the difference estimate, the SD of the difference between the metered estimates and the engineering estimates is determined by:

$$SD_{diff} = \sqrt{\left[\frac{\Sigma(X_i - (Y_i + Diff))^2}{(n - 1)} \right]}$$

where X_i is the metered estimate for observation i ; Y_i is the engineering estimate for observation i ; Diff is the difference adjustment factor; and n is the sample size. With the data in Table 14, the resulting estimated SD for the difference estimator is 3,742 kWh. The precision, with a 90% CI, is $\pm 11\%$.

For these hypothetical data, incorporating prior engineering estimates to create a difference estimate produces a program energy-savings estimate that is nearly four times more precise than the conventional approach using only the meter data for the sample.

While the difference estimate uses the difference between the meter estimate and the engineering estimate, the ratio estimate uses the ratio of the metered estimate to the engineering estimate. The applications of these two methods is similar; the choice between the difference and ratio approaches depends on the relationship between the metered estimates and the engineering estimates (EPRI 1991a; Wright and Townsley 1990). If the size of the difference between the two is independent of the size of the savings (i.e., a large engineering estimate does not imply a large difference between the engineering estimate and the metered estimate), then the difference estimator is more precise. In the extreme, if the engineering estimate is always a fixed magnitude different from the metered estimate, then the SD of the difference estimate is zero.

If the size of the difference between the engineering and metered estimate is related to the size of the engineering estimate, then the ratio estimate is generally more precise. If the engineering estimate is consistently different by a certain percentage (e.g., it always overstates savings by 20%), then the SD of the ratio estimate is zero. In practice, if a difference estimate can be calculated, then a ratio estimate can also be calculated. The analyst can then calculate both difference and ratio estimates and select the approach that produces the smallest standard error of the estimate.

In the above example, the predicted energy savings were not normalized for floor area. The normalization method presented above can be combined with statistical audit procedures to further increase the precision of estimates. In this case, both the engineering estimates and the meter data would have to be expressed in terms of savings per square foot.

SAE Estimates as Ratio Estimates

SAE estimates are a type of ratio estimate for which the paired data encompass a field-derived statistical estimate and a prior engineering estimate of savings. SAE applications fall into two categories: (1) auditing the accuracy of engineering-based program-saving estimates and (2) disaggregating whole-building-metered loads into individual end-use loads (EPRI 1991a). The example presented here focuses on DSM program savings.

SAE estimates are produced by using engineering estimates as normalizing variables or, when more than one DSM measure is being considered, as independent variables in a regression equation. The basic non-SAE model is:

$$\Delta \text{Energy}_t = \text{Energy}_t - \text{Energy}_{t-1} = \beta_1(\text{Part}_1) + \beta_2(\text{Part}_2)$$

or

$$\text{Energy}_t = \beta_1(\text{Part}_1) + \beta_2(\text{Part}_2) + \lambda(\text{Energy}_{t-1}) ,$$

where λ is the coefficient on the lagged energy variable, Energy_{t-1} , and Part_1 and Part_2 are (0,1) binary variables that indicate participation in more than one DSM program (i.e., programs 1 and 2), where the variable takes on the value of 1 if that customer is a participant and is zero otherwise. Alternatively, Part_1 and Part_2 could refer to different DSM measures.

Under the SAE approach, the basic model is:

$$\text{Energy}_t = \beta_1(\text{Eng}_1) + \beta_2(\text{Eng}_2) + \lambda(\text{Energy}_{t-1}) .$$

Now the participation variables are engineering estimates (i.e., Eng_1 and Eng_2) of the savings from different DSM measures or savings associated with participation in programs 1 and 2, respectively. In either case, program-specific or measure-specific engineering estimates of savings must be available. SAE estimates can also be applied in a multivariate format. The only difference is the addition of other explanatory variables, X_i , to the above equation.

The interpretation of the regression coefficients in the SAE model is different from the interpretation of those in the first model. A coefficient of 0.5 indicates that, on average, only 50% of the predicted engineering estimates were in fact realized, based on the measured consumption data.

The use of engineering estimates of savings as independent variables in the regression equation instead of dichotomous 0-1 variables has several potential advantages. First, it is a way to compare statistically-derived estimates with engineering estimates. Second, the direct incorporation of the engineering estimates into the equation brings more information to bear on the estimation problem. If one building, according to the energy audit, has greater potential savings than does another building, then this information is reflected in the variable representing participation (i.e., the engineering-savings estimate). If this information is hierarchically accurate (that is, the building does have a higher savings potential and does achieve higher savings), then the precision of the estimated savings should increase. If the engineering model is so inaccurate that it is unable even to hierarchically rank buildings by savings potential, then the SAE model estimates may be less precise than simply using (0,1) participation variables. This condition is not very restrictive, and SAE estimates will likely increase the precision of savings estimates in most instances. However, this thesis can be tested statistically. Simply estimating the equation with (0,1) indicator variables and also estimating the model with the engineering estimates will show which specification produces the more precise estimate. If the engineering estimates are accurate enough to hierarchically rank customers by savings, then the SAE formulation will produce savings estimates with smaller SDs and higher *t*-values than will the non-SAE model formulation.

MULTIVARIATE-SAVINGS-MODEL SPECIFICATIONS

The comparison approaches discussed above require data that most utilities typically have and produce useful information about the impacts of a DSM program. However, with these approaches, the researcher cannot know whether the participant and nonparticipant groups are truly comparable. The only data available to confirm the similarity of the two groups are billing data and other in-house information. Adding customer-specific information (e.g., from surveys) may increase the confidence in impact estimates. In some cases, such additional information is obtained from energy audits of samples of participants and nonparticipants (CMP 1990; BPA 1990). To date, most multivariate models have been developed for residential programs; however, applications to commercial-sector programs are increasing (Consolidated Edison 1991; Central Hudson Gas and Electric 1991; San Diego Gas and Electric 1991; Train and Ignelzi 1987).

The availability of additional customer-specific data allows for the use of multivariate models that can incorporate more structure in terms of engineering relationships, building characteristics, and customer behavior and attitudes. This additional information can be important when program savings are expected to be small compared to the customer's total energy consumption. A model to isolate program impacts on the order of 5% of total consumption requires that all available information be used, including a well-defined structure and as much information from engineering principles as possible.

Evaluators can estimate total energy-use models (cross-sectional models) and change models (pooled time-series/cross-sectional models). One total-energy-use model and two forms of change models are presented in Exhibit 6.

Exhibit 6. Specifications of energy-use models

Three types of energy-use models can be used for DSM-impact evaluation, as illustrated here. Assume that there are three end uses [space heating (SH), AC, and water heating (WH)], that program participation is denoted by a 0,1 indicator variable (Part), and that participation occurs at the end of 1990.

1. Cross-Sectional Model

$$E_{1991} = \beta_0 + \beta_1 SH_{1991} + \beta_2 AC_{1991} + \beta_3 WH_{1991} + \beta_4 Part + \epsilon_{1991}$$

2. Constrained-Change Model

$$E_{1991} - E_{1990} = [\beta_0 - \beta_0] + [\beta_1 SH_{1991} - \beta_1 SH_{1990}] + [\beta_2 AC_{1991} - \beta_2 AC_{1990}] \\ + [\beta_3 WH_{1991} - \beta_3 WH_{1990}] + \beta_4 Part + [\epsilon_{1991} - \epsilon_{1990}]$$

or

$$\Delta E = \beta_1(\Delta SH) + \beta_2(\Delta AC) + \beta_3(\Delta WH) + \beta_4 Part + \Delta \epsilon$$

3. Flexible-Form Model

$$E_{1991} = [\beta_0 - \lambda_0 \beta_0] + [\beta_1 SH_{1991} - \beta_1 \lambda_1 SH_{1990}] + [\beta_2 AC_{1991} - \\ \beta_2 \lambda_2 AC_{1990}] + [\beta_3 WH_{1991} - \beta_3 \lambda_3 WH_{1990}] + \beta_4 Part + \lambda_4 E_{1990} \\ + [\epsilon_{1991} - \lambda_5 \epsilon_{1990}]$$

Here, $\lambda_0, \dots, \lambda_5$ are adjustment factors. In all cases, the effect of the program on energy use is given by the coefficient on the participation variable (β_4).

Total-Energy-Use Models

Preparticipation data simply are not available in some cases, such as for new-construction programs. For these programs, the only option is to use a cross-sectional framework in which the postparticipation energy use of a participant group is compared with the energy consumption of a nonparticipating group. This comparison can be done with a multivariate-regression model in which total energy use is modeled and a participation variable is used as an explanatory variable.

The total-energy-use model should include all measurable factors that influence electricity consumption. Any omitted factor that varies systematically across participants and nonparticipants will bias the estimates of program impacts. Although different specifications of models can be used, one approach that is theoretically sound involves use of a conditional-demand specification. Total electricity demand is modeled as a function of the electricity used in each end use.

A model of residential energy use illustrates the approach. The household's total electricity demand at time t (E_t) is the sum of the electricity used for space heating (SH_t), air conditioning (AC_t), and water heating (WH_t) and an error term (ϵ_t):

$$E_t = SH_t + AC_t + WH_t + \epsilon_t \quad (4-2)$$

In practice, electricity use for each end use is itself a function of variables that represents customer behavior and engineering principles (EPRI 1991a; Parti and Parti 1980). Applications of the conditional-demand model for estimating DSM-program savings with a total energy model can be found in Wisconsin PSC (1989), CMP (1990), and Violette and Ozog (1989).

Change Models: Pooled Cross-Sectional/Time-Series Models

The total-consumption models discussed above are cross-sectional (i.e., they analyze total energy consumption at one point in time). Models that include changes in energy use over time are likely to be more appropriate for DSM-program evaluations. Pooled cross-sectional/time-series models have the following form:

$$\Delta E = \beta_0 + \beta_1 \Delta SH + \beta_2 \Delta AC + \beta_3 \Delta WH + \epsilon \quad (4-2a)$$

where Δ denotes the change in the value of that variable (e.g., $\Delta E = E_t - E_{t-1}$).

For each end-use variable, a change can result from changes in one or more of the individual components of that variable. For example, the energy used for SH can change because of changes in the heating equipment, building insulation, occupant behavior, or weather. In some cases, no changes will occur in the individual variables so the composite change variable will not appear in the model. Equation (4-2a) is referred to as a constrained model because it constrains the coefficient (i.e., the energy-use intensity) for each end use to be constant between the t and $t - 1$ periods being analyzed.

A more flexible form of the change model uses a lagged dependent variable as an independent variable (i.e., E_{t-1} appears on the right-hand side of the equation) and introduces adjustment factors (λ_i) for each end use.

$$E_t = (\beta_0 - \lambda_0\beta_0) + \beta_1(SH_t - \lambda_1SH_{t-1}) + \beta_2(AC_t - \lambda_2AC_{t-1}) + \beta_3(WH_t - \lambda_3WH_{t-1}) + \beta_4(PART) + \lambda_4E_{t-1} + \epsilon_t \quad (4-3)$$

The λ_i ($i = 1, \dots, 4$) allow energy use to change with time, both in total and by end use. This equation can be used to model the change in annual energy use, the change in energy use between one month and the same month a year later, or the change from one month to the next month.

This flexible change model results in a unique pattern among the coefficients. Equation 4-3 shows that, for each end use, the coefficient of the lagged variables (e.g., the SH_{t-1} term) equals -1 times the adjustment factor λ_i . This pattern is illustrated in Table 15, which shows an example of the flexible change model for a residential audit program. Here the energy savings is 546 kWh (the coefficient on the participation variable). The coefficients on the preparticipation (1985) variables are of a similar magnitude, but with an opposite sign to the coefficients on the same variables for 1987.

The flexible-form-change model is a useful specification when pre- and postprogram consumption data are available. This model should serve as one of the basic specifications used in multivariate analysis of DSM impacts.

SELF-SELECTION BIAS AND FREE RIDERS

Several methods of addressing self-selection bias and free riders use the types of multivariate models discussed above. The estimation problems associated with self-selection and free riders stem from the same source, a comparison group that is not truly representative of what the participant group would have done had the program not existed (Keating 1989). For example, if nonparticipants are more likely than participants to expand floor space, then the pre/post comparison-group energy use would encompass this change. The baseline trend in energy use estimated from the comparison group would be too high.

Potential bias because of free ridership reflects one aspect of self-selection bias. If the comparison group is under-represented by customers that have a propensity to conserve when compared to the participant group, then free ridership may be an issue. In this case, the participant group contains more customers who would have undertaken conservation actions even if the DSM program had not been offered. Therefore, the comparison-group change in energy use under-estimates the natural conservation that would have occurred had the program not been offered. Viewed in this fashion, self-selection bias and free ridership are aspects of the same estimation problem, with free ridership being a subset of self-selection biases.

The most appropriate procedures for addressing self-selection may not allow the analyst to separately identify the impacts of free riders independent of the broader self-selection bias. The specific estimation of free-rider effects is required by some state regulations; however,

Table 15. Electricity-savings, unconstrained-change model^a

Independent variable	Coefficient	(t-values)
Intercept	1256.42	(4.73)
Participation variable	-545.94	(-4.30)
Number of people at home during the day in 1987	527.65	(4.99)
Number of people at home during the day in 1985	-389.97	(-3.63)
SH ^b times heating degree days, 1987	0.00047	(3.83)
SH ^b times heating degree days, 1985	-0.00057	(-4.81)
SH ^b times average price, 1987	-0.0215	(-12.24)
SH ^b times average price, 1985	0.0237	(12.03)
SH ^b times income, 1987	0.00020	(7.46)
SH ^b times income, 1985	-0.00020	(-7.35)
Company #1 indicator variable	-1234.83	(-5.93)
Company #2 indicator variable	-910.23	(-4.78)
Electricity consumption, 1985	1.03	(77.72)
Number of observations	1169	
R ²	0.90	
Mean of dependent variable	12,460	

^aThe dependent variable is postparticipation electricity consumption (1987). The preparticipation period is 1985, with 1986 being the participation period.

^bSH is a variable representing the estimated surface area of the home.

Source: New Jersey Conservation Analysis Team (1990).

this may be difficult to accomplish, even with the sophisticated estimation techniques designed to address self-selection.

Two general approaches for addressing self-selection within multivariate models are available. The first approach is based on simultaneous-equations methods; the second uses the properties of the cross-sectional/time-series change models to reduce the likelihood of self-selection bias.

Simultaneous-Equations Approach

Self-selection can be viewed as a situation in which the amount of energy a customer uses is a function of program participation and the decision to participate is a function of energy use. Thus, participating in a DSM program and energy use are joint decisions.

Approaches for addressing self-selection based on a simultaneous-equations framework require the estimation of two equations, a discrete-choice participation model and an energy-use equation such as the regression models discussed above. Among the approaches for explicitly addressing self-selection bias, one easily applied method is the selectivity-correction-factor technique. This technique, developed by Heckman (1978) and refined by Dubin and McFadden (1984), involves the estimation of a participation model and a multivariate-regression model of energy use. In theory, if the researcher can incorporate into the energy-use equation all the variables that characterize participation, then self-selection bias will be minimized. Many of the variables that influence the decision to participate are difficult to observe (e.g., attitudes and beliefs). The selectivity-correction technique uses a participation model to estimate the probability of participation. Then, based on the estimated probability of participation, a selectivity-correction term is developed. The general form of this correction term is:

$$C_i = \left[\frac{\hat{P}_i \ln \hat{P}_i}{(1 - \hat{P}_i)} + \ln \hat{P}_i \right], \quad (4-4)$$

where \hat{P}_i is the estimated probability of participation from the participation model. The new energy-use equation is:

$$\Delta \text{Energy} = \beta_1(\text{PART}) + \beta_2(C_i) + \beta_3(x_i) \quad (4-5)$$

This is a multivariate regression model, as discussed in the previous section, with one change: C_i is added as an explanatory variable. Calculating the C_i is straightforward:

- Estimate a discrete choice participation model with a sample of participants and nonparticipants, data on customer attitudes, and other factors believed to influence the decision to participate in the program.³
- Use the participation model to estimate the probability of participation for each customer in the sample.
- Use the probabilities of participation for *each* customer in the sample to calculate the C_i for each customer.
- Include the calculated C_i variable in the energy-use regression equation.

³A discrete-choice model uses discrete outcomes as the dependent variable. Here, the dependent variable is 1 if the customer is a participant and is 0 if that customer is a non-participant. The independent variables include factors that influence customer decisions to participate, including income, expected energy savings resulting from participation, and awareness of the program.

The selectivity-correction approach introduces the term C_i into the equation, which allows the researcher to test for bias. Violette and Ozog (1989) pointed out several aspects of this technique that influence the interpretation of the findings. Most importantly, a large, statistically significant coefficient on the selectivity correction term does not necessarily imply that the estimates of program energy savings have a large bias. In some cases, little correlation exists between the magnitude of the coefficient on the selectivity term and the size of bias in the energy-savings estimate. The only way to examine this is to estimate the model with and without the selectivity factor and then compare the resulting savings estimates. Finally, the bias can be either positive or negative (i.e., it can increase or decrease estimated program savings).

Implicit Correction for Self-Selection: Pre/Postregression-Analysis Techniques

A second approach to self-selection uses a change model based on pooled time-series/cross-sectional data. Under a reasonable set of assumptions, the change models presented in the previous section can reduce self-selection bias (Heckman and Robb 1985; EPRI 1991a). This technique exploits the fact that an evaluation is only concerned with how energy use changes over time and not what determines total energy use. By investigating only the factors that alter energy use, one can mitigate the effects of self-selection bias without using sophisticated models.

To understand this approach intuitively, assume that participants in a conservation program are more likely to consume less energy without the program than nonparticipants are. If this propensity to consume less energy is constant over the short run, then a difference model cancels out this propensity to consume less energy, eliminating the bias.

The implication of this result is that by collecting data on at least one period of preprogram energy use, researchers can produce an estimate of DSM impacts that may not be affected by self-selection. However, this approach does not correct for self-selection if participants' and nonparticipants' change in energy consumption exhibits a systematic difference with time. For example, if participants consistently take more conservation actions over time than nonparticipants, then participants not only will have lower levels of energy consumption, but also will have more rapidly changing energy consumption. Figure 8 depicts such a situation. Because the rate of change in energy use over time is different for the two groups, this difference between participants and nonparticipants does not cancel out in a two-period model. Under these circumstances, a pre/postmodel will not fully correct for self-selection bias, and the use of the self-selection correction factor may be warranted.

Two methods for mitigating self-selection bias have been presented. The choice of which approach to use depends on several factors. First, the selectivity correction technique can be used with both cross-sectional and pooled cross-sectional/time-series data, while the change in energy use approach can only be used with pooled cross-sectional/time-series data. Thus, the selectivity-correction technique is particularly useful for new-construction programs. Also, the potential for self-selection bias is much greater in cross-sectional models.

ELECTRICITY USE PER CUSTOMER (kWh or kW)

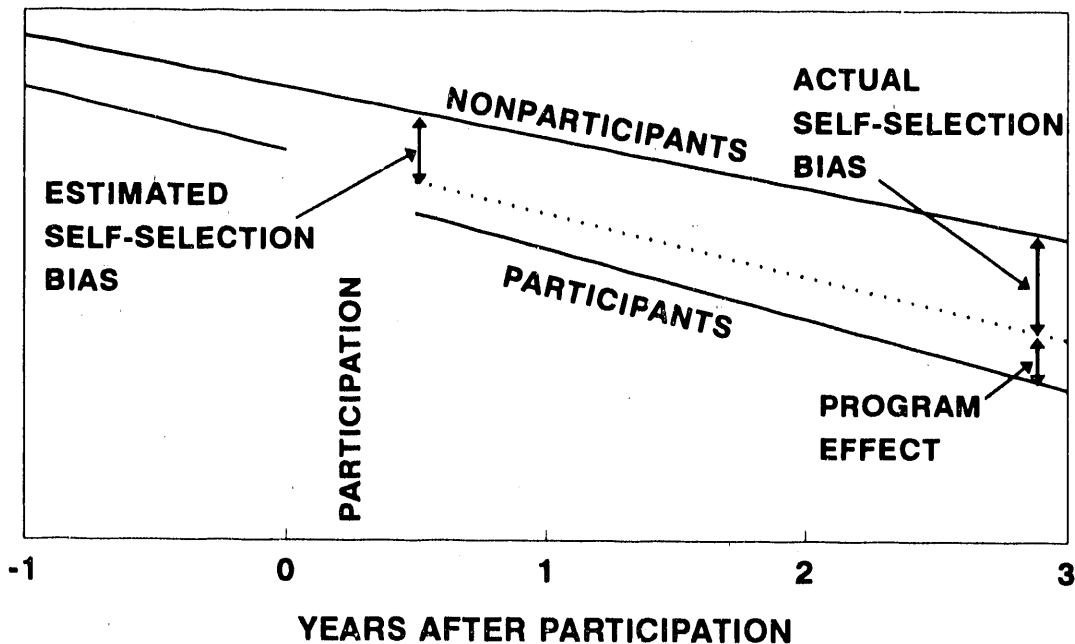


Fig. 8. Self-selection bias in pre/postmodels.

SELECTED SPECIAL ISSUES

Data Analysis and Estimation Accuracy

Some of the confidence levels and tolerance limits set as targets for DSM research are based on those developed for end-use research where the objective was to measure a consumption level, not a change over time. These precision requirements may not be appropriate for DSM evaluation. One criteria for determining the appropriate level of accuracy should be whether the information is adequate for making good business decisions regarding investments in DSM programs. The confidence needed for this may be different than that required for other utility investments because of the ability to monitor DSM programs periodically and the high ratio of variable-to-fixed costs of DSM programs compared to a power plant. If savings are estimated to be 10%, then is it necessary to have 90% confidence intervals with 10% tolerances, i.e., must the utility know that the true savings falls between 9% and 11%: Or might an estimate of 10% savings, where the analyst is 90% confident that actual savings will fall between 6% and 14% (i.e., a $\pm 40\%$ precision), might meet the criteria for good business decisions, particularly if this number can be updated frequently as additional program experience is gained. In this case, a requirement that savings estimates achieve a $\pm 10\%$ precision seems high. Also, if program costs are low, the program may be cost-effective at 6% savings. Ultimately, the specification of accuracy required for decision making will be decided by the utility and other interested parties.

Multiple-Analysis Methods

Multiple analysis methods are frequently used for evaluation. For example, experimental designs sometimes include engineering estimates, metering data (whole-premise or end use), and billing data analyses (CMP 1990; Northeast Utilities 1991a and 1991b; the New Jersey Conservation Analysis Team 1990). The use of multiple methods can encompass different approaches. Three are discussed below.

- **Triangulation** involves the use of more than one approach to estimate the impacts of a DSM program. See the subsection on Combinations of Data Sources in Chapter 3 and Hirst (1991) for more information.
- **Leveraging data** differs from triangulation in that it uses data from one method as input to another method to increase the precision of the impact estimates. This approach uses data to reduce the variance in the parameter being estimated. Difference and ratio estimates are examples of leveraged approaches in which data from two analyses are combined. For example, in SAE methods, the engineering estimates of program or measure savings are used as inputs to the statistical model (EPRI 1991a); see Chapter 3 for additional discussion of this topic.
- **Prior information** on the variable of interest (e.g., program savings) is used as a starting point. The uncertainty around this estimate is quantified from previous studies or expert judgement. This approach uses this information as an input to a Bayesian-estimation framework. Information on the impacts of DSM programs and measures can come from prior evaluations, engineering models, and expert judgement. The use of this information with a systematic approach to updating and adapting existing estimates based on new evaluation data could usefully augment conventional estimation approaches. In some cases, this may be a successful primary estimation method when the objective of the evaluation is to update existing estimates. This approach is discussed further later in this chapter.

The use of multiple methods and data sources allows the evaluator to use the available information in a cost-effective manner. Information in a DSM evaluation can be leveraged through a number of methods.

- Using end-use metering information in engineering models for calibration, validation, and simulation of a hypothetical baseline (Kaplan, Jones, and Jansen 1990; EPRI 1991b)
- Using engineering relationships to specify statistical models (King and Gavelis 1990)
- Using engineering estimates and/or metering to augment statistical estimates (EPRI 1991a and 1991b; Empire State Electric Energy Research Corp. 1990)
- Combining population and sample information

- Combining expected or forecasted savings with statistical models by using statistical audit procedures or using SAE estimates
- Using prior information in the form of savings estimates and distributions (EPRI 1991a)

Procedures That Accumulate Information Over Time

One concern with current methods is that each evaluation essentially starts over, ignoring the results of previous evaluations. This procedure may be appropriate for medical research where falsely rejecting the null hypothesis carries an extreme penalty. However, the appropriateness of this assumption for DSM evaluation is an open question. The use of prior information in a Bayesian framework assumes that information should be cumulative. On this basis, the researcher can start with existing estimates and then revise these estimates based on new data.

The key requirement for a Bayesian analysis is the establishment of an acceptable and unbiased "prior estimate" of impacts or consumption levels. While this requirement may be a difficult one to meet, these techniques might usefully augment traditional estimation approaches. The use of prior information as constraints, as initial estimates of savings, or in the form of a likely distribution of savings may all be useful. These approaches assess the likelihood of observing a drawn sample, given the conditionally assumed prior estimates. The use of these methods may be one way to achieve the levels of accuracy that some analysts are hoping for in evaluation. Software is now available that allows the analyst to start with an existing estimate that is bounded by wide CI (EPRI 1991a; Pollard 1986; Schmitt 1969).

Free Drivers and Comparison-Group Contamination

The issue of free drivers is increasingly recognized as important in DSM evaluation (Chapter 8). Several approaches for addressing this issue have been suggested. One approach is to change the estimate of baseline energy use from which the impacts are measured. For example, it may be appropriate to use a historical baseline from the early years of the program. Free drivers are more likely to be a significant problem for programs that have been in existence for several years and have achieved high participation levels. If baseline energy use for nonparticipants is taken from one of the earlier years, then the concern about nonparticipant consumption being influenced by the program is reduced. Because it can be assumed that some energy-efficiency improvements were likely to occur without the program, this approach typically would provide an upper bound on savings. However, if the time span is only two or three years, this improvement could be expected to be minor.

A second approach involves the use of survey methods to determine whether nonparticipants have changed their energy use as a result of the program, as discussed in Chapter 8. Nonparticipants could be queried as to whether their behavior has changed because of the program. Trade allies could be queried to determine whether they believe the market for

the DSM actions promoted by the program has moved in such a way that nonparticipants are impacted (CMP 1990).

While free drivers are one form of comparison-group contamination, several other concerns regarding contamination have been raised as well. These concerns include whether or not to include in the comparison group customers that participated in DSM programs other than the program being offered. A second issue is whether to exclude from the comparison group customers that participated in the same program prior to the analysis period (Chapter 6). In some instances, these customers cannot be cost-effectively identified and excluded from the comparison group. The question is whether this inclusion poses a potentially important bias. In most cases, this will not be an important problem. If a change model is used, then a number of factors will reduce the potential influence of these factors on impact estimates. Information should be collected on all programs in which the customers have participated to allow for testing of the significance of this factor on estimates, but the use of a pooled cross-sectional time-series experimental design should reduce any bias this might cause.

End-Use and Whole-Premises Metering Versus Billing Data

The finer the temporal and end-use resolution in the consumption data, the better the analysis (Chapter 7). The statistical problems and methods discussed earlier in this chapter are equally applicable to end-use, whole-premise, and billing data. Using difference- and ratio-estimation methods to increase precision is more relevant with end-use data because of their ability to reduce the required sample sizes to achieve a given level of precision.

There have been only a few metering studies that have been specifically designed to estimate the savings resulting from DSM programs (Northeast Utilities 1991a; Violette 1990). Pre- and postparticipation metering are required to accurately capture the change in consumption. Several utilities have incorporated end-use metering projects in evaluation plans. As these studies are completed, considerable information on the applicability of metering will be available.

Engineering Methods

Engineering approaches can provide information on equipment and system performance characteristics and operation profiles of measures installed through the programs. These techniques have seen widespread application in DSM-program impact evaluation as well as in program planning and screening (EPRI 1991a).

Early comparisons between impact-evaluation methods using consumption data (e.g., billing histories) gave estimates that were substantially lower than the engineering estimates. However, recent work using engineering methods based on onsite inspections after installation of DSM measures show much closer agreement between methods.

Engineering methods can serve several useful functions in impact evaluations.

- **Engineering estimates may provide independent, stand-alone estimates of program impacts.** Engineering methods provide a quick method of developing before- and after-the-fact estimates, as well as ongoing monitoring of program impacts. As an evaluation method, however, engineering assumptions and estimates need to be benchmarked against billing-data analyses and end-use metering studies.
- **Engineering estimates can verify statistical methods.** Statistical approaches can produce implausible estimates for certain programs, often because of data-quality problems. Engineering estimates can help assess plausibility and provide backup estimates.
- **Engineering methods are sometimes the most cost-effective method.** Engineering methods are inexpensive and can serve as a primary method when the value of information does not justify more-expensive statistical and end-use-metering approaches.
- **Engineering estimates can provide time differentiation of impacts.** Statistical estimates are limited by the temporal disaggregation of available consumption data, often a one-month billing period. Valuation of impacts in benefit/cost analyses often requires estimates of consumption impacts differentiated into on-peak and off-peak periods and load-shape impacts by hour of the day. Savings from billing analyses can be allocated to these time periods with engineering models.
- **Engineering approaches can estimate savings for other fuels.** Billing data for natural-gas or fuel-oil consumption may not be available to an electric utility, yet the program may influence the consumption of these fuels. In some cases, it may be more cost-effective and expedient to rely on an engineering estimate.
- **Engineering methods can offer measure-specific resolution.** Statistical approaches may be constrained in determining impacts by individual measures. The available data for a billing analysis may be on a package of measures, and multicollinearity may prevent the identification of measure-specific impacts. Engineering estimates can allocate aggregate savings to individual measures or combinations of measures, including interactive effects. For example, a DSM program may offer a package of hot-water conservation measures. Engineering methods can apportion the estimated total savings for the package to individual measures (e.g., low-flow showerheads, pipe insulation, and water-heater wrap).

CONCLUSIONS

Performing impact evaluations will continue to challenge evaluators, who will apply methods that draw from many areas of expertise. These areas include survey research, sample design, statistics, market research, consumer theory, economics, and engineering. Future DSM evaluations will likely be characterized by several factors:

- Increased emphasis on discrete-choice modeling of DSM decisions. Such analyses develop better estimates of program impacts through improved baseline usage estimates. Also, these methods provide valuable insights that can be important for program marketing and understanding customer decision processes.
- More end-use and whole-premises metering. The costs of metering are declining, and experimental designs are being developed to use this information more efficiently.
- Nearly universal analysis of electricity billing data, even if additional metering is being conducted. Billing data are readily available, and the additional information required for even the most detailed analyses of billing data are generally collected through customer surveys. Showing how monthly energy consumption changed is the starting point of most evaluations. Additional methods can be used to explain why this change is or is not an accurate estimate of program impacts, but most decision makers want to see this basic calculation.
- Use of multiple approaches, in which the output of one method is used as the input to another method. Examples of these approaches include difference estimators, ratio estimators, and SAE estimates.

REPORTING AND USING EVALUATION RESULTS

Steven Nadel

INTRODUCTION

This chapter discusses what evaluation material to report, how to report that information, and how to use the evaluation results developed in the other chapters. This chapter emphasizes the orderly and complete reporting of all major types of evaluation information and the diverse uses for evaluation data, both internal and external to the utility conducting the program.

REPORTING EVALUATION RESULTS

Sometimes evaluation results are reported in a single document; other times several discrete reports are issued. The most common reporting arrangements are separate process- and impact-evaluation reports on each program or annual reports that summarize all available information on a utility's DSM programs.

In addition to basic process- and impact-evaluation data (Chapter 1), other types of evaluation data are important and useful, including data on participation rates, program costs, measure life, and free riders. Often, these data are included as part of evaluation reports, but these data are frequently omitted from such reports entirely. Much of these data are available in databases that track program activity. Because these data are continually compiled, they can be reported at almost any point, although annual reporting is most common. Table 16 provides a checklist of the information that should be included in comprehensive evaluation reports; most of these topics are discussed below.

Table 16. Key ingredients of a comprehensive evaluation report

Table of contents	Program costs
Program description	Program energy and load reductions
Evaluation method	Measure life
Participation	Cost-effectiveness
Number of participants	Process-evaluation results
Participation rates	Summary and recommendations
Free riders	

Program Description

Each report should include a description of the program being evaluated to provide a context for reviewing the results. The description may be a brief summary (with citations to a more in-depth description), or a complete description may be included in the evaluation report or in an appendix to the report.

The description should include: a history of the program (e.g., when it began and ended); program goals, including the target audience; an indication of whether the program is a pilot or full-scale program; the program's components (marketing, incentives, information, technical assistance, etc.); the program's operations (audit, rebate application, etc.); and changes made to the program since previous evaluations or program descriptions were published.

Evaluation Method

Every evaluation report should discuss the methods used. For process evaluations, the description of methods may be brief, indicating who was interviewed, how these people were selected, what questions were asked [commonly a copy of the interview guide(s) is included in an appendix], what other data were used and their sources, and who conducted the interviews and other evaluation activities.

For impact evaluations, the description should cover sample sizes, sample selection, comparison-group selection, data-cleaning rules (what criteria were used to eliminate or modify poor-quality data), data attrition (how many customers in the original sample were eliminated from the analysis and for what reasons), and the analysis procedures used. This information allows the reader to assess the validity of the approach and the results obtained. When discussing data attrition, summarizing the results in a table is often useful (see Table 17). In discussing the method, explaining why particular approaches were selected is also useful. The method write-up often includes citations to other sources that describe particular evaluation approaches in detail.

The methods used to track and calculate other program data should be reported, including how the participation rate, free-rider proportion, direct and indirect costs, and energy and demand savings were defined and calculated. Because different evaluators use different approaches and definitions for key terms, it is important to specify which approach and definitions were used (Berry 1989 and 1990; Nadel 1991; Hirst and Sabo 1991).

Participation

Data on the number of participants is used to assess the overall size of a program and to calculate indices, such as participation rate and cost per customer. The number of participating customers is typically calculated in one of three ways:

Table 17. Data attrition in the low-income segment of CMP's Weather Shield and Attic Attack Program^a

	Number of cases	
	Participant group	Comparison group
Original database	1,904	442
No billing data available	274	---
Master meter; less than 300 days of data; vacancies; duplicate accounts	226	33
Greater than 50% change between pre- and postparticipation years	86	36
Remaining observations	<u>1,318</u>	<u>373</u>

^aThis table shows the original sample size for the participant and comparison groups and the number of cases that were lost at each stage of data cleaning, including missing data ("No billing data available"), unusable data ("Master meter, ..."), and data that distort the analysis ("Greater than 50% change ...").

Source: Stucky et al. 1990.

- The number of applications processed (including multiple applications submitted by a single account)
- The number of unique account numbers participating in the program (including multiple account numbers used by the same home or business)
- The number of unique customers participating in the program

While the last definition is probably the most useful, it is frequently impractical to implement because most record-keeping systems cannot identify all accounts associated with a particular home or business. The second definition is easier to calculate because identical account numbers are generally easy to find. For this reason, it is likely to be the preferred definition for most applications. Use of the first definition should generally be avoided because many programs encourage customers to participate more than once, which makes impossible the calculation of how many eligible customers have not participated in a program.

Three other important issues in calculating the number of participating customers are whether to count pending applications or only completed applications, whether to count all participants or only participants who are not free riders (see Chapter 8), and whether to

count audit recipients or rebate recipients in combined audit/rebate programs (not all audit recipients apply for a rebate). Because some pending jobs are never completed, counting only completed applications is usually best. Generally, in calculating participation, no adjustment is made for free riders; instead, the free rider proportion is tracked separately. For combined audit/rebate programs, the number of audit recipients and rebate recipients are usually counted separately because both figures are useful.

Evaluators should also categorize participants by market segment, such as customer electricity use (peak demand or annual kWh use), business or house type, income or economic activity, geographic region, and the DSM measures installed. Categorizing participants makes it possible to compare participants to eligible nonparticipants to see which groups participate and which do not.

Participation rate, the number of participating customers divided by the number of eligible customers, indicates the proportion of customers served by a program and, by subtraction, the proportion that have been missed and hence need to be the target for future efforts. The number of eligible customers may be all residential or C&I customers, or it may be a subset (e.g., C&I customers with peak demand greater than 500 kW).

Sometimes, the number of customers eligible for a program is large, but a much smaller number is specifically targeted. In such cases, evaluators should report two participation rates, one based on eligible customers and one based on targeted customers. In calculating the number of eligible customers, the same rules should be used as are used for calculating the number of participants (Berry 1990; Nadel 1991; Hirst and Sabo 1991).

For some programs, participation rates may be based on factors besides participating and eligible customers. For equipment rebate programs, such as motor, ballast, and refrigerator programs, the participation rate may be defined in terms of pieces of equipment that received rebates relative to annual sales of that type of equipment in a utility's service territory.

Both annual and cumulative (since program inception) participation rates should be reported. In addition to reporting participation rates, interpreting the rates obtained in terms of the utility's short- and long-term goals for the program is useful. For example, is the program on schedule?

As noted in Chapter 8, estimating free riders is an imprecise science. Therefore, in reporting free riders, reporting a range and which point within the range is most likely (for use in calculations that require a single-point estimate) is often appropriate. Because free-rider rates are likely to vary from measure to measure, evaluations of programs that promote multiple measures should present free-rider estimates by measure.

Program Costs

Information on program costs is needed for budgeting as well as for determining program cost-effectiveness. Costs associated with a program fall into three general categories: direct

utility costs, indirect utility costs, and nonutility costs (Berry 1989; Hirst and Sabo 1991). Direct utility costs include monies paid to customers and contractors for the purchase and installation of DSM measures. Indirect utility costs are all other utility costs, including funds for inhouse staff, marketing, evaluation, consultants, etc. Nonutility costs are the monies paid by customers and other involved parties, *after* crediting any utility incentive.

Direct costs are usually the easiest to obtain because they are commonly tracked in the program's database. Indirect costs may be tracked at the individual program level or may be tracked at the department level. When costs are tracked at the department level, they should be allocated to individual programs. In reporting indirect costs, indicating which costs are included (e.g., staff, marketing, evaluation, etc.) and which, if any, are not included is important. Nonutility costs (e.g. net customer costs to purchase and install a measure) have traditionally not been tracked by utilities. However, because these costs are needed to calculate cost-effectiveness from the societal perspective, some utilities now collect these data. Customer costs are commonly obtained in one of three ways: requiring customers to submit invoices, asking customers to report their costs, and compiling estimates of typical customer costs. The first method is accurate, but places an extra burden on customers. The second method is less accurate but less burdensome. The third method usually involves detailed data collection for a sample of customers and extrapolation to the entire population of participants. This method can be accurate if samples are carefully drawn and accurate information is obtained from customers. In reporting data on customer costs, reporting the method used to compile the data is important.

Program Savings

Data on kWh and kW savings are needed for analysis of program savings and cost-effectiveness. However, kWh and kW figures are difficult to interpret unless savings data are referenced to preprogram-consumption data. For example, savings of 1000 kWh/year, commendable for a customer using 3000 kWh/year, are trivial for a customer using 1,000,000 kWh/year. The most common way to reference savings is to calculate savings as a percentage of the average preprogram, whole-building electricity use or demand of participating customers.

In reporting savings, it is important to distinguish and report total savings and net savings (Chapters 1 and 4). Energy savings may be based on engineering estimates or on statistical analysis of metered energy use and demand (Chapter 4). In reporting savings estimates, noting which method(s) was or were used to develop the estimates is important. Also, for statistical analyses, reporting information that indicates how accurate the savings estimates are likely to be is useful. Perhaps the most useful measure of statistical accuracy is the CI. Unlike other statistical measures, such as significance levels, *t*-values, or SD, CIs are easy for nonstatisticians to understand. When reporting the CI around a savings estimate, the confidence level (e.g., 90%) should also be reported (Table 18).

Demand savings vary by the hour of the day and the day of the year (Chapter 7). Rather than reporting 8760 hourly values for each year, evaluation reports usually provide more limited data, such as demand savings coincident with a utility's summer and winter peaks.

Table 18. Electricity savings from the Northeast Utilities EnergyCHECK program^a

Annual electricity use and savings (kWh/ft ²)	Participants (n = 54)			Nonparticipants (n = 75)		
	Mean use	95% CI	SD	Mean use	95% CI	SD
Preimplementation use	12.73	±3.06	11.49	12.81	±4.35	19.23
Postimplementation use	12.48	±3.12	11.70	13.79	±4.53	20.02
Change in use	-0.25	±0.88	3.31	0.98	±0.76	3.38
Net program impact (difference between participants and nonparticipants)	-1.23					

^aThis table shows net and total savings and includes data on sample size, CI, and SD. Although the table does not show percentage savings, data on preprogram electricity use is provided, permitting percentage savings to be easily calculated. Data is reported on a per square-foot basis because participant buildings were slightly larger than buildings in the comparison group.

Source: Peters, Oswald, and Horowitz 1990.

Some utilities report maximum demand savings or average demand savings. Maximum demand savings is the largest demand reduction achieved by a measure, although this reduction is not necessarily achieved at the time of the system peak. Average demand savings are kWh savings achieved by a measure divided by 8760 hours (the number of hours in a year). For example, replacing a 40-W lamp with a 34-W lamp has a maximum demand savings of 6 W. If 70% of the lamps are on at the time of the summer peak, coincident summer peak savings are 4.2 W [6(0.7)]. If the lamp operates for 3500 hours each year, average demand savings are 2.4 W [6(3500/8760)].

Many cost-effectiveness models separately examine program costs and savings by season (e.g., summer, winter, and spring/fall) and by time of day (e.g., peak, off-peak, and shoulder periods). These models generally require estimates of *average* demand savings during each of the periods involved. In reporting these values, which months, days, and hours are included in each period should be indicated.

Even more information on demand savings can be provided in graphic form. For example, demand savings can be plotted as a function of time of day or day of the year. Time-of-day graphs typically provide hourly information. Often, graphs for several days are provided (e.g., for the peak day and for an average day during the summer and winter). Graphs of this sort, while useful for all programs, are particularly useful for load-management programs because such graphs show not only the peak demand savings but also the increases in demand during other periods (Fig. 9).

Reports of kWh and kW savings should indicate whether savings are at the customer or generation level, and if the latter, what transmission and distribution losses and reserve

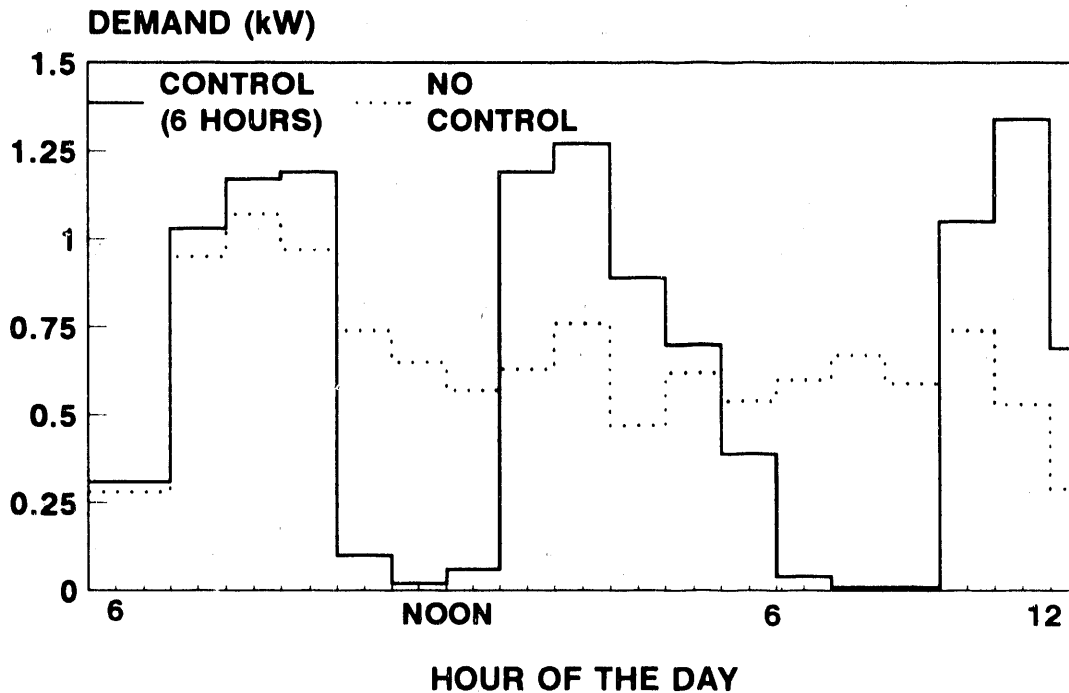


Fig. 9. Hourly load profiles for residential customers participating in a water-heater-control program run by Massachusetts Electric (1991). The solid line shows demand on days when the water heaters were turned off from 10 am to 1 pm and again from 6 pm to 9 pm. The dotted line shows demand for the same households on days that the water heaters were not controlled.

margin values were used to calculate the generation-level savings estimates. Transmission and distribution losses are typically 5 to 10% of the total energy and power generated. Reserve margins are typically 20%. Customer-level savings are needed to calculate cost-effectiveness of measures from the customer perspective (discussed later in this chapter). Generation-level savings are needed to calculate the total kWh and kW impacts at the utility level. Either type of savings can be used to calculate program cost-effectiveness from the utility and societal perspectives if avoided costs are calculated at the same level (customer or generation) as the energy savings estimates.

Annualized savings are savings achieved during a 12-month period. If savings are measured for a shorter period, adjusting the figures to estimate what the savings would be over an entire year may be useful. Examples of such situations include cases when less than a year of electricity consumption data is available or when measures are installed midway through a year. Although using annualized data is usually simpler, actual savings are needed in some cases (e.g., to estimate the lost revenue attributable to DSM programs during a specific period). In these cases, both annualized and actual savings estimates should be reported.

In summary, energy and demand savings should be presented in both absolute and percentage terms. Each savings result should be clearly marked as to whether it is net or total, statistical or engineering, customer-level or generation-level, and annualized or not.

In reporting demand savings, summer and winter coincident peak savings should generally be reported. In addition, maximum savings, average annual savings, and average savings during specific seasons are often useful.

Measure Life

The life of DSM measures, an important variable for determining program cost-effectiveness, is also critical for capacity-planning purposes because, when measures wear out, they must be replaced by new resources (Chapter 6). Estimates of the effective life of each DSM measure promoted by a program should be reported along with the source of each estimate.

Program Cost-Effectiveness

Cost-effectiveness is usually assessed by comparing the benefits and costs of a program. If the ratio of benefits to costs is greater than one, a program is cost-effective. To standardize these calculations, the California PUC and the California Energy Commission (1987) developed several tests to calculate the benefit/cost ratio for DSM programs. These tests examine program cost-effectiveness from the perspectives of participants, nonparticipants, the utility, all parties (called the total resource cost test), and society (the same as the total resource cost test with the addition of quantified externality costs). Krause and Eto (1988) explain and demonstrate the use of these tests.

Evaluation reports should include the results of these five tests for each program analyzed. Benefit/cost ratios can be calculated on a cumulative basis (since program inception) or for the most recent year. In addition to providing benefit/cost ratios, evaluation reports should include an explanation of the key assumptions used in the calculations, such as discount rates, avoided costs, and environmental externality values.

Costs per kW and kWh provide a quick approximation of program cost-effectiveness, although these indices cannot substitute for a detailed benefit/cost analysis. Costs per kW are calculated by dividing program costs by kW savings. To be most useful, these calculations should include both direct and indirect costs. Savings should be estimated for the time of the system peak. Costs per kWh, also called the cost of conserved energy, are generally calculated on an average basis over the life of the measures installed. This is done by levelizing, over the average measure life, the annual program costs with a discount rate equal to the utility cost of capital. The cost of conserved energy is the annualized payment divided by the kWh savings in one year (Meier 1982). For example, if a program costs \$1 million, saves 4 million kWh annually, has an average measure life of 10 years, and a 6% real cost of capital, then the annualized program cost is \$135,868, and the cost per kWh is 3.4¢/kWh (\$135,868 divided by 4 million kWh).

Costs per kW and kWh can be compared to reference values based on utility avoided costs (¢/kWh) and to the cost of new power plants (\$/kW). Whenever utility DSM-program costs are substantially less than these reference values, the program is likely to be cost-effective. Of course, a definitive estimate of cost-effectiveness requires a full benefit/cost analysis. In particular, using the ¢/kWh index does not give any credit for kW savings (and vice versa).

Programs that are not cost-effective based on kW or kWh savings alone may be cost-effective when both benefits are included. The costs for peak-clipping programs should be compared to the cost of new peaking power plants (in \$/kW), while the costs of energy-efficiency programs should be compared to the cost of new baseload power plants (in ¢/kWh or \$/kW).

Process-Evaluation Results

Process-evaluation results may be reported in a separate process-evaluation report, or they may be one element of a larger evaluation report. In either case, process-evaluation results should be reported in the context of other evaluation findings. That is, where appropriate, the process evaluation should address issues raised by other evaluation activities and vice versa. In this manner, the results of different evaluation components, when taken together, provide a complete, integrated evaluation (Chapter 9).

Summary and Recommendations

The summary may be the most important section of an evaluation report because utility executives and PUC commissioners may read only that part (Exhibit 7). In addition, many people will read the summary before deciding whether to read additional sections of the report. Thus, the summary must include key evaluation findings and recommendations. The summary should be written for senior management and should strike a balance between brevity and thoroughness. Recommendations and action items should be prominently featured. The summary should interpret evaluation findings in light of short- and long-term program objectives. For example, how do participation rates compare to estimates made at the beginning of the year? In addition to providing a written summary, summarizing data on the program (e.g., costs, savings, and participation rates) in a single form may be useful to readers so they can readily find the data they need (Hirst and Sabo 1991).

Recommendations from an evaluation usually fall into two general categories: ways to improve the program being evaluated and follow-up work to provide more and better evaluation data. In reporting recommendations, each recommendation should be highlighted so that managers can quickly find the recommendations that interest them. The accompanying text should describe and summarize the rationale for each recommendation, and also identify who (e.g. program manager or evaluation supervisor) should be responsible for reviewing and implementing each recommendation. If evaluation recommendations are acted upon before the evaluation report is completed, the report should indicate the disposition of each recommendation. More often, the disposition of recommendations will have to be reported in subsequent evaluation reports, and the disposition of previous recommendations reported in the current evaluation report.

USING EVALUATION RESULTS

Evaluation results can be used for both internal (within the utility) and external purposes. In this section, we discuss many of these uses and conclude with a brief discussion on internal versus external reporting.

Exhibit 7. Evaluation summary for the C&I lighting program of NEES (1988)

This summary describes the program, participation, savings, costs, and cost-effectiveness. These findings are interpreted in light of long-term goals. Several recommendations are also discussed. This summary has only a few weaknesses: (1) participants are reported in terms of number of applications, not number of customers; (2) participation rate is not reported; and (3) the type of energy savings reported is not described (in this case, net generation-level savings based on engineering estimates; demand savings are coincident with the summer peak).

The C&I Lighting Rebate Program promotes the installation of energy-efficient lighting products through dealer incentives. Dealers are paid rebates for sales of qualifying products to customers of NEES's affiliated retail companies. Products currently eligible for the program are energy-efficient fluorescent lamps, ballasts, fixtures and reflectors, and high-intensity discharge and compact fluorescent lamp retrofits. 1987 was the start-up year for the program.

In 1987, 104 dealers (out of a target group of 179 dealers) submitted 1449 rebate requests totaling \$516,696. Of the money rebated, 36% went to fluorescent lamps, 11% went to fluorescent ballasts, and 52% went to fluorescent fixtures. The majority of rebate requests came from 18 dealers. Nearly half the customers listed on rebate requests come from the large C&I revenue class, despite the fact these customers account for only 6% of NEES's C&I customers. For the program to be successful, more dealers need to participate in the program, and participating dealers need to handle an increased volume of rebates. One major target of these efforts should be small and medium-sized C&I customers.

Energy savings from equipment sold under the program in 1987 are estimated to be 1.3 MW and 4.8 GWh per year. These savings figures have been adjusted to eliminate estimated savings by "free riders" — customers who would have purchased eligible equipment even if the program were not offered. Free riders are estimated to account for approximately 45% of fluorescent lamp rebates and 50% of fluorescent fixture rebates.

The program is functioning smoothly. Dealers appear to be having little difficulty with the rebate forms, and rebate checks are being issued promptly. Overall, a survey of dealers indicates that the majority of dealers are very satisfied with the program. Only 6% of the dealers surveyed were dissatisfied with the program.

Based on the program's performance in 1987 and based on projections of future program participation levels and costs, the program is projected to just meet its target of 57 MW of peak demand savings. Under these projections, the program's cost-benefit ratio is estimated to be 0.21. Even if participation levels are half of projections and the number of free riders increases significantly, the program still has a cost-benefit ratio of 0.50 or less. However, if participation levels are low or free riders increase, the program will not meet its peak demand savings target. In order to allow for these contingencies and still meet the program's demand savings goal, it may be desirable to add additional types of products to the program (beyond those added in December, 1987) and/or to take steps to increase savings from products currently eligible for the program.

Internal Uses

Internally, evaluations are used for many purposes, as illustrated in Table 19.

Table 19. Internal uses of evaluation results

Use	Example(s)
Program modification	Process evaluation shows how to attract new market segments to a program. Impact evaluation shows that some measures save less energy than expected.
Resource planning	Impact evaluation shows the likely future effects of programs to be included with power plants in utility resource plan. Process evaluation identifies role of trade allies, leading to creation of a new program.
Load forecasting	Evaluation data is used to strengthen behavioral relationships within forecasting models. Impact evaluation results are used to adjust forecasts for future effects of utility DSM programs.
DSM bidding and performance contracting	Impact evaluations of utility-run programs provide a yardstick against which to compare bids. Evaluations suggest which programs a utility should run itself and which ones to operate through contractors.
Rate design	Time-of-use response to interruptible and other rate programs can be used to design new rate programs.
Market research	Process evaluations identify differences among market segments in participation, satisfaction with the utility, its programs, and other characteristics.

Program Modifications. Evaluation results can be used to modify programs. For example, as a result of an evaluation of a pilot program, a utility might decide to implement a full-scale program, incorporating program modifications suggested by the evaluation. Process evaluations often include recommendations for program modifications that can be quickly assessed and implemented.

For example, BPA ran a pilot program in which industrial customers submitted proposals for partial funding to install DSM measures. Measures were required to have a payback period of three years or more. A process evaluation found that the three-year payback criterion exceeded the investment threshold of many firms. The evaluation also found that, because of the competitive nature of BPA's selection process, many firms were unwilling to invest in proposal preparation unless projects were already under consideration at the plant, which meant that a high proportion of free riders was likely. Also, the two-month period during which proposals were accepted did not correspond to the capital-budget cycle at most plants (Peters and Gustafson 1987). BPA subsequently changed the incentive to a ¢/kWh-saved basis and allowed proposals to be submitted at any time (BPA 1988).

In addition to adopting short-term recommendations, evaluation results (e.g., participation rates and savings per customer) should be compared to program expectations. If results are significantly less than expectations, further analysis is needed to decide whether the initial expectations were reasonable and if program performance can be improved. For example, an evaluation of BPA's Institutional Buildings Program found that savings were 40% less than prior engineering estimates. Likely reasons for the discrepancy were identified, and suggestions were made as to how the program could be modified to improve future savings (Keating and Blachman 1987). Several of these recommendations were incorporated into the program, and many of the remaining recommendations were incorporated into new programs subsequently developed by BPA (Keating 1991).

Evaluation results should also be compared to studies of the conservation potential within a utility service-area. For example, if a conservation-potential study indicates that 500 MW of savings are available from motor-efficiency improvements but motor savings after two years of program operation are only 20 MW, then additional steps may be needed to tap more of the potential. This happened in NEES's initial motor program; low savings relative to the potential led to a complete revamping of the program (Stout and Gilmore 1989). Following the program changes, participation rates increased substantially (Nadel 1990).

Utilities should compare evaluation results with results from similar programs operated by other utilities. Comparing results for several different indices is useful; such indices include participation rates, percentage kWh and kW savings, free-rider proportions, ratio of indirect to direct costs, cost per kWh saved, cost per kW saved, and the ratio of evaluation estimates of energy savings to engineering estimates. If these indices are calculated with the procedures discussed above and in Nadel (1991) and Hirst and Sabo (1991), they are usually comparable from program to program. If an index value for a particular program is better than that for other similar programs (e.g. high participation rate), it may indicate that a program is going well. If an index for a program is below average, it indicates that additional research is needed, first to see if there is a reasonable explanation for the poor index value and second to see if program improvements are justified to improve the index value.

Resource Planning. Evaluations of existing programs often provide useful information for planning new programs and for preparing long-range resource plans. To illustrate, Nadel and Tress (1990) developed long-range DSM plans for three New York utilities. For each

of the 21 programs examined, estimates of participation rates, utility and customer costs, savings per customer, and measure life had to be made to compute energy savings, costs, and cost-effectiveness. Wherever possible, these estimates were based on the results from evaluations of successful utility programs.

Forecasting. Evaluations can be used to improve utility load forecasts by providing data needed for forecasting models and by helping to estimate the impact of DSM programs. Most utilities incorporate DSM measures into their forecasts using a two-step process. First, baseline energy-use and demand forecasts are prepared that incorporate the effects of DSM actions taken by customers in the absence of utility programs. Second, the impacts of utility-operated programs are specifically modelled and subtracted from the baseline forecast.

In computing baseline energy use, future efficiency trends in the absence of utility programs must be estimated, including the impacts of market forces and government efficiency standards. Data collected for evaluations can help in this regard, particularly data on free riders and energy-use trends among nonparticipants. For example, baseline data collected to evaluate an appliance rebate program can be used to estimate the energy consumption of new residential appliances. Similarly, baseline data collected for a commercial new construction program can be used to estimate the energy use intensity (kWh/ft²) of new buildings. These values are required inputs for many forecasting models. Some forecasting models include consumer-choice algorithms in which adoption of DSM measures is modeled as a function of measure costs, savings, and current stage on the technology-diffusion curve. Evaluation results can be useful in estimating these model coefficients.

DSM Bidding and Performance Contracting. Many utilities contract with energy service companies (ESCO) and other contractors to provide energy savings. In such cases, the utility must determine whether the energy savings and cost estimates provided by prospective contractors are reasonable. Also, the amount of energy savings actually achieved must be estimated, particularly if payments to contractors are based on kW and kWh savings. DSM evaluations can help with both tasks. Evaluations of existing programs provide useful yardsticks for assessing the reasonableness of bids and proposed contracts, and past evaluation can suggest cost-effective ways to measure the energy and demand reductions provided by contractors.

Rate Design. Data collected on electricity demand during different hours of the day and different seasons can help in the design of time-of-use and interruptible rates. An evaluation of NEES's Stand-By Generation Program (which provided payments to customers to operate emergency generators during peak periods) provided insights on how to improve program participation and savings. These results were subsequently used to revise the rate schedules for the company's Stand-By Generation and Interruptible Service Programs (Nadel, Cress and Ticknor 1989). Similarly, Southern California Edison used performance data on storage cooling systems to design special time-of-use rates to encourage such systems (Nadel 1990).

Market Research. Evaluations collect information on the market segments targeted for the program, including data on customer characteristics and preferences. For example, Southern California Edison maintains a database of market-research information on its largest

customers. As evaluations are conducted on programs serving these customers, additional information for the database is likely to be produced.

External Uses

Regulators use evaluations in their review of DSM plans and performance. Intervenors in the regulatory process use evaluations for the same reasons. In addition, DSM-program planners and implementors from other utilities use such results to learn from the experiences of others.

Regulators may want to review evaluation plans prepared by utilities for reasonableness (are the approaches proposed in line with modern evaluation practice?) and thoroughness (are all needed data included?). Such reviews help ensure that the information regulators need will be collected, reducing the chances that disagreements on procedural issues will hamper the recovery of DSM-program costs. For these reasons, the Connecticut and Massachusetts PUCs require utilities to submit detailed evaluation plans for approval.

Regulators might require utilities to collect and report specific types of evaluation data. For example, the Massachusetts Department of Public Utilities developed a specific format for annual reports of DSM results (Hirst and Sabo 1991). Regulators might even specify specific evaluation methods. For example, the New York PSC directed its utilities to work with commission staff to develop a standard method to evaluate three DSM programs mandated by the PSC (XENERGY 1990).

Of course, PUCs will review the results of evaluations. Typically such reviews are ad hoc; commissioners and commission staff read evaluation reports and ask questions. One approach to a more systematic review is for regulators to compare evaluation results to utility short-term and long-term goals, estimates of the available conservation resource, and results of similar programs operated by other utilities. As discussed earlier, using several indices can make this task easier (Nadel 1991).

Occasionally, when regulators are dissatisfied with utility evaluations, or when they want an independent review of evaluation results, they may hire an evaluator to conduct such a study. For example, the Wisconsin PSC has done this on a few occasions (Nichols et al. 1990).

Finally, regulators can use evaluation results to help guide cost-recovery determinations for DSM programs, including recovery of program costs, lost revenues (reductions in revenue collection due to reduced kWh sales caused by DSM programs), and financial rewards for successful implementation of DSM programs (Chapter 1). Evaluations are generally the source of program costs and savings estimates upon which cost recovery is based.

Typical intervenors in PUC proceedings include public interest groups, state consumer counsels, and large industrial customers. Each of these parties will selectively choose evaluation results that reinforce its own case. For example, in a recent proceeding before the District of Columbia PSC, witnesses for the local consumer council, citing a recent review of DSM-evaluation results, argued that Potomac Electric Power Company's proposed DSM

programs had significantly lower participation rates and savings than the most successful programs offered elsewhere in the country and that the utility should substantially increase its targets for DSM savings. Witnesses for the utility, citing the same study, argued that proposed programs had participation rates and savings near the industry average, and hence its targets were reasonable. In this case, use of evaluation findings allowed the different parties to agree on the facts (the utility's targets were near the average in the industry, and not among the most ambitious) but to disagree on what actions to take. The agreement on the facts made it easier for the PSC to make a decision in the case. The commission encouraged the utility to exceed its proposed savings targets, but did not require it to do so (District of Columbia PSC 1991).

Evaluation results can be used by other utilities, ESCOs, and state energy offices. Unfortunately, obtaining evaluation data from others is not always easy, because most evaluation results are not widely distributed or publicized. Several compilations of evaluation results have been published (Nadel 1990; Blevins and Miller 1989a and 1989b; Northeast Region Demand-Side Management Data Exchange 1989; Keating and Hicks 1990; Energy Program Evaluation Conference 1991). These publications can be useful but suffer from limited availability (the Northeast Region Demand-Side Management Data Exchange data are generally available only to members), limited coverage (only some programs are included), missing data, and time lags between updates.

Internal vs. External Reporting

Because evaluations are used both internally and externally, the question inevitably arises whether internal results should be reported externally. A number of considerations affect this determination. First, in many jurisdictions, once results are put in writing, even in an internal memorandum, regulators and intervenors can request the information during the next regulatory proceeding for which the data are relevant.

Second, if information is withheld, such as disappointing results for a particular program, and regulators discover it, the ramifications are likely to be far worse than if the results were voluntarily reported in the first place. An open sharing of information with regulators is likely to improve a utility's credibility, resulting in long-term benefits to the utility.

Based on these considerations, internal results should generally be reported externally. The advantages of external reporting generally outweigh the disadvantages. Although most results should be reported externally, they should first be reported internally, starting with staff who worked on the program that was evaluated and then proceeding to utility management. Such a successive reporting strategy allows errors to be caught before results receive wider distribution. Also, a successive review process allows each reviewer to consider the findings in private before making a public response. With this type of process, program staff and utility management are less likely to be defensive about evaluation recommendations and are more likely to accept recommendations that make sense.

Also, some information is particularly sensitive, with little to be gained by public reporting. Information that falls into this category can include the names of employees, customers, and

trade allies who made mistakes. Such information is of little concern to others, and by publicizing the information, much animosity is generated that may cause more harm than good. This type of information should generally be reported verbally only to those people within the utility who can constructively use the information (McRae 1990).

CONCLUSIONS

A thorough evaluation requires collecting, analyzing and reporting a substantial amount of data, including process, impact, and tracking information. This information should be presented in a clear, well-organized manner, so that those who use the report can easily find specific information. When information is reported, the meaning and derivation of each data element should be clearly described (for example, that the savings reported are net savings at the customer level, as determined with conditional demand analysis).

While much is known about how to report and use evaluation results, additional work still needs to be done. Some of the definitions for specific types of evaluation data are imprecise (e.g., program participant and eligible customer) and could benefit from standardization. A recent study on terminology and reporting formats should help substantially in this regard (Hirst and Sabo 1991).

When evaluation data are reported, they often appear in utility or consultant reports that are not widely distributed. To make this data available to a wider audience, an index of evaluation reports should be prepared and regularly updated. Logical organizations to undertake such an effort include the national laboratories or EPRI. Furthermore, existing databases of evaluation results should be expanded to include more programs and data and should be regularly updated. To address this need, Lawrence Berkeley Laboratory is planning to develop and regularly update a database of DSM results (Association of DSM Professionals 1991). In addition, Hirst (1990b) proposed that EIA collect and compile annual data on individual DSM programs just as they presently collect annual data on individual power plants. EIA has begun to collect limited data on utility DSM efforts, but has yet to compile data on specific DSM programs.

PERSISTENCE OF ENERGY SAVINGS

Kenneth M. Keating

INTRODUCTION

Persistence of the energy and load reductions from DSM programs is a crucial issue for program planners and a difficult one for evaluators. The planned value of savings from these programs, and hence the cost-effectiveness, depends on the continued impact of the program over the projected life of the program measures. Unless the savings from the program continue, the alternate resources deferred by the program will be needed sooner than expected. Without persistence, the DSM resource loses its long-term value. Put simply, if it is not there when you need it, it is not worth much.

In the initial development of DSM programs, persistence, a nagging issue for electric-system planners, was assumed by program operators to be reasonably constant. In many instances, experience with DSM measures was lacking, making it difficult to determine the effective life of a measure. Most planners assumed that knowing the physical life of the measures installed was sufficient to determine persistence. As DSM programs have matured and as the money spent on such programs has increased dramatically, utilities and regulators have begun to look closely at persistence. Evaluators are being asked to validate the persistence assumptions of program planners as part of integrated resource planning.

Although early evaluation reports suggested a slight decline in the net program savings of some programs (Hirst and Keating 1987), limited data supported changes in the dominant assumptions of basic persistence. Now, program and evaluation experience indicate that persistence is a pressing and difficult issue. DSM-program managers can no longer install a measure, declare victory, and walk away.

UNDERSTANDING THE ISSUES

To answer the question "Do the impacts of DSM last?", evaluators need to be clear about which impacts are being examined. Persistence has two dimensions. One dimension concerns the lifetime of the DSM measures installed by a program, the manner in which these measures are operated, or both. The second dimension concerns the overall definition of persistence. Defined at the utility level, persistence is equivalent to the long-term temporal changes in net program impacts; defined at the societal level, it is equivalent to the

long-term temporal changes in total impacts. Note that the persistence of both total and net impacts includes technical and operating characteristics.

The net-program perspective requires that persistence be measured with an eye to what the participants would have done in later years had there been no program. If nonparticipants in the comparison group adopt energy-efficient practices and measures, this perspective in effect reduces the value of the savings in later years. It reflects the probable actions of the participants had the program not existed. This concept is theoretically appealing, but the results obtained with a comparison group can be confounded by spill-over effects from the program. In essence, nonparticipants may adopt program-recommended actions because of the program even though they do not formally participate in the program; see the discussion of free drivers in Chapter 8.

In considering total impacts, what others are doing is irrelevant as long as the measures in the participants' facilities continue to perform over time as they did when first installed. The key issues here are the technical performance of the DSM measures and the way that these measures are operated and maintained. These technical- and operating-performance estimates are relative to the performance of the conventional measures replaced. Consider a situation in which electricity use for a high-efficiency refrigerator increases by 5% after two years because the householders did not clean the compressor coils. If the same increase in electricity use would have occurred with a conventional refrigerator, no loss of savings is associated with the high-efficiency unit.

Figure 10 illustrates the different perspectives for a residential retrofit program. In this example, the program's net impacts decline over time as the general population adopts more-efficient measures and practices. From the "efficiency" perspective, some measures deteriorate (e.g., deterioration of weatherstripping after five years, reducing savings until it is replaced in year nine). In either case, the amount of savings is the area under the respective curves, not the first-year savings continuing for 20 years.

Technology and behavior affect persistence. These two factors interact and are often difficult to separate. Examples of technology failure include equipment malfunction, breakdown, or loss of efficiency over time. Behavioral issues include removing measures, failing to provide maintenance, or overriding controls. Technological and behavioral aspects interact when measures, such as efficient lights, are removed because they were inappropriately recommended and provide insufficient light or when the high-efficiency technology does not produce the requisite comfort. For example, people may remove low-flow showerheads if they do not like the "feel" of the shower they get with such units. It is also possible that a manufacturer could increase production *as a result of* having a more efficient operation. This behavioral response is called takeback and will erode net program impacts. Figure 11 illustrates the various threats to persistence discussed here.

Program managers should be concerned about the actual life of many commercial-sector technologies (and some residential ones also, such as low-flow showerheads and compact fluorescent lights). DSM-program planners often base estimates of measure lives on

SAVINGS PER PARTICIPANT

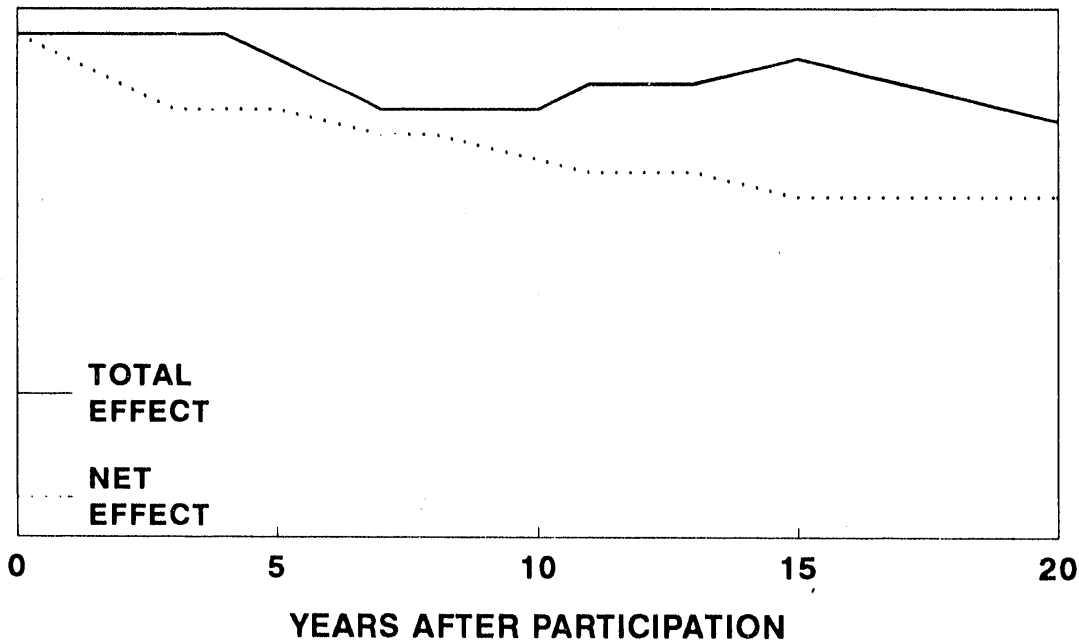


Fig. 10. Schematic of persistence of energy savings from the total (societal) and net (program) perspectives. In this example, net savings decline with time as nonparticipants, driven by increasing electricity prices, adopt some of the measures that participants previously adopted. The total savings decline when measures deteriorate and then increase when these measures are replaced.

manufacturers' tests or laboratory tests. These estimates often ignore the effects of repeated cycling, improper maintenance, and other factors that might affect the lifetime and performance of the measure. These estimates are referred to as "test measure lives." However, the appropriate standard for estimating program expectations is "effective measure lives," the amount of time that half of the measures remain in place and are effective (the median lifetime) and half are gone.

The differences between these two concepts are not just academic. A survey of manufacturer representatives, maintenance staff, and DSM experts on estimates of the effective lives of 52 DSM measures gave responses that were usually less than the manufacturer specifications (Gordon et al. 1988).

Effective life involves more than the physical life of the equipment (Exhibit 8). Conservation measures can fail because they are not maintained properly, they are poorly designed, or they are overridden or disabled. For measures installed in commercial buildings, the primary reason for lack of persistence may be their removal during renovations. For example, electronic ballasts may last 20 years; however, that does not benefit the DSM program if they are put in a building that changes tenant or function and the ballasts are removed and

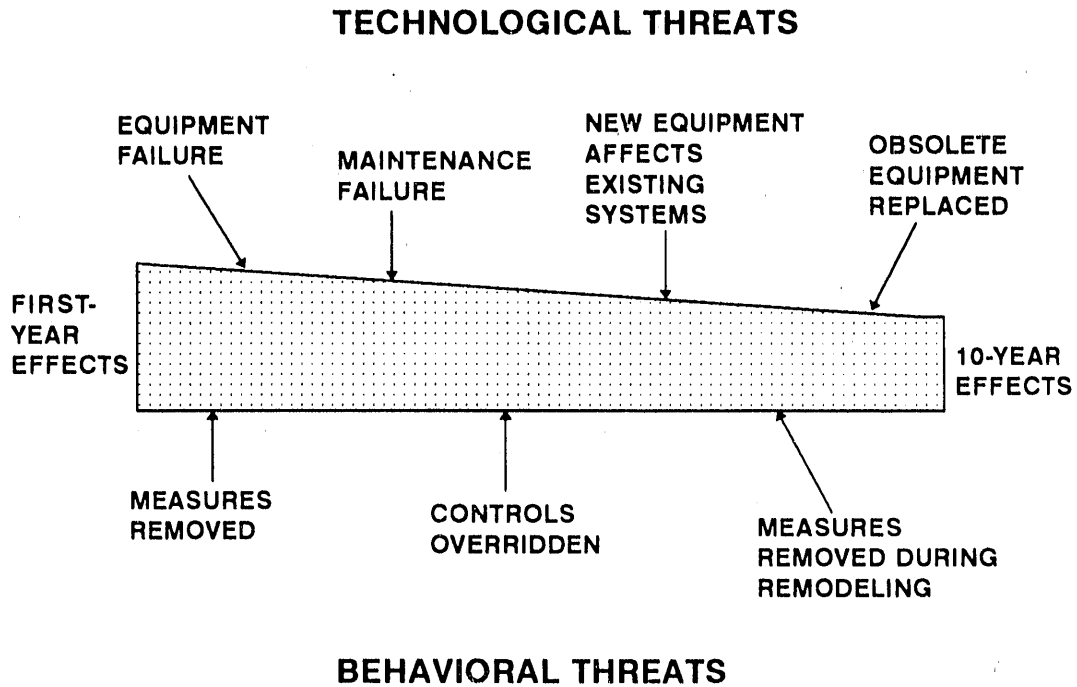


Fig. 11. Technological and behavioral factors that affect the long-term energy savings of DSM programs.

replaced with less efficient fixtures in a few years. Data show that commercial buildings typically undergo some type of remodelling or renovation every few years (Petersen 1990; Petersen and Sandler 1991; Puget Energy Services 1991; Kunkle and Johnson 1991; Hickman and Steele 1991).

Persistence is not an either-or question; it is a matter of degree – how much of the impact remains over time. DSM programs install a mix of measures in a variety of buildings. Some measures in some buildings will fail or be removed. However, program evaluations deal with aggregates of buildings, and although technological or behavioral failures may dramatically affect savings for individual buildings, their effect may be limited for an entire DSM program.

Exhibit 8. Frequent changes in commercial buildings shorten effective lives for many DSM measures.

In an extensive research effort for BPA, 100 professional remodelers were asked to estimate the frequency of remodelling in different types of buildings, and which end uses were likely to be affected. In addition, the researchers visited 300 commercial buildings that had undergone remodeling or renovation to find out what had been changed, and what happened to the old equipment (Skumatz et al. 1991). The results suggested that the effective life of many common DSM measures was only half that of the manufacturer estimate. For example, lighting fixtures might be replaced every four or five years in small office buildings and restaurants.

METHODOLOGIES AND ISSUES IN PRIOR RESEARCH

Limited research is available on the persistence of DSM savings, but horror stories abound: power outages causing electric timeclocks that control water heaters to lose so many hours that the heating elements are on at system peak; floor insulation being ripped out to fight termites; or pipes freezing, cracking, and soaking the insulation in walls. DSM opponents point to these occurrences as proof of the ineffectiveness of DSM measures. DSM proponents argue that each of these glitches lost only a tiny amount of a diverse resource, compared to a loss of 800 MW caused by a bad valve at a power plant. Both sides are right. Perfect persistence is an impossibility. However, the erosion of savings may be modest. A key evaluation issue concerns how savings vary with time. A critical planning issue concerns how to design programs so that their impacts are more persistent.

Two general approaches to research on persistence have been used:

- The total-impact perspective is associated with the study of measures in place; follow-up site visits or surveys to verify continued installation and proper functioning of measures have been the emphasis in several studies.
- The net-impact perspective is associated with analysis of electricity-billing records of participants and those of a comparison group of similar nonparticipants. This approach may not work well for large C&I customers because a suitable comparison group may not exist.

Studies of Measures in Place

Persistence of efficiency impacts may be tracked by verifying the continued presence and proper operation of program-installed measures when the measures operate on a known and unchanging schedule (e.g., motors on 24-hour operation, exit lighting, and street lighting), when there are few interactions with other energy-using systems or occupants, and where the savings are well established (e.g., those of water-heater tank wraps or low-flow showerheads).

Surveying samples of participants is an inexpensive approach for tracking some measures. Central Hudson Gas and Electric conducted a short-term study of persistence of WH efficiency improvements. In its household surveys, between 87% and 94% of the measures installed were reported to still be in place up to 29 months after installation (TechPlan 1990). The measures included tank wraps, low-flow showerheads, and pipe insulation.

A two-step survey process was used to study the persistence of residential lighting measures in New England. In a process evaluation, participants reported the removal of 12% of the high-efficiency compact fluorescent bulbs within seven months of their installation (Applied Management Sciences, Inc. 1990b). A year later, NEES conducted follow-up site visits to 96 of these low-income homes. The company found that 25% of the bulbs had been removed, representing 20% of the displaced wattage. Most occupants removed the bulbs because they were dissatisfied with the quality, color, or quantity of light (Tolkin 1990). This erosion of savings was caused by the interaction of behavior and technology. These results

do not necessarily mean that compact fluorescent lamps will always show a loss of savings; the results may provide evidence of potential weaknesses in program design, measure selection, or program delivery.

A more expensive two-part study was undertaken on buildings that participated in the BPA Institutional Buildings Program in Washington State. A program engineer visited 49 sites (74 facilities) three years after the buildings were retrofitted (Kunkle 1990). The second stage of the research was a telephone follow-up to the facilities one year later (Kunkle and Johnson 1991). A major advantage of the follow-up was that it tracked what happened to major systems when problems developed. In 27 facilities, there had been 41 problematic measures, centering on control systems, HVAC, and heat-recovery systems. Of the 12 problem control systems, five were disabled, three repaired, three replaced, and one was left unrepaired. Of the 12 problem HVAC retrofits, three were disabled, four repaired, one replaced, and four continued to operate inefficiently. Although the poorly performing systems were less than 25% of the systems installed (and only 5% of the lighting retrofits had problems), significant erosion of savings occurred during the first six years after installation.

Studies of measures in place face limitations. Problems arise in quantifying and extrapolating the effects on persistence when some measures are missing or when they operate less than optimally. To illustrate, if the efficient motors examined during a site visit are not being maintained properly, exactly what kWh impact should be inferred for these motors? What should be inferred for other motors installed through the program but not examined during the site visit? In addition, unless this type of verification is extended to a sample of nonparticipants, the program savings may be underestimated because no account would be taken of efficiency losses for standard equipment. Measuring persistence without a comparison group implicitly assumes that standard equipment experiences no deterioration in performance over time.

Analysis of Electricity-Consumption Data

Analysis of billing data is often associated with the net-effects perspective because it has traditionally involved the use of comparison groups of eligible nonparticipants. This approach is essential for attributing the changes in consumption to the program and for separating these effects from those caused by other (nonprogram) factors. This approach also adjusts for the effects of free riders.

This approach has been used in several studies of short-term persistence, but rarely to examine persistence for more than a few years. BPA studied three retrofit-program cohorts (1981, 1985, and 1986) for three postparticipation years (Hirst et al. 1985; Keating and Hirst 1986; Haeri 1988; White and Brown 1990). The results were mixed. The savings for some cohorts remained essentially constant, while the savings for others declined with time.

Analysis of data for three postparticipation years for participants in the Hood River Conservation Project showed a complete loss of savings. But most of the erosion was caused by the participants' switching SH fuel from wood to electricity (Schoch-McDaniel 1990). This finding demonstrates the importance of using a comparison group, which was not

possible in Hood River, where 90% of the homes participated. If a comparable group of nonparticipants had been available for analysis, they might have shown the same pattern of fuel switching because of the dramatically improved economic conditions in the Hood River area.

A parallel study of the persistence of the demand impacts (kW) of the Hood River Conservation Project indicates that the effect on weekday and annual peaks did persist extremely well for three years (White et al. 1991). These Hood River results suggest that the performance of the retrofit measures persisted largely unchanged; net savings declined because of operating changes (primarily the shift from wood to electricity for heating).

In the commercial sector, a comparison group is essential to control for the frequent fluctuations in building energy use. In particular, changes in business cycles can have large effects on energy use (Keating and Oliver 1991).

Two long-term studies of persistence used billing data from participants and a comparison group. Both involved residential retrofit programs in the Pacific Northwest, one by Seattle City Light (Sumi and Coates 1988) and one by BPA (Horowitz et al. 1991). Each followed samples of participants and nonparticipants for six years, and both weather-adjusted electricity consumption with PRISM (Fig. 12). Two results are evident: the trend in energy savings is downward, and the decline is erratic (i.e., net savings vary from year to year). The

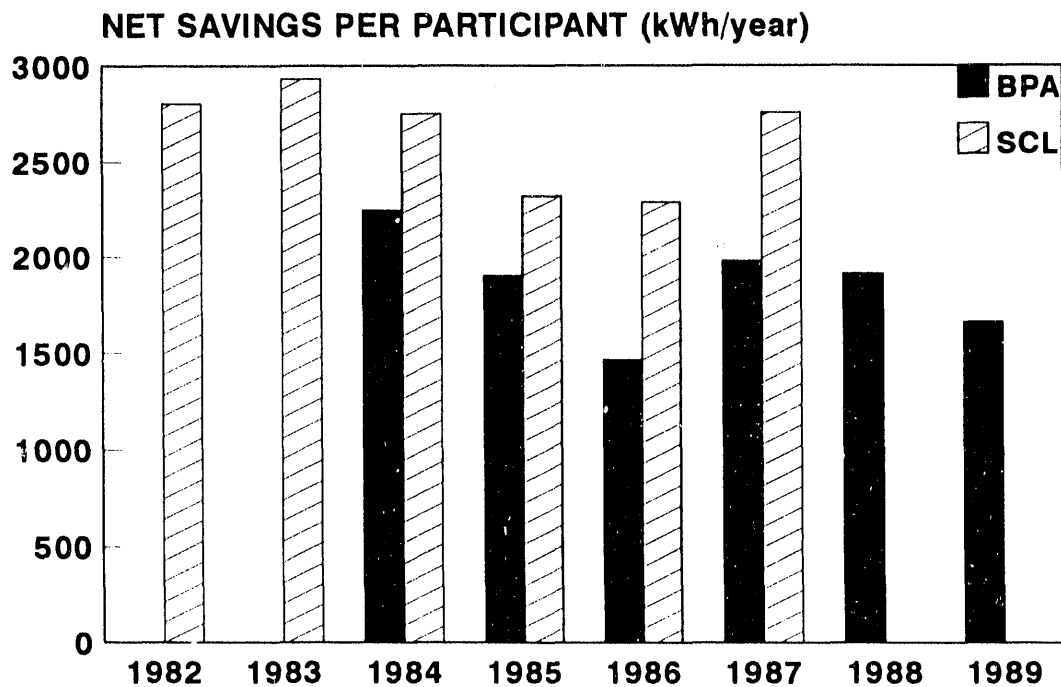


Fig. 12. Net electricity savings for several years produced by two residential retrofit programs in the Pacific Northwest.

average decline in savings for the last five years compared to the first year amounts to 7% for the Seattle program and 21% for the BPA program. These studies indicate substantial persistence, but also some erosion of savings. The eighties were characterized in the Northwest by unprecedented electricity-price increases for the first half of the decade, followed by declining real prices. Much of the erosion of savings was caused by reduced consumption by the comparison groups, about 2000 kWh during the six years studied, primarily because of electricity-price increases.

The reports from BPA and Seattle show that these data can be aggregated and analyzed many ways. For example, evaluators in Seattle were asked to provide a forecast of future changes in electricity savings, not for any particular cohort, but in general. Therefore, their report includes a weighted average savings for all cohorts for each year (i.e., the average savings in 1983 for homes weatherized in 1981 and 1982; the average savings in 1984 for homes weatherized in 1981, 1982, and 1983; and so forth). This practice reduces the year-to-year fluctuations within cohorts and shows an average erosion of savings of 27% during the six-year period (Sumi and Coates 1988).

These studies exemplify two problems with the use of billing analysis to study persistence: attrition bias and the confounding effects of free riders and free drivers. Sample attrition occurs when households move, billing data are lost, or nonparticipants become participants. Attrition bias occurs if the remaining cases are systematically different than the cases that were lost. The BPA participant and nonparticipant samples suffered an attrition rate of 55%. Six years after retrofit, they had only half as many cases with consumption data as they did one year after retrofit for that cohort. Because the first-year savings for the group with six years of postprogram data was much smaller than that for the original sample, attrition bias was apparent. The study used survey data to correct for the small portion of the bias that could be quantified.

The Seattle study, on the other hand, shows how to avoid attrition bias in a longitudinal study. Seattle reported only 10% sample attrition because Sumi and Coates (1988) used a retrospective approach. That is, they started their research in 1987 and selected their samples of participants and nonparticipants from among those for whom they had continuous billing records back to the year before program participation for the participants and to 1980 for the nonparticipants. This meant that all the people who moved or joined the program late were already excluded. Because this study examined the same households for the entire period, the results are not affected by attrition bias. On the other hand, the samples of customers that Sumi and Coates were left with were special groups and were surely not representative of all customers. Their approach, however, should be considered where possible.

Controlling for free riders and free drivers in persistence research from the net-program perspective is daunting (Chapter 8). The effect of free riders on net program savings is usually accounted for by the use of a comparison group to estimate what the behavior of the participants would have been without the program. However, what if the comparison group takes action *because of* the program? This behavior, a spillover effect of the program, is called a free-driver effect.

Free drivers are customers who adopt program-recommended actions without participating officially in the program. They may take action because

- The program changed the stock of equipment available
- The program changed perceptions among customers about the usefulness of conservation
- They know about the program, but do not want the hassles of formal participation
- The service industry they deal with is now interested in efficiency
- Competitive pressures from participants have influenced them

Free drivers are certain to accompany programs that aim to change construction practices or the basic marketplace. The dilemma is that free drivers are interpreted as free riders when billing data are used to estimate net program savings. Because this confounding is more likely to occur the longer the program exists, the issue is particularly acute for persistence research, which looks at cohorts over many years.

Potentially-Fruitful Approaches

Discussions of research approaches and their pitfalls often associate approaches with perspectives: examination of measures in place with the total perspective and billing history analysis with the net-program perspective. But research need not follow this pattern. The study of measures in place can incorporate the concerns of net program effects through two options:

- The use of parallel site visits to nonparticipating buildings to see what is happening outside the program
- The use of surveys to identify free riders and free drivers

The combination of the two options is likely to be the most fruitful approach because site visits by themselves cannot separate free riders from free drivers. However, surveys, whether alone or with site visits, are weak ways to get at customer motivation for adopting energy-efficiency actions.

On the other hand, quasi-experimental designs that use billing analyses can be used with the efficiency perspective with one important change. If researchers limit the comparison group to include only previous participants, they can take advantage of the power of quasi-experimental designs to control for changes in electricity prices, business cycles, and other factors while minimizing the possibility that the results will be affected by free riders or free drivers. Previous participants will be less likely than the general populace to adopt additional efficiency improvements, having already adopted such actions because of the DSM program. Also, self-selection bias is largely eliminated because these customers, like later

participants, chose to participate in the program. Nevertheless, sample attrition and attrition bias will remain, particularly in the commercial sector.

An, as-yet untried alternative is an econometric approach. A dynamic econometric model that defines the relationships among electricity use and electricity prices, measures of economic activity, and program participation could be developed. This model could then be used to predict what consumption would be in future years if energy efficiency were held constant and the actual values of electricity prices and other explanatory variables were entered into the model. If the consumption observed in the out-year was higher than predicted, then the efficiency effects could be said to be eroding, and the effect could be quantified.

Finally, many projects that are in idiosyncratic settings (e.g., industrial-process retrofits) will have to be studied on a case-by-case basis to measure persistence. In these situations, the evaluation may involve more than measurement. It may form a partnership with maintenance and monitoring functions so that persistence problems found in the evaluation provide feedback to the maintenance staff.

CURRENT AND PLANNED RESEARCH

Several persistence studies are under way or in the planning stage. The California Energy Commission sponsored a workshop in early 1991 to review ways to study the issue state-wide. The BPA study (Skumatz et al. 1991) reviewed above is the first step in a major study of persistence in the commercial sector. It is limited to studying only one threat to persistence, the removal of measures caused by remodellings, renovations, and business changes. If energy-using systems (such as lighting, HVAC, and motors) are frequently removed, the expected lives for efficiency measures will be much less than their physical lives. Further research will be done to determine the extent of maintenance failures, measure failures, and the importance of consumer training. In general, the amount of research devoted to persistence is expected to grow. EPRI (Hanser 1991) and the New York State Program Evaluation Task Force (Cummings 1991) each have made persistence a high priority.

PROGRAM IMPLICATIONS AND CONCLUSIONS

Concerns about persistence should affect programs in both planning and implementing cost-effective programs. Evidence continues to accumulate that program planners have generally overestimated savings (Nadel and Keating 1991). Empirical research raises questions about the validity of using manufacturers' claims for physical measure lives as a basis for projecting persistence. Planners need to become more conservative in estimating the effective life of program-induced savings.

If managers want long-term effects, they need to be concerned about the longevity of the results. One approach is to focus program resources on new construction, remodelling, and renovations. This approach allows utilities to pay a lower incentive based on the incremental

cost of the efficiency improvement (rather than the total cost, which might be paid for a retrofit). It also installs DSM measures when the length of time to the next remodelling is the longest.

A second suggestion is to make sure that the installed equipment is commissioned properly – that controls work, that systems work together rather than out of synchronization, that air flows are balanced, that the measures provide the desired amenity level, and that the occupants know how to operate the measures properly. This will reduce the likelihood that occupants will override control systems, remove measures, drop out of the program, or "bad mouth" the program to other potential participants.

A third approach is to provide continuing education and maintenance. Many times when occupancy or staff change, the new people do not know how to deal with the efficiency measures and may ignore their potential to save energy (Komor et al. 1989). Follow-up can be accomplished by making maintenance contracts a part of the measure package, by flagging vulnerable technologies for yearly follow-ups, and by providing general education for building owners and maintenance workers. Performance contracts for third-party measure installations that base payments on savings over time can be effective for some classes of buildings and measures, but it can be expensive, and it removes the customer contact from the utility. Nevertheless, it is potentially effective for targeted segments of the DSM market or as an interim approach until the utility can establish its own expertise.

Persistence is a genuine problem of undetermined scope. Its effects on cost-effectiveness, program planning, and resource reliability are clear. It is now time to address persistence in earnest.

USE OF LOAD RESEARCH IN EVALUATION

George Fitzpatrick

INTRODUCTION

Load research is the collection of electricity-usage data through a metering device associated with an end use, a circuit, or a building. The goal of this activity is to produce cost-effective, statistically sound, detailed databases for resource planning, evaluation, and rate-design studies.

Load research has been employed by electric utilities since the 1960s to better understand the characteristics of electric loads, the timing of their use, and the amount of electricity consumed by users. During the 1970s, load-research activities expanded as the needs for the data increased beyond those for traditional cost-of-service studies. As early as 1976, load-research information had become a direct input into at least one utility's load forecast (New York Power Pool and the Empire State Electric Energy Research Corporation 1976). Even earlier, class load research had been employed in the construction of weather-normalized class-by-class system peaks (Long Island Lighting Company 1972). The enactment of the Public Utility Regulatory Policies Act of 1978 spurred utilities to make additional investments and kept manufacturers of load-research equipment and consultants busy meeting the demand for load data.

There are two generic types of load-research information. Direct load-research information is the measured class or end-use demand (kW), energy (kWh), or reactive power (kvar). The sampling, collection, editing, and analysis of these measurements are the primary activities of direct load research.

Indirect or corroborative load research is the data that the load researcher uses to explain and understand direct load-research measurements. These data include appliance-saturation data, demographics, DSM-process-evaluation data, weather data, equipment-nameplate data, audit data, etc.

Load research is expensive. Therefore, care must be used in starting or expanding such efforts. The design and conduct of a load-research program requires knowledge of research objectives, sample designs, and common load-research needs across DSM programs.

LOAD-RESEARCH PROGRAM ACTIVITIES

Load-research programs have eleven major activities (Table 20), each of which is discussed below. While these activities are discussed in a linear fashion, the activities are, in fact, continuous and often take place in parallel, particularly as priorities change and new requirements arise. A successful program emphasizes planning and design while striving to maintain the efficiency of load-research operations.

Table 20. Key activities associated with load research

Program planning	Meter placement and replacement
Sample design	Installation record keeping
Sample selection and validation	Meter operation and maintenance
Recorder selection	Data collection
Meter quality control and quality assurance	Data validation and editing
	Data analysis

Program Planning

In this activity, overall program objectives are defined; the data requirements (both direct and indirect) are specified; the common needs of utility users of load data are identified; program timelines are determined; and staffing, equipment, and software requirements are specified. This activity is a vital first step, but it is also an activity that should be revisited frequently during the conduct of a load-research program.

Planning should include investigating the experiences of other utilities with load-research programs, including visits to view their load-research operations. If visits can be arranged, they should include discussion with the load-research manager and DSM-program managers using load-research data, the viewing of field installations, discussions with installers and maintenance personnel, and observation and discussion of procedures for recruiting customers, installing equipment, data collection, and data editing.

There are other important sources of information on load research. The Association of Edison Illuminating Companies conducts seminars on load-research methods for utility staff with minimal experience as well as regional conferences for advanced practitioners. The American Public Power Association Rates and Load-Research Committee may also be contacted for valuable guidance. In-house seminars conducted by consultants can address specific topics, such as analysis software, editing strategies, and development of applicable statistical skills.

Given the significant lead times required to purchase and install load-research equipment (a minimum of four to six months), the evaluator must identify load-research data

requirements well in advance of when load research begins. Load-research planners find it useful to have evaluators develop a timeline showing when different types of data are required. With this information, the load-research planner can anticipate purchasing and installation requirements. For example, if a DSM program has as its focus summer-peak reduction, then evaluation plans must be finalized by December of the previous year. Knowledge of the evaluation cycle of programs that require load research is essential for competent planning.

Sample Design

Sample design was discussed in Chapter 3. But an important distinction must be drawn between a load-research sample and a sample to be used for evaluation. In the case of the load-research sample, the researcher is interested in predicting the time-varying load or class of loads. For example, the goal might be to predict the diversified demand of air conditioners with a 95% accuracy. If the average air conditioner load is 3.25 kW and the average duty cycle at the 3:00 PM peak is 0.6, then the average predicted load would be 1.95 kW, and the sample size would be chosen to make this accurate to ± 0.09 kW.

In contrast to the load researcher, the evaluator usually is interested in the *change* in load. Thus, if the program manager's goal is to reduce AC load by an average of 10%, then the sample chosen has to be large enough to predict an average reduction of 0.195 kW with an accuracy of ± 0.009 kW, assuming the estimate is to be accurate within 5%.

A second difference between a load-research sample and an evaluation sample concerns external validity (see Chapter 2). The load researcher may only be interested in the behavior of a class of customers, such as users of central AC. The evaluator is likely to be interested in understanding the air-conditioner usage patterns of subgroups of air conditioner customers or in generalizing the results (external validity) to a larger group of customers. Thus, a simple random sample of air-conditioner users may not be sufficient for an evaluator who needs to have the sample of air-conditioner users stratified by key variables, such as total monthly energy consumption, business type, size of business, household income, size of household, type of structure, size of structure, or air-conditioner sizing.

Billing, load-research, and survey data can be used to help identify appropriate sample sizes. For example, if the target for a load-control program is customers with central AC who use more than 1500 kWh/month in the summer, sample-size requirements can be estimated for certain cycling strategies with estimated diversified demand curves developed from load-research data. End-use data are invaluable for developing baseline information for all DSM programs.

A key challenge is to develop an overall plan that integrates samples at as many points as possible to achieve the most economical pattern of metering. Sample designs can be developed by creating new samples or enhancing existing samples. For example, a utility's in-place class load-research program might be modified and/or increased in size to allow for the metering of a representative sample of one or more key end-uses. This might require adding a transponder to take advantage of channel capacity in an existing recorder, adding

more channels to an existing recorder, or adding households to the sample. This strategy optimizes the use of available equipment and takes advantage of other data collected from load-research customers.

Some important cautions about this approach should be noted. First, the integrity of the load-research sample must be maintained while providing an adequate evaluation sample. Existing load-survey customers may have an unmetered load available, but their customer profile may not fit the sample design requirements, thus leading to bias. Upgrading the metering and acquiring additional information from the customer may require further intrusions upon customers' time and privacy, causing customers to drop out. Thus, gains in economy may be offset by the need to deal with sample bias.

Sample Selection and Validation

Sample selection is the translation of a complex sample design into a practical set of activities leading to the installation of metering devices. This activity has several outputs:

- The primary samples (lists of customers to be contacted) for all programs
- Multiple backup samples for all programs
- Primary and alternate customer lists for meter installations
- In some cases, customer lists for personal or mail surveys

The responsibility for control of the sample should be clearly specified and should be known to all program participants. Any question or problem regarding a sample should be referred to that person.

A customer chosen to be in a sample may have to be dropped for any of several reasons: a difficult or impossible installation, no telephone line accessible from the metering point, a customer's refusal to participate, or a customer not having the appliance of interest.

The randomness of a sample must be ensured when a customer must be replaced. Installers tend to want to complete an "easy" or "next-door" replacement installation rather than to complete an installation that preserves the statistical integrity of the sample. The sample-design coordinator should keep a backup selection file and should specify any replacements on a customer-by-customer basis until the full sample is completed.

The problem of sample replacement is even more acute in the case of a stratified sample. Purists may argue that random replacement within a stratum is sufficient, while pragmatists may argue that the replacement sample point should have characteristics that are as close as possible to those of the sample customer being removed. The following method may satisfy both perspectives:

- When drawing the initial sample, draw ten separate samples, each with the same number of respondents as specified in the original design.
- With the original-sample design parameter(s), validate each sample by comparing means and SDs.
- Select the best "fit" sample, by stratum, as the installation sample.
- Take the remaining samples that fall within acceptable confidence limits (e.g., 99% to 90%) and combine these samples into a master backup list maintaining the stratification and the design criteria (e.g., where the design criterion is peak-month billed kW, segment the "master backup" by both stratum and 10-kW increments).
- When replacing sample points, randomly draw from the master backup sample from within the appropriate design segment.

This method ensures that the overall central tendency, variance, and distributional shape are closely maintained over the course of the load-research experiment. A variation on this concept that provides even greater fidelity to the original design entails the use of appliance survey data to develop a replacement matrix. This method provides greater matching precision while maintaining statistical integrity.

Recorder Selection

Equipment selection is driven by a number of criteria. These criteria may include the type of measurements to be made (e.g., whole-building versus end-use data), the frequency with which measurements must be made, and the future uses of the equipment.

Whole-building meters can be placed in an existing standard watt-hour socket. These devices provide both billing and load-research data. These meters can be read by a handheld recorder, a handheld device aimed at the meter while driving by, or a modem. These meters can contain logic and sufficient memory to store many days of readings taken at intervals ranging from a minute to weeks.

Multichannel end-use meters are usually installed independent of the billing meter. They are capable of taking data from 2, 4, 8, or 16 channels. Transponders at the metering points may be hard-wired to the recorder or may communicate with the recorder with some form of power-line carrier and the building wiring. The recorder usually contains sufficient memory to store up to several days of data taken at intervals of one minute to several hours. The data stored in the recorder is usually captured by the use of a handheld recorder, a modem that calls a master station periodically over the customer's telephone lines, or a modem connected to a leased phone line.

Which of these options to use depends on the trade-offs between data requirements; capital costs; installation costs; operational costs (such as the use of a meter reader, the customer's phone line, or a leased line); analysis costs; and restoration costs to the customer's premises.

For example, using a power-line carrier may have lower installation and restoration costs than hard-wiring transponders, but it will likely have higher capital costs. Whole-building metering is generally less costly than multichannel metering but may require more sample points to obtain the same accuracy.

An important issue is the coordination of master-station software with recorder selection. Most utilities now have inventories of recorders from different vendors. Different vendors use different communication protocols. If the vendor for the master-station software is different from that of the recorder, then the master-station software supplier should communicate with the recorder supplier about appropriate modifications to the software. The products of some independent suppliers of master-station software can communicate with a variety of recorders, thereby integrating recorders from different vendors into a coherent system.

If software is to be developed (or modified) in-house, then early communication with the recorder supplier will facilitate development. Early attention to this detail may suggest some easy modifications to recorder output data or data formats to meet particular in-house analysis requirements.

Care in recorder selection can produce significant savings for the load-research program. A common problem is the mismatch between current meter inventory and the requirements for the next load-research experiment. Plans for the purchase of new equipment should allow future experiment cycles to use equipment and installation and removal labor efficiently.

Spare recorders should be available at all times. With the popularity of stratified samples, the loss of a few recorders, especially in the same stratum, can render entire periods of data statistically unsound. Utilities considering joint load-research efforts should investigate the possibility of integrating samples and sharing spare equipment.

Meter Quality Control and Quality Assurance

Meter quality control and quality assurance is an essential ongoing activity that ensures the reliability and the cost-effectiveness of the load-research activity. Reliable metering reduces the need for oversampling and lowers costs. The level of reliability depends on whether or not the data will also be used for billing.

Meter problems have several potential sources, including metering equipment that is faulty upon arrival from the factory, equipment that is improperly installed, faulty record keeping (for example, misrecording information about the location of a meter or channel information), and failure of meters in the field for hardware or environmental reasons.

The utility should establish a test bench and test all metering equipment supplied by manufacturers. All of the equipment to be used at a site should be tested together. The site-specific settings should be made and tested while the equipment is on the test bench. Appropriate labels (barcodes) should be attached to the individual devices. For example,

if an air conditioner, a water heater, and the total household load are to be monitored, then the channels should be preassigned, current transformers and transponders associated with the channels should be preset, and the devices should be labeled. A standard protocol should be used for assigning channels to end-uses. This practice will help reduce errors, minimize problems in the field, and aid trouble shooting. Everything should be double-checked and recorded, and then the equipment should be boxed for transport to the site.

Equipment performance should be monitored constantly. Each new installation should be polled as soon as is practical after installation, and the resulting data should be checked to see if the unit is functioning properly. Most master-station data-collection software interrogates metering devices and reports power outages, malfunctioning transponders, or high rates of communication errors. These reports should be reviewed every time the meters are polled. Range checks should be performed on the data. The data should also be reviewed to identify channels where values are always "0" or where values are valid but constant. Every instance of invalid data or suspected equipment malfunction should be followed up to identify and catalog the causes. Equipment performance should be discussed frequently with the field staff and their supervisors. These procedures can lead to the detection and correction of manufacturing and installation problems before significant data loss occurs.

The researcher has to seek the assistance of the most qualified field installers and meter-maintenance personnel. The metering and billing departments should be involved in decisions about metering devices and the planning of metering operations. Without such consultation, the priorities of field personnel are likely to be set by other organizations, and the quality of load-research data will suffer. Regular meetings should be held with field personnel to review equipment performance and to reinforce their understanding of the importance of the load-research program to the utility.

Meter Placement and Replacement

Meter placement and replacement are continuous activities, but periods of intense effort occur at the beginning and end of each research cycle. Proper staffing and effective planning can minimize the cost of this activity. Predefined placement and replacement procedures ensure consistency and efficient installation while maintaining the overall statistical quality of the experiment.

The resource requirements for meter placement depend on the type of metering equipment, the familiarity of the field staff with the equipment, the capability of staff to do maintenance and repair, and the type of communications. The meters for a sample cannot be placed overnight. Mobilization of the necessary staff must be well planned. If hard-wiring is required and/or installation contractors are to be used, then the customer must be informed of this. Contractual arrangements must be in place to ensure the availability of the premises for a timely installation (even though descriptive information may have been given to the customers and an authorization for the installation was obtained). Contractual instruments to be used with customers should be kept simple. Good customer relations dictate the inclusion of an offer to correct any complaint or to remove the installation if the customer

so desires. Resolution of complaints should be handled by the utility rather than by a contractor.

Spare equipment must be on hand and in good repair to permit efficient and timely replacement of malfunctioning units. The data-collection software should be able to document any unusual equipment outages (e.g., during replacement) and the reinitialized values of any data stored in the accumulators of the replacement meters.

Installers should treat customers' premises with care. The customers' premises must be restored to their initial condition upon completion of the program. Lack of attention to this detail can irreparably damage relations with customers. Arrangements with contractors should explicitly include restoration of customer premises, and an inspection of the final result should be conducted, preferably with the customer in attendance. Removal of equipment requires as much attention as does installation.

Installation Record Keeping

Keeping installation records is a very important but often overlooked aspect of load research. Many a utility has had to send someone to each installation to retrieve a vital piece of information needed by analysts or to determine if any of the equipment at the site is of a type recalled by a vendor. Many crews have spent hours trying to locate a piece of equipment that was "in there." More than one program manager has wondered if a particular set of failures is specific to a particular model or lot of equipment.

These examples illustrate the need for a meter-installation tracking system. Such a system needs to include the model, lot number, and serial number for each piece of equipment and its utility identification number. Having similar information about component parts is also useful. The settings for each piece of equipment should be recorded as should installation information, such as date installed, date serviced, reason for service, and name of installer.

Installers should make a map of the location of each device installed at each location. Alternatively, the installers may photograph the installation and surrounding area with identifiable features. This map or photo will help to locate a device when it is to be maintained or retrieved. Maps and photographs can be scanned into a database so that the information can be maintained and viewed electronically.

Meter Operation and Maintenance Activities

The level of company involvement in maintenance will depend on the particular operation and maintenance capabilities of the meter-maintenance staff and the type of equipment to be used. Diagnosis and/or repair of logic circuits should not be left to the company computer hacker. If the load-research program is experiencing equipment problems, the program director may want to meet with the maintenance staff and the supplier to ensure that equipment is checked upon installation and that any limitations on in-house maintenance are identified. Contracting with the supplier to maintain equipment may be required, and if so, the supplier's ability to provide timely and adequate services or

replacements should be checked. A storage room full of equipment waiting to be repaired is not only an inefficient use of capital but potentially a problem affecting the quality of data provided by a load-research program. Maintenance of communication equipment may be entirely separate and may not be covered by arrangements for metering-equipment maintenance. A warranty may not be a suitable substitute for equipment repair if the program must be slipped a year while awaiting warranty resolution of equipment problems.

Data Collection

Data collection may be accomplished in several ways. Onsite collection and remote polling with the telephone are the most popular methods. The selection of the most appropriate medium involves assessing the trade-offs among cost, reliability, and program requirements. It is also a function of the geography and the rural/suburban/urban mix of the population.

Onsite data collection typically is done by using a hand-held microcomputer interfaced to a recording device through an optical port. Onsite data collection has a number of advantages, including eliminating coordination problems with the telephone company, the opportunity to visually inspect the recorder site and surroundings at the time of data collection, and sophisticated technology to permit error-free data collection.

Recent breakthroughs in hand-held microcomputer technology have enhanced the attractiveness of this method. Some units use standard microcomputer operating systems and literally guide the meter reader through the data-collection process. Incorrect identification numbers, incorrect visual meter readings, and incorrect "time in" and "time out" observations are eliminated. Further, the ruggedness of these devices ensures safe storage and transport of data to the utility's computing system and master file. The main drawback is the cost of dedicated employees. However, given proper planning and the relative ease of the data collection, this operation can be performed part time by meter personnel or by dedicated full-time employees adhering to a flexible data-collection schedule. Load-research recording devices and the portable data-collection equipment now have sufficient memory capacity to ensure that data will not be lost if collection schedules are not strictly followed.

Telephone data-collection methods are probably most cost-effective in unusual situations, such as distant or dangerous locations. Telephone polling permits frequent communication with metering devices and presents opportunities to reduce data losses. Unfortunately, many companies do not take advantage of this feature because they do not screen the data obtained from load-research programs in a timely fashion.

Load-research equipment typically features redundant data-collection methods. For example, units with telephone modems may contain optical ports for manual data collection. Data-collection costs vary depending upon the method selected. Telephone meter reading costs approximately \$300 per recorder per year, while onsite manual readings through optical ports will cost about 50% more. However, cost is clearly not the only consideration, and given the cost of lost load-research data, reliability and redundancy should be equally stressed.

Data Editing and Validation

Data editing is divided into two types. Housekeeping editing is the minor editing that involves adjusting start and stop times; summation of total kWh consumption over the appropriate period; and correction of improper IDs, meter multipliers, etc. Load-data editing involves identifying and replacing missing or bad data and correcting data for outages.

Editing involves a combination of human expertise and efficient software. Depending on the storage algorithm, as much as 3.5 megabytes of data can be produced per day by 200 four-channel recorders recording data at 5-minute intervals. These data cannot possibly be visually inspected. Thus, software must be used to evaluate data quality.

Sometimes, load-research data collected for a DSM evaluation is also used for customer billing. This creates special problems for the load-research editing function because two sets of data for the same customer, one for billing and one for the evaluator, must be developed. The billing application requires rapid turnaround of edited data, so the procedures for modification or replacement of missing or damaged data must be formalized.

Validation of load data is an audit function that ensures that the data entering the master file is consistent with data that has gone before. Validation is an important aspect of data collection and sets the stage for any editing that may occur.

Since the introduction of solid state electronic load-research recording devices, the incidence of data loss has dropped dramatically. During the days of magnetic tape recorders, typical month-to-month data loss was 10% to 15%. With the maturation of electronic devices, typical data loss may average 3% or less. However, even minimal data losses can be critical for stratified samples. Therefore, the utility should develop decision rules for editing data. Where sufficient sample sizes exist and the samples have been meticulously maintained to preserve the original objective of the sample design, minimal editing may be the best course of action. However, if extensive editing is required, automated data editing systems are available to aid the process.

How to handle data if an outage occurs at a customer site has always been subject to differences of opinion. There are three basic approaches. The first is to simply insert zeros in the appropriate time slots. The second method is to remove the record from the analysis for the outage period. The third is to construct a hypothetical record for the outage period. While some may argue that outages are a normal part of a utility's business, it is probably better to recognize that every sample customer represents hundreds, perhaps thousands, of fellow customers and that the outage is not indicative of what is occurring on the system. Further, if the evaluator is interested in the impact of a device, such as an efficient air conditioner, then representing the data with zeros does not help to meet this objective. The best course appears to be to simulate the data or to remove that part of the case from the analysis.

Outages not only influence data for the period of the outage but also for a period following the outage. An early-morning outage might cause an electric water heater to run for an extended period during the late morning. Thus, it may be necessary to modify or remove data beyond the period during which the outage occurred.

Data Analysis

Both primary and secondary load-research data are required by DSM evaluations. Taken together, these data form the basis for drawing conclusions from the impact and process evaluation (Fig. 13). Here we mention a few considerations not covered in Chapter 4.

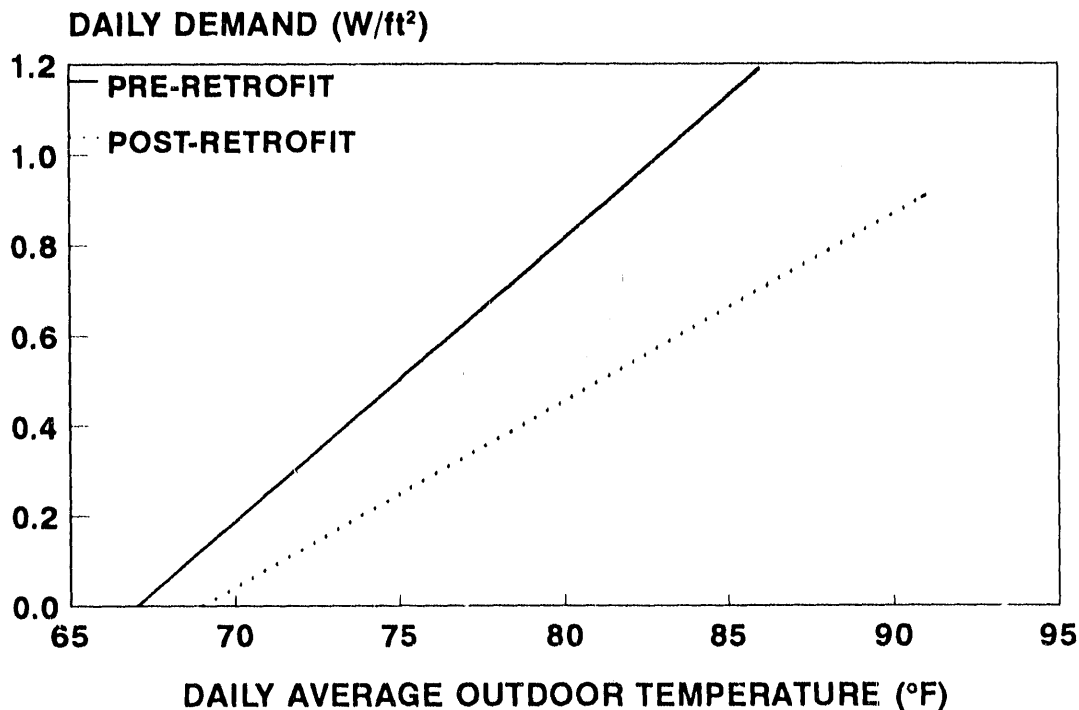


Fig. 13. Daily electricity demand for air conditioners in Austin, Texas, based on 15-minute load data collected at 14 homes. The air conditioners in these homes were replaced in late 1987 with high-efficiency units. These replacements cut peak demand by 38%.

Load-research programs often founder because of the failure to anticipate the requirements of data management. An active load-research activity can produce enormous quantities of data. It is not unheard of for data to be "lost on some tape," or for some preprocessing to be done several times because analysts do not know what preprocessing has been done previously, or for data to be lost because it is not adequately protected.

Every load-research activity should have a data-management plan. In addition to identifying how data are to be validated, such a plan should identify how the data are to be initially

processed, whether they are to be stored by sample identification number or end-use or both, what quantities of data are to be stored, where they are to be stored, how files are to be named, how they are to be backed up, and any special requirements for securing access. A data dictionary should also be established. Finally, a data-management plan should identify the types of analysis that are to be conducted and the schedule for those analyses.

Data files and analyses should be tracked in a special database. The database should tell the location of the data, when it was collected, when it was processed, and how it was processed.

Analysis of load-research data typically takes a significant amount of machine time. For example, it can take a half hour or more to process a hundred thousand records on a fast minicomputer. The analyst will probably want to work with test samples to develop and test software before processing the entire set of records. Microcomputers and workstations can speed this work because of their friendly interfaces and the fact that they are dedicated machines. Some utilities now use microcomputers almost exclusively for this work, and specialized software packages are available for such machines.

Samples should be monitored to see how they change over time. Commercial firms add and release personnel, people take new jobs, household occupancy changes, and each of these factors affects energy use. Because metering samples generally are small, a few such changes may affect the character of the sample or stratum significantly, thereby influencing the generalizability of the data. For example, the estimated effects of a load-control program may be misleading if a local company adds a second shift and many persons in the sample change from day shift to night shift.

The analyst can take several actions to detect such changes. Billing data totals can be compared. If more than a year's monitoring data are available, the analyst can compare consumption data between years and relate the observations to any expected changes from the program. A follow-up survey of customers can reveal whether changes have taken place. If changes have occurred, restratifying the data may be necessary for meaningful analysis. The researcher should be prepared to evaluate all data for consistency before completing an analysis. Control groups must also be monitored for changes.

THE COST OF LOAD RESEARCH

Estimating the true cost of a load-research sample point is difficult. Economies of scale generally apply but may be lost as the size of a program changes and the use of "lumpy" resources, such as people and computers, changes. Metering costs can vary significantly depending on the number of end uses monitored at a sample location, local communications, and the difficulties of hard-wiring transponders. The first end-use point at a location is often costly, but the incremental costs for additional end-use points are less. Shared resources (e.g., programmers and field maintenance staff) may help mitigate the costs, but the effectiveness of this strategy depends on the availability of expertise and the disposition toward cooperative efforts within any particular company.

Tables 21 and 22 provide cost estimates for a range of customer sample sizes. These estimates contain assumptions about the number of end-use sample points per customer, which will vary from program to program and which can significantly affect metering costs. Also, a utility may engage a contractor as a way of trading off staffing and internal processing costs, particularly when smaller samples are involved.

Training costs have not been included. They may be significant for a small program but may be mitigated by the use of contractors, especially if the program is to begin small and grow.

Table 21. Assumptions used to estimate costs in Table 22

Meters	Points monitored		Installed meter cost (1991 dollars)
	Total load	End-use points	
Meter configuration			
Residential	1	2	2,300
Commercial	1	2	3,700
Industrial	1	3	5,000
Annual charge rate			22.0%
Annual salaries			1991
Staff classification			Annual cost
Manager			40,000
Data editor			20,000
Analyst			27,000
Programmer			30,000
Field staff			26,000
Maintenance staff			26,000
Overhead loading rate			30.0%
Annual salary escalation			4.5%
Central equipment and facilities			Cost
Type			
PCs, printers, etc., each			6,000
Vehicles, each			14,000
Communication equipment, per meter			30
Minicomputer			Estimated
Supplies			Estimated
Purchased software			Estimated
Annual charge rate			20.0%

Table 22. Load-research costs (1991 dollars) as a function of sample size

Sample size (meters):	150	450	750	1,500
Program size description				
Meters				
Residential	100	300	500	1,000
Commercial	30	100	150	300
Industrial	20	50	100	200
Staff (prorated for shared staff)				
Manager	0.5	1.0	1.0	1.0
Data editor	1.0	2.0	2.0	3.0
Analyst	1.0	2.0	3.0	4.0
Programmer	0.5	1.0	2.0	2.5
Field staff	1.0	2.5	3.5	5.0
Maintenance staff	0.4	1.5	2.5	4.0
Equipment				
PCs	2.0	4.0	6.0	8.0
Minicomputer	0.0	0.0	1.0	1.0
Vehicles	1.0	2.0	3.0	5.0
Summary of first costs, dollars				
Meters	441,000	1,310,000	2,205,000	4,410,000
PCs	12,000	24,000	36,000	48,000
Minicomputer	0	0	70,000	70,000
Vehicles	14,000	28,000	42,000	70,000
Communication equipment	4,500	13,500	22,500	45,000
Purchased software (capitalized)	40,000	40,000	70,000	100,000
Total first cost	511,500	1,415,500	2,445,500	4,743,000
First cost per meter	3,410	3,148	3,261	3,162
Annual costs (1991 dollars)				
Salaries including overhead	153,920	348,400	490,100	672,100
Supplies, telephone (estimated)	1,000	2,500	5,000	8,000
Office space	5,720	13,000	18,200	25,350
Present value of costs (20 years @ 11.5% discount rate)				
Meters	748,006	2,221,969	3,740,032	7,480,063
PCs	18,504	37,007	55,511	74,014
Minicomputer	0	0	107,937	107,937
Vehicles	21,587	43,175	64,762	107,937
Communication equipment	6,939	20,817	34,694	69,388
Purchased software	61,679	61,679	107,937	154,196
Salaries including overhead	1,597,648	3,616,298	5,087,106	6,976,217
Supplies, telephone	10,021	25,053	50,106	80,170
Office space	57,321	130,276	182,386	254,037
Total present value	2,521,705	6,156,273	9,430,472	15,303,961
Present value per meter	16,811	13,681	12,574	10,203
Levelized total annual costs, dollars				
Annual cost per meter	2,161	1,774	1,631	1,323
Annual cost per sample point	696	570	521	422

ORGANIZING LOAD-RESEARCH PROGRAMS

The success of a load-research program in meeting the needs of DSM evaluation depends upon the effectiveness of communication between evaluators and load-research program managers. Many utilities already have a load-research section, often located in the rate department. Utilities that are just undertaking or expanding their load-research activities often have a concern about where the load-research function should be located.

Two strategies are possible. One is to have a load-research steering committee. This approach often is used for a mature load-research organization. The load-research steering committee meets monthly or quarterly. This committee is comprised of representatives from the traditional areas of load-research use within the utility (e.g., rates, distribution engineering, forecasting, and cost of service) as well as representatives from the residential, commercial, industrial, and agricultural DSM-evaluation efforts. Members present their needs for load research to the committee. The load-research staff analyzes these needs and designs studies to meet as many of the needs as possible. Some needs may be met without load research or by using other types of analysis, such as conditional-demand analysis. Formalization of the committee within the utility structure places the proper emphasis on the importance of such optimizing procedures.

An alternative approach is to make load research part of the DSM-evaluation function. This strategy is appropriate for utilities that are instituting or expanding small load-research programs. This organizational choice is appropriate for a number of efficiency and cost reasons. Many of the skills required of load-research personnel, including statistics, econometrics, sample-design expertise, and engineering analysis, are also required of evaluation personnel. Load-research personnel are constantly called upon to meet the data requests of evaluators. Direct organizational responsibility is advantageous in ensuring a timely response.

APPLICATIONS OF LOAD RESEARCH TO EVALUATIONS

From a load-research perspective, DSM programs fall into five basic varieties: load-control programs; direct installation and rebate programs; informational programs, including audits; time-of-use rate programs; and interruptible-rate programs. The value of load research varies according to the program being evaluated. The extent to which load research is used should be a function of the estimated level of net resource savings expected from a particular program and the extent to which an adequate analysis of impacts can be achieved through load research or some other less costly approach (Long Island Lighting Company 1991; Rochester Gas & Electric 1991).

For load-control programs, load research is the most effective way to assess the impact of control strategies and to measure program free ridership (Exhibit 9). The size of samples is dependent upon the extent to which the use of a particular appliance or end-use is discretionary and the type of load-control strategy employed. For example, the effect of load control on pool pumps can be reliably determined with a small number of monitoring

Exhibit 9. Load-control programs often require load-research data for evaluation

PG&E evaluated its residential air-conditioner-control program with meters that recorded air-conditioner electricity use every 30 minutes for a sample of about 200 participants (Malcolm 1989). These load-research data were collected for control days (when the units were turned off from 2 pm until 6 pm) and for noncontrol days (Fig. 14). Noncontrol days were chosen for their similarity to control days on the basis of outdoor temperatures. These data show a substantial reduction in demand at the time of PG&E's system peak, 4 pm.

A regression model was developed to estimate load reductions as a function of outdoor temperature and the activation of controls (MacDonald 1987). The net reduction depends strongly on outdoor temperatures (Fig. 15) because of two factors. First, the fraction of free riders (households participating in the program that do not have their air conditioners on at the time of control) decreases with increasing temperature. Second, the air conditioner duty cycle (percentage of time the units are on) increases with outdoor temperature. These results indicate that the program should be offered only to customers who use more than 600 kWh/month in the summer to reduce the incidence of free riders, and it should be dispatched only at temperatures above 100°F.

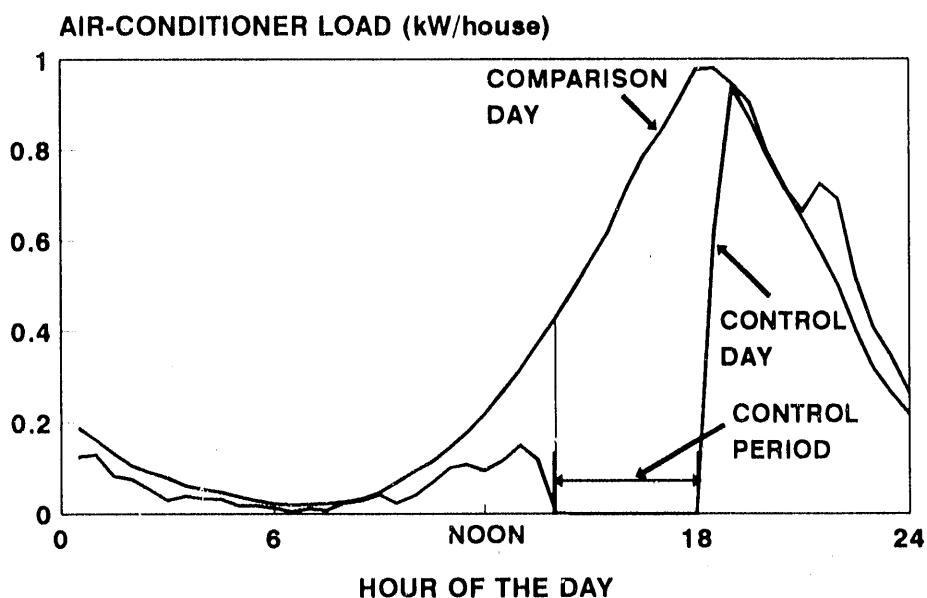


Fig. 14. Electricity demand for residential AC on control days and on noncontrol days. On control days, the units were turned off for four hours, from 2 to 6 pm.

Exhibit 9 (Continued)

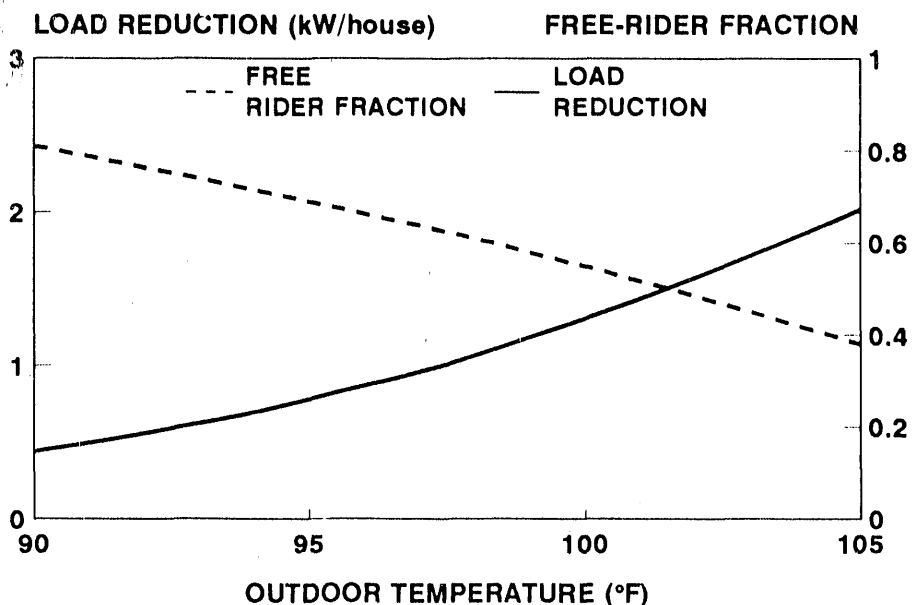


Fig. 15. The load relief achieved by PG&E's residential air-conditioner-control program increases with outdoor temperature.

devices strategically placed according to pump size and, perhaps, household size. However, the reliable measurement of central AC impacts requires more attention to statistical theory as well as recognition of the variations in customer response based upon such factors as geographic location, age of the head of the household, and age of the AC equipment.

Direct-installation programs (e.g., retrofits) and residential and C&I rebate programs (e.g., for light bulbs, refrigerators, or water heaters) have been used for a wide range of appliances and end-uses. The use of load research to measure these program impacts has been thought to be cost-prohibitive. Further, direct load research has limited applicability for assessing the impact of rebated measures.

Residential and C&I informational programs cover a wide spectrum of activities. They range from bill inserts informing customers about DSM programs to public appeals on radio and television to curtail load or cut back on energy use to customer audits that provide specific measure recommendations. Load research generally is not a viable measurement tool for information programs.

Residential and C&I time-of-use rate programs are excellent candidates for load-research-based evaluation. Load-research data are usually already available for C&I customers.

Load-curtailement programs have great potential to achieve significant peak reductions. Billing analysis is of little use in evaluating such programs unless it is accompanied by direct load-research data. While company-specific load research is usually preferable, borrowed data can be used in concert with billing analysis to estimate load-curtailement impacts in some instances. This information, coupled with weather and survey data, can be used to develop load-impact estimates for the customer, class, building type, and total.

JOINT UTILITY END-USE MONITORING

The decision to institute end-use monitoring to support DSM or other utility activities is a difficult one. Collecting end-use load data faces many barriers: cost, obtaining customer cooperation, the size of samples, and the many end-uses to monitor.

For these and other reasons, utilities sometimes hesitate to undertake end-use load monitoring on their own. As DSM programs become larger, some utilities will investigate alternative concepts for economically collecting end-use load data. One of these approaches is the joint utility project.

Joint utility projects can cut the costs of end-use monitoring. The major concern associated with end-use data analysis within a utility is the variability of the end-use loads. Adding multiple service territories may increase this variability, and joint projects must therefore be carefully considered. For example, weather, size of dwellings, type of equipment, construction practices, insulation, and family size can vary substantially within and among service territories, making SH a poor choice for joint end-use metering. Water heating may be a better prospect for joint projects because factors like family size, number of children, water source, and the presence of dishwashers and clothes washers can be identified and controlled. Estimates for the different service territories can be developed based on this information. The best prospects for joint end-use load research are those end-uses that are little affected by short-term weather and those end-uses that have little customer intervention (e.g., refrigerators, freezers, and water heaters) (Applied Energy Group, Inc. 1988).

CONCLUSIONS

Load research is a key element in the DSM-evaluation process. The costs of monitoring can be significant, but the need for measurement of DSM-program impacts makes this a necessary expenditure. This chapter identified a number of aspects of monitoring that the evaluator needs to attend to: planning, sample design, sample selection and validation, installation, quality control, and data analysis.

Direct monitoring has alternatives. Intelligent use of class, end-use, and borrowed load-research information, coupled with attribute information and statistical modelling, can often provide cost-effective alternatives to metering. The evaluator needs to carefully consider the problems and the alternatives before entering into a load-research program.

FREE RIDERS AND OTHER FACTORS THAT AFFECT NET PROGRAM IMPACTS

William Saxonis

INTRODUCTION

Accurately attributing energy savings and other benefits to DSM programs is a major goal of any comprehensive evaluation. To gain a better understanding of program impacts, the evaluation community has focused attention on program participants who would have taken the identical energy-conservation actions without the DSM program. This type of participant is referred to as a free rider.

Free riders must be examined because they represent a cost to the program but offer no direct benefits in return. A common practice among utilities is to adjust for free riders after gross capacity and energy savings are calculated. DSM programs that are highly successful from a gross-benefits perspective may prove to be less attractive when free riders are considered. For example, if 50% of participants are free riders, net program savings might be only half the gross savings. Program costs, however, would remain constant. From a societal perspective, free ridership is not important because they provide benefits to society even though their actions are not attributable to the DSM program.

In addition to helping determine net program impacts, free-rider data can serve as a useful tool in designing DSM programs for maximum effectiveness. By considering the impact of free-rider rates on program design, utilities could employ DSM resources more effectively and offer programs that are more responsive to the needs of the target audience.

While free ridership is theoretically an issue for all DSM programs, past research concentrated on programs offering rebates for the purchase of energy-efficient equipment and lighting. Appliance rebate programs have received considerable evaluation attention, resulting in at least 20 evaluation reports that include free-rider estimates. C&I rebate programs have also been frequent candidates for free-rider analysis.

To accurately identify the effects of free riders, understanding exactly what constitutes free ridership is important. Because programs can affect participants in different ways, three major free-rider categories are distinguished (Table 23).

A pure free rider is a participant who would have taken the *identical* energy-conservation actions at the same time if the program never existed. By contrast, a nonfree rider is a participant who takes the desired actions as a *direct* result of the program.

Table 23. Categories of free ridership

Category	Key characteristic
Pure free rider	Would have taken the identical action without the program
Incremental free rider	Influenced by the program to take action but not to the extent advocated by the program
Deferred free rider	Takes the action promoted by the program sooner than originally planned

An incremental free rider is a participant whose behavior is influenced by the program but not to the full extent advocated by the program. For example, a small business plans to install energy-efficient lighting without being aware of its utility's commercial-lighting program. After learning that the utility is offering a rebate for such lighting systems, the business purchases lighting that is 15% more energy efficient than the equipment it had originally planned to buy. In this case, the net program impact is the difference in electricity use between the lighting the business had planned to install and the even more efficient lighting it did install.

A deferred free rider is a participant who would have taken the same actions promoted by the program, but is influenced by the program to take the actions sooner than planned. For example, an evaluation of PG&E's Customized Rebate Program found that 68% of the participants planned to implement the same measures advocated by the program even if the utility did not offer a rebate. This group also said that the program encouraged them to install the measures from six months to more than five years sooner than originally planned (Pacific Consulting Group 1986). The energy savings between the time the customers installed the measures and the time they planned to take the actions are legitimately attributable to the program.

Unfortunately, defining free riders does not end with these three categories. For example, an incremental free rider may also be a deferred free rider.

FREE-RIDER MEASUREMENT

Despite the increased attention focused on free ridership, its measurement is an inexact science. Most of the literature on the topic has been critical of the reliability of the estimates. Lui and Fang (1990) concluded that "free-rider estimation can and should be improved." And Kreitler (1990) noted that "the standards for how to conduct evaluation research into free ridership need to be strengthened."

Although efforts are underway to improve measurement of free ridership, the lack of a reliable method to determine free-rider rates remains a problem. In fact, research has frequently produced confusing and inconclusive data (Table 24). For example, five utility C&I high-efficiency-motor programs identified free-rider rates ranging from 3% to 88% of total participants.

Table 24. Examples of free-rider estimates

Program type	Free-rider rates
Residential	
Refrigerator (rebate)	59–89%
Air conditioner (rebate)	19–79%
Heat pump (rebate)	40–60%
Loan programs	22–70%
Low-income programs	6–45%
C&I	
Lighting	5–85%
Motors	3–88%
Multiple end uses	5–70%

Sources: Kreitler (1990), Lui and Fang (1990), and Nadel (1990).

Several factors explain the variation in these estimates. In some cases, deficiencies in the research design cause the variance, but in other cases differences in program design, market segments, and reporting methods cause it (Table 25).

SURVEYS TO MEASURE FREE RIDERSHIP

Because programs and the environments in which they operate differ, the methods for determining the free-rider impact must be tailored to each program. Experience, however, provides valuable lessons for enhancing free-rider measurement with survey instruments, data from the marketplace, comparison groups, and modeling techniques.

Survey Program Participants

The most frequently used approach to measure free riders is to survey participants. Usually this survey asks participants one or more questions about what they would have done if the program were not available. This approach has a low cost, is easy to administer, and can provide results quickly. In some instances, such questions are part of the program-application form.

The key to successfully using this approach is to design carefully the questions aimed at determining free-rider rates. Kreitler (1990) claimed that a serious problem in accurately determining free-rider rates is that "inadequate attention has been devoted to the wording of survey questions, leading to misunderstanding on the part of respondents and misinterpretation on the part of analysts."

Table 25. Different methods of reporting free-rider rates

Wisconsin Electric refrigerator and air conditioner turn-in programs (single family)

Results

- 28% pure program influence
- 7% mixed program influence
- 28% mixed windfall^a
- 37% pure windfall

Long Island Lighting residential air-conditioner rebate program (condenser-unit replacement only)

<u>Energy Efficiency Ratio Range</u>	<u>Free-rider rate</u>
10.5 - 10.99	10%
11.0 - 11.49	25%
11.5 - 11.99	22%
12.0 - over	15%

Northern Illinois Gas Company gas-furnace rebate program

<u>Category</u>	<u>Free-rider rate</u>
Pure free rider	71%
Incremental free rider	18%
Nonfree rider	4%
Do not know	7%

^aWindfall reflects what would have occurred in the absence of the program and is equal to the sum of naturally occurring conservation and free ridership.

Sources: Wisconsin Electric (1991); Applied Energy Group (1989); and Lui and Fang (1989).

Target the Questions to Address Free Ridership

Some evaluations estimate free ridership by asking consumers to indicate if the program influenced them to take energy-conservation actions. While the responses may be useful in

assessing the participant's feelings toward the program, they are a poor way to quantify the effects of free riders.

The principal flaw is that the question is too indirect to measure program-induced change. Participants may identify the program as being influential, but may also say that the level of influence was not sufficient to convince them to do something they would not have done anyway. This problem occurs in numerous evaluations. In an appliance rebate program operated by the New York State Energy Office, 71% of the participants claimed that the rebate influenced their appliance purchase, but only 34% indicated that they would have purchased a different appliance if the rebate were not available (Saxonis 1991).

In an evaluation of a commercial-sector program operated by Long Island Lighting Company, two questions were asked to determine free ridership. Participants were asked directly, "Would you have purchased that particular high-efficiency unit without the rebate offer?" Participants were also asked, "Did the availability of the rebate have any influence on your decision to purchase any of the technologies listed ... ?" In many cases consumers responded "Yes" to both questions. These responses are inconsistent. A sample of customers responding "Yes" to both questions was surveyed by phone to clarify their answers. The results of this survey found that almost all of the customers responding yes to both questions were free riders (Applied Energy Group 1989). Obviously, if only the second question was used to estimate free-rider rates, the results would be misleading.

Ask Questions That Provide a Complete Picture

Some evaluators simply ask participants if they would have taken the same action if the program were not available. For example:

- Would you have purchased this appliance if the rebate were not available?
- Would you have installed this measure without the program?

While this type of question directly asks about free ridership, a single question is usually not sufficient to fully address the issue. Asking questions similar to these will prompt only a yes (free rider) or no (nonfree rider) response. Incremental or deferred free riders cannot be identified with such simple questions.

An additional concern is that if the question is not well worded, the responses will be difficult to interpret (Kreitler 1991). For example, some consumers may confuse the purchase of the appliance in general with the purchase of a particular appliance model.

A better approach is to more thoroughly investigate the customer's decision to participate in the program. This approach involves asking several questions to validate responses and to gain a more complete understanding of participant behavior. In an evaluation of a C&I program operated by Madison Gas and Electric, customers were asked what they would have done without the program along with several questions to further probe and verify free-rider levels:

Would you have done the(se) exact same measure(s) *at the same time you did* if the program had not existed?

1. No
2. Yes----> So you are saying the program had no impact on your decision to do the(se) thing(s)?

1. No
2. Yes

Interestingly, depending on the measures installed, 10 to 25% of the respondents answered "No" to the second part of the question. The evaluators reported that the estimates of free riders would decrease for the entire sample from 10 to 15% to 8 to 13% "if we adjust for responses to this question" (Vine et al. 1990). A second question was asked to further probe participant actions:

What do you think you would have done if the program had not existed?
Would you have:

- a. done the(se) things at a later date?
 _____ No _____ Yes _____ Don't know
- b. (If multiple things installed) done fewer things at the same time?
 _____ No _____ Yes _____ Don't know
- c. (If can vary in efficiency) installed less efficient equipment at the same time?
 _____ No _____ Yes _____ Don't know

The responses to this question showed that many participants were encouraged by the program to install more measures and more-efficient equipment than they had originally planned. This approach provides a more accurate assessment of net program impacts, including free ridership.

Tailor Questions for Your Program

Different programs require different types of questions to measure free riders. For example, most consumers view a refrigerator as a necessity. The majority of the participants in refrigerator-rebate programs purchase a refrigerator because the old unit broke or they were moving to a new home. A rebate might encourage a consumer to purchase an efficient refrigerator, but rarely would someone buy a refrigerator because a rebate is offered (New York State Energy Office 1989; Synergic Resources Corporation 1987).

C&I incentive programs frequently promote actions that are considered beneficial but may not be necessities (e.g., replacing incandescent with fluorescent lamps). Under these circumstances, it is important to ask participants if they decided to take the conservation

action before or after they heard about the program. Obviously, if they decided to take the action before hearing about the program, they are free riders or incremental free riders.

Sometimes, asking customers if they decided to take the action before becoming aware of the program can be challenging, especially if the DSM program involves several components (e.g., information, audit, and installation of measures). For example, a small business receives a utility-sponsored audit from an independent engineering firm. Later, the business talks to several contractors to learn more about the features and costs of several measures recommended in the audit, including some that are eligible for the utility rebate. Because of the various influences in this hypothetical program, it may be difficult for the participant to accurately identify and assess the impact of the DSM program as distinct from the independent influences of the engineering firm or the contractors. The customer might say that the utility program had no effect because a contractor convinced him/her to adopt the measure. In reality, the DSM program may have played a pivotal role in the firm's decision.

This problem was encountered in a C&I incentive program operated by Central Hudson Gas and Electric Corporation. The evaluation (Applied Management Sciences 1990a) states,

A contractor or dealer who is aware of the program and intends to use the rebate as a marketing tool, might persuade the customer to invest in efficient equipment. When the customer fills out the marketing data on the application, however, it appears to him that he has intended to invest in efficient equipment from the beginning because that was the only type of equipment the dealer tried to sell him. In such a case, a self-reported free rider is not really a free rider.

Timing of the installation is also important in estimating free-rider rates. Did the program allow the consumer to take the action sooner than planned? If so, by how long? Table 26 shows sample questions for a C&I program operated by PG&E (Pacific Consulting Group 1986).

Identify Free Ridership for As Many Program Components As Possible

In programs that promote adoption of several DSM measures and practices, asking questions that pinpoint program elements where free ridership is the highest and where it is the lowest is important. This permits the utility to use evaluation results to fine tune the program by modifying or eliminating program measures that have high free-rider rates. A NEES program designed to encourage energy-efficient improvements in C&I structures illustrates this point. Rather than simply asking participants if they would have taken any actions without the program, respondents were asked to identify specific measures (e.g., energy-efficient lighting, heating-system improvements, and insulation) they would have implemented without the program. Although 53% of the participants indicated they would have installed some of the measures, the measure-specific free-rider rates ranged from 30 to 83%. This detail can be valuable in better understanding the program as well as in improving program design (Charles River Associates 1991).

Table 26. Sample questions for a C&I program

Question: How did the availability of the 1984-85 Customized Rebate Program affect your decision to make the equipment change for which you got the rebate? Would you say ... ?

Responses: You wouldn't have made the change without the program.
You would have made the change anyway.
You would have made the change anyway but the program speeded up the change.
Don't know.

Question: Did the 1984-85 Customized Rebate Program allow your company to undertake this rebate project *earlier* than originally planned?

Responses: Yes
No
Don't know

Question: (If Yes) How much earlier did you make the change?

Responses: Less than 6 months
6 months to 1 year
2-3 years
4-5 years
More than 5 years
Don't know

Limitations of Survey Methods

Even with well-designed survey questions, the evaluator must deal with recall and other problems of self-reported data. Specific issues include fading memories, a tendency to tell surveyors what they want to hear, and a reluctance to admit that they would have done something different from what they did.

Questions on the energy-related actions of participants may be asked several months after their involvement in a program. It may be unrealistic to expect participants to remember all the details of their participation. As a result, evaluators should survey participants as soon as possible after they receive program services.

Some respondents are reluctant to admit taking advantage of a program that failed to influence them. Others may not know what they would have done without the program; they may provide answers that they think will please the interviewer and will be consistent with positive goals, such as protecting the environment.

Do people learn, believe, and then act, or do they act and then make their beliefs consistent with their actions? Once a purchase is made, the consumer's perception of the product not selected may become more negative, and the perception of the product purchased may become more positive. This phenomenon makes it unlikely that the consumer will indicate a desire to have taken a different action without the program service (Calder 1973). Such responses will result in an upward bias in free-rider estimates.

MARKETPLACE DATA TO ESTIMATE FREE RIDERS

Surveying program participants is not the only way of determining free-rider rates. Other methods include analysis of market data, use of comparison groups, and discrete-choice modeling. These topics are discussed in this and the following two sections.

The impact of DSM programs is usually not limited to the customer that actually receives the service. If consumers are receiving rebates for purchasing efficient appliances, retailers, distributors, and manufacturers may act to ensure that the qualifying products are available in the stores. Moreover, they may promote the program by using the utility incentive as part of their own marketing efforts. For programs that advocate measures that require professional installation, a network of architects, engineers, contractors, installers, and suppliers are affected. These "trade allies" (service providers, equipment distributors, and retailers) can be surveyed to yield estimates of free ridership independent of those obtained from program participants.

For appliance-rebate programs, the net impact of the program appears to be higher (and therefore estimates of free ridership are lower) when sales data (rather than participant surveys) are used to estimate free riders. Dealers participating in the New York State Appliance Rebate Program reported a significant shift in the number of energy-efficient refrigerators and room air conditioners they stocked and sold during the time the rebates were available. This occurred at a time when only a modest increase occurred in the efficiency levels of appliances sold nationally. A majority of the dealers attributed the increase in efficiency levels directly to the program. Anecdotal reports suggest that appliance distributors shipped less-efficient appliances to other regions (Saxonis 1989). Considering that the primary purpose of the program was to increase the sale of efficient appliances, the program may have achieved its goal despite an apparently high free-rider rate (as estimated from participant self-reports). The appliance dealers appeared to have successfully reduced the opportunity for consumers to purchase inefficient models.

In the case of rebate programs, evaluators have asked retailers if they sold more of the product qualifying for the incentive during the program than before the program. After adjusting for general trends in product sales, the difference between the before and after sales provides an indication of the net impact of the program. This approach has the advantage of examining what actually occurs as contrasted to asking consumers to speculate on what actions they would have taken if the program never existed.

A major barrier to collecting this type of information is the reluctance of retailers to provide data because such information is often considered proprietary. National trade organizations, such as the Association of Home Appliance Manufacturers and Gas Appliance Manufacturers Association, provide some product and sales statistics, but they usually reflect only national trends and fail to identify local sales of specific models.

Despite these difficulties, several approaches have been used successfully to collect marketplace data. A key to success is developing a positive relationship with the organizations from which you are requesting the information and stressing the importance of the evaluation effort to their business. Wisconsin Electric reported that, with persistence (i.e., several phone calls and a letter from a senior vice president), they received sales data from an initially reluctant group of lighting distributors. They also encouraged cooperation by paying the distributors \$200 each to help defray the cost of collecting and reporting the data (Brooker and Fichtner 1991). Wisconsin Electric also obtained sales data by paying appliance dealers \$5 per sale to collect data on purchases of both qualifying and nonqualifying air conditioners as part of an evaluation of their Smart Money program (Brugger and Brooker 1991).

Some utilities require dealers to provide sales data as a prerequisite for participating in their program. For example, the San Diego Gas & Electric Earthwise Appliance Program requires participating dealers to provide both preprogram and program sales data. Most of the major-appliance dealers in the San Diego area (about 60) have agreed to participate. Some dealers have even offered to provide names and addresses of consumers who purchased appliances not eligible for the rebate. In addition, the utility plans to conduct a survey of appliances displayed in the dealers' showrooms for additional data validation (Wiggins 1991).

Another approach is to survey dealers with a voluntary and confidential survey. An evaluation of an appliance rebate program in New York asked dealers to indicate in ranges (e.g., 10-20%, 30-40%) the percentage increase or decrease in inventory of equipment that would qualify for the rebate. Ranges were used to reduce dealer reluctance to provide detailed sales and inventory data. Dealers were also asked if inventory changes were made because of the program (New York State Energy Office 1989).

The effectiveness of this type of data collection is limited because ranges sacrifice precision. For example, if the dealer indicated that sales increased 10 to 20%, the actual percentage could be as low as 10% or as high as 20%.

An additional concern is that dealers may overestimate program impacts to encourage continuation or expansion of the program. It is important to consider that the availability of the DSM program may benefit their business.

Nonresponse bias is a potential problem for all of these approaches, if not all the affected dealers participate. It may be possible that the dealers most enthusiastic about the program will respond to the survey and the less enthusiastic dealers will not.

COMPARISON GROUPS TO ESTIMATE FREE RIDERS

The use of a comparison group is a common evaluation technique that can be employed to account for free riders. The difference between the energy-conservation actions of a sample of program participants and an appropriately selected comparison group should provide an accurate estimate of the net program impacts. A major advantage of this approach is that it compares actions as opposed to asking participants to estimate what they would have done if the program were not available.

Finding a group that is comparable to the program participants but not contaminated by the DSM program can be difficult. Specifically, participants choose to participate while nonparticipants either choose not to participate or are unaware of the program. As a result, systematic differences may exist between participants and nonparticipants with respect to energy-use patterns and attitudes towards energy efficiency.

Addressing the potential for self-selection bias in the comparison group is important. This point was vividly illustrated in a process evaluation of NEES's C&I incentive program. The evaluators attempted to estimate the level of free ridership by comparing the activity of participants to nonparticipants. They found that 13% of the nonparticipants installed measures similar to those being promoted by the program, suggesting a free-rider rate of the same percentage.

The report correctly noted that differences between participants and nonparticipants would likely inject a significant bias in their free-rider estimate. Specifically, "participants appear to be different from nonparticipants in several interesting ways. Participants tend to look to outside sources of expertise for advice while nonparticipants are do-it-yourselfers." This finding suggests that free-ridership estimates that are based on comparisons to nonparticipant actions may be flawed, because participants may need the program to undertake the retrofit measures, while nonparticipants may not (Freeman Research Resources 1991).

In some cases, the differences between participants and nonparticipants will be less dramatic and will not seriously affect results. If differences are significant, evaluators can correct for self-selection bias. Methods for doing so include comparing energy consumption between participants and nonparticipants using multiple comparison groups and simultaneous-equation models. The models generally require the estimation of two equations, a discrete-choice participation model and an energy-use equation (Chapter 4).

Another approach is to monitor the activity of a nonparticipating region with similar characteristics to the test market. Wisconsin Electric used this approach to evaluate a program to encourage C&I customers to purchase energy-efficient fluorescent lamps and ballasts. Equipment sales were monitored in the area where the rebate was available (Milwaukee) and in a region with similar characteristics (e.g., population and electric rates) but without a lighting-efficiency program (Fig. 16). Cincinnati was selected as the comparison region. Sales of efficient lighting fixtures remained nearly constant in Cincinnati,

suggesting that the increase in the sale of efficient lighting in Milwaukee during the program was directly attributable to the program (Brooker and Fichtner 1991).

Deferred free riders should be considered when comparison groups are used (as well as with other methods). In a comparison-group analysis, the effects of deferred free riders would appear as a decrease in net (program-induced) energy savings over time; see Chapter 7.

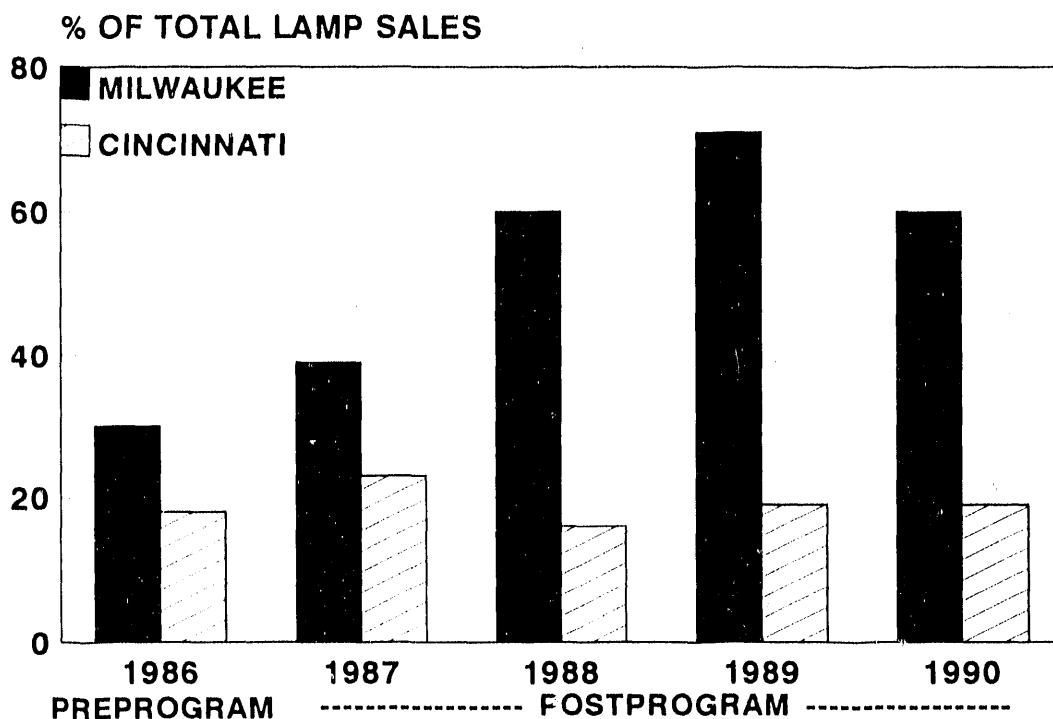


Fig. 16. Comparison of sales of energy-efficient lamps in Milwaukee, Wisconsin, where a program operated and Cincinnati, Ohio, where no such program operated.

STATISTICAL MODELS TO ESTIMATE FREE RIDERS

Some evaluators have used statistical models to estimate free ridership (Chapter 4). Generally, this method involves analysis of the energy-related actions, characteristics, and attitudes for samples of participants and similar nonparticipants. Simulations are developed to predict the likelihood of the adoption of program-sponsored measures with and without the program. The two estimates are then used to calculate the free-rider ratio.

Niagara Mohawk used this approach to estimate the free-rider rate in its Low Cost Measures Program (Regional Economic Research 1991). This program offered low-flow showerheads, pipe wrap, water-heater blankets, and compact fluorescent bulbs to residential customers. With data from a survey of participants and nonparticipants, an equation was developed to estimate the likelihood of the adoption of these measures. Characteristics like

income, household size, education level, and attitudes toward conservation were used to adjust for differences in participants and nonparticipants (self-selection bias).

The estimated model (Table 27) was used to simulate the predicted adoption rates of program participants under two scenarios: customers participate in the program and they do not. This dual simulation involves solving the model for the participant sample with different values for the participation variable (yes vs no).

Table 27. Example of statistical model used to estimate free ridership in a residential program

$$\text{ADOPT}_{ik} = A_k(\text{OWN}_i, \text{INC}_i, \text{NUMINHH}_i, \text{EDUC}_i, \text{PART}_i, \text{ATT}_{1i}, \dots, \text{ATT}_{ji}, \text{MILLSRAT}_i),$$

where:

ADOPT_{ik} = a binary variable reflecting the adoption of measure k by household i

OWN_i = a binary variable reflecting home ownership

INC_i = income of household i

NUMINHH_i = household size

EDUC_i = education level

ATT_{ji} = an attitude index (j) relating to conservation

PART_i = participation of the household in the program

MILLSRAT_i = Mills ratio (used to correct for self-selection)

Because the dependent variable is binary, a logit formulation was used, which assumes the logistic functional form:

$$\text{ADOPT}_{ik} = 1/(1 + e^{-h_k(x)}) ,$$

where: $h_k(x) = h_k(\text{OWN}_i, \text{INC}_i, \text{NUMINHH}_i, \text{EDUC}_i, \text{PART}_i, \text{ATT}_{1i}, \dots, \text{ATT}_{ji}, \text{MILLSRAT}_i)$

Source: Regional Economic Research, Inc. (1991)

The specifics of the model will vary with program type and availability of data. For example, in a residential loan program, factors like the number of people over age 75, the number of people at home at night, and the importance of comfort were considered (Ozog 1991). In modeling a commercial-sector program, variables would likely include the type of business and its energy-consumption history (Regional Economic Research 1990).

While these techniques hold promise, they are hindered by the quality of the data on which they are based. These data usually suffer from the same problems as other evaluation techniques. Regional Economic Research (1991) noted, "The adoptions model was estimated using the results of the tracking survey. While the survey yields a relatively small sample size, it is the only available consistent source of information on adoptions and nonparticipants. In the future, with more data points, our free-rider estimates may change."

FREE DRIVERS

The opposite of a free rider is a free driver. A free driver contributes to the goals of the program (e.g., reduce energy consumption) but is not formally a program participant. A free driver is affected by the program either through a conscious awareness of the program or because of program-induced changes in the marketplace. Free drivers require evaluators that use comparison groups to consider whether the comparison group is actually taking the conservation actions because of the program (i.e., what is traditionally considered a free rider may turn out to be a free driver); see Chapter 7.

An example of a free driver is a customer who purchases a product that qualifies for a rebate but does not claim the rebate. Northern States Power Company's analysis of its appliance rebate program showed that only about 40% of the customers that purchased a qualifying appliance applied for the rebate (Brian C. and William Lesh Inc. 1986). In some cases the customers may not have been aware of the program, and in other cases may have decided it was not worth the effort to apply for the rebate.

Free drivers also occur in new-construction programs. Research conducted on such programs in Maine and Wisconsin found that they affected the actions of nonparticipants. Specifically, some builders constructed homes to program standards for competitive reasons but chose not to participate in the program. As a result, an examination of nonparticipants would not accurately represent what would have happened if the program did not exist (Violette, Ozog, and Wear 1991).

Research on free drivers is limited. One way to identify free drivers is to ask nonparticipants if they are aware of the utility's DSM program and, if so, whether the program influenced their adoption of program-recommended actions. Unfortunately, customers may be influenced by a program even if they are unaware of its existence. For example, wholesalers and retailers may stock more of the energy-efficient appliances because of a utility's rebate program; nonparticipants may purchase these units even if they are unaware of the program simply because the efficient ones are in greater abundance and more prominently displayed in the showroom.

Another approach is to use comparison groups consisting of communities outside the area in which the program is offered. Comparing the distributions of efficiencies for program-sponsored measures in the participating area with those in the comparison area might show the overall effect of the program, including (implicitly) the contributions from free drivers.

Another approach is to survey the trade allies involved in a program. Their responses to program-induced changes, both at the distributor level and the customer level, might identify the degree of free ridership in a particular program. For example, a retailer not participating in a program might stock the more efficient units anyway to remain competitive with the participating retailers, an example of free ridership among trade allies.

DISCUSSION AND CONCLUSIONS

It may not always be necessary to conduct rigorous analysis of free ridership because the importance of free riders varies with program type. Free-rider measurement is usually more important in determining program cost-effectiveness where the benefit/cost ratio is close to one.

An appliance rebate program offering generous rebates for the purchase of energy-efficient products would be expected to be sensitive to free-rider rates in determining net program impacts. Not surprisingly, most of the free-rider research has been targeted toward this type of program. On the other hand, a low-cost energy-outreach program designed to encourage people to insulate their attics could tolerate a high level of free ridership because the cost of the program is low and the benefits are high. Even more costly programs, such as those offering energy audits, may be able to tolerate free-rider rates of 50% or more and remain cost-effective. In BPA's industrial programs, program designers were generally not concerned about free riders because, according to Keating (1990):

The savings are so inexpensive that the cost-effectiveness limit wouldn't even be approached at 50% free riders and there was a clear possibility that without the utility's intervention and quality control, the measures that might have been installed would not have been as comprehensive or aggressive. [In addition] the utility did not want to risk the loss of such large blocks of savings (1-8,000,000 kWh/yr) if the evaluators were wrong about the willingness of the participants to install the measures without the program.

DSM programs aimed at low-income customers have experienced low free-rider rates. Research suggests that low-income households don't have the money to implement major conservation measures without assistance. Most studies of such programs do not incorporate a free-rider test, and those that have often found rates below 15% (Lui and Fang 1989).

Free-rider levels may also be unimportant if a program is designed to move a specific market (e.g., encourage retailers to stock energy-efficient light bulbs and high-efficiency appliances). If this goal is achieved, the free-rider rate is irrelevant.

One way of dealing with free riders is to design programs to reduce their impact. Free-rider research has been valuable in this regard. For example, in a commercial lighting program offered by NEES, the free-rider rate was quite high, about 65%. Rather than eliminating the program, the program was modified. The first change was to require preinstallation inspections to ensure that potential participants did not already have efficient lamps and ballasts and merely wanted to replace worn-out equipment. This type of customer would readily fit the description of a free rider. The utility also encouraged the use of "advanced" lighting measures that typically have free-rider rates considerably lower than simple conversions from incandescent to fluorescent fixtures. The free-rider rate for regular fluorescent lamps was 65%, but for compact fluorescent lamps only 5%. In general, free-rider rates tend to be higher for products with high market shares and lower for products with low market shares (Nadel 1990).

Research conducted on a rebate program for C&I customers operated by Southern California Edison shows how different segments of the target market reacted to the program. The program gave rebates for the installation of energy-saving equipment. The customers who asked to be audited (as opposed to a utility representative's initiating the audit) had a greater tendency to install recommended measures (Train, Ignelzi, and Kumm 1985). These results show that free ridership is a function not just of program design but also of the type of customer participating in the program. Evaluation results can be used to help utilities better target their programs to customer groups with low free-rider rates.

Markets, customer attitudes, and actions are not constant. A free-rider rate of 10% today may be 50% tomorrow. Despite the problems associated with free-rider measurement, if questions are asked over a long time with the same measurement techniques, important trends can be detected even though the absolute accuracy of the estimate may be questioned.

Measuring and dealing with the free-rider issue is not simple. The data can be inconclusive and contradictory, which makes it risky to depend on a single method of collecting and analyzing free-rider data. The challenge for the evaluation community is to improve techniques for free-rider measurement. The challenge for those developing DSM programs is to reduce the free-rider impact in the design stage.

PROCESS EVALUATION OF DSM PROGRAMS

**Benson Bronfman
Jane Peters**

BACKGROUND

Process evaluation is the review and assessment of program implementation. Originally applied to federal housing and education programs, process evaluation plays a central role in the evaluation of DSM programs. The major characteristic of DSM programs that distinguishes them from power plants is the distributed nature of DSM-program activities and effects. DSM programs are centrally planned, but are implemented by field staff, contractors, and customers.

Initially, the focus of process evaluation was on program operations and structure, and such evaluations provided program managers with a documented history of the program. Currently, the goal of process evaluation is program optimization through

- Improvement in implementation efficiency
- Assessment of market segments and targeting of specific segments
- Improvement in quality of measure installation
- Identification of program-design issues
- Interim accounting of program progress through reviews of the program database
- Examination of special issues, such as measure life and program comprehensiveness

Despite the creative uses of process evaluation, too often process evaluations are narrowly defined to include only customer response and satisfaction. While satisfaction is clearly important, programs could achieve high satisfaction but acquire few DSM resources. Measuring customer satisfaction, in and of itself, does not represent the current practice of process evaluation.

Evaluation includes a continuum of activities designed to measure program impacts and to improve program delivery. Because of this range, the attribution of activities exclusively to "impact" or "process" practice is somewhat arbitrary. Impact evaluations often include surveys of consumer satisfaction and behavior. Process evaluations frequently document program activities (e.g., measures installed and their actual costs) and estimate cost-

effectiveness. This review of process evaluation takes a broad view of the field, including some activities that might belong on the impact side of the ledger.

As DSM-program performance has become more closely tied to integrated resource planning and incentive regulation, process evaluation has assumed more importance. First, because many utilities have only recently come to the DSM-resource-acquisition philosophy, many programs are designed and implemented quickly. Even with expert assistance in program design, early process evaluations are critical to fine-tune programs, assess the market potential, and identify market segments not reached. Second, because process evaluations are in the field early and because impact evaluations often do not produce results for a year or two, process evaluations are instrumental in producing early reports on program effects and effectiveness. This activity may include assistance in designing the program database, assessing the reliability of program data, and periodic reporting of implementation data (e.g., the number of installations, estimated and actual costs, and projected savings).

Process evaluations are generally conducted once a year for each program. The topics to be addressed vary from year to year, depending on the stage of the program and the types of issues that are currently important to the program. Start-up evaluations often address program design, contract-negotiation procedures, and internal organization. Short-term evaluations can be conducted to address special issues, such as changes in building codes, the effects of utility reorganization, or changes in the marketplace. Comprehensive evaluations may include a variety of projects designed to address specific problems from start-up to special issues, finally combining all the results into a comprehensive report on the program.

Process evaluations should be conducted with guarantees of confidentiality to all parties providing information. This guarantee will help demonstrate that the evaluation will be used to improve the program. It also serves to assure customers that the utility is concerned with the program as a whole and not with their particular response to the program. By assuring all contacts that the information is confidential, greater objectivity and more insightful reports will be obtained.

Process evaluations can be conducted by utility staff or by contractors. Utilities have often conducted their own process evaluations because of the sensitive nature of the information and because process evaluations are primarily for internal use. However, the use of contractors is becoming more accepted because contractors can provide an alternative and confidential channel for communication among program staff, participants, contractors, and utility decision makers. An additional benefit of contractors often occurs after a utility reorganization. Contractors frequently serve as the corporate memory during these changes.

KEY ISSUES

Process evaluations are broadly defined as the assessment of program implementation. A comprehensive process evaluation examines a program at different stages. This section discusses the planning and design stage as well as the implementation stage (Exhibit 10).

Exhibit 10. Process evaluation of a DSM-bidding program

In 1988, Orange & Rockland Utilities proposed to run an experimental bidding program for DSM resources. The New York State Energy R&D Authority assisted in the design of the program and sponsored an evaluation of the program to inform the other utilities in New York (Environmental and Energy Services Company and Pacific Energy Associates 1989; Peters, Barry, Horowitz, and Gordon 1990).

The evaluation included two phases. The first addressed program planning and design, contractor selection, and contract negotiations. The second addressed the first year of implementation, including an analysis of project savings and costs. The evaluation addresses seven major issues:

1. **Program Goals and Design.** Program goals provide a framework to determine whether and how policy intentions are framed within the program design; whether the program is being implemented as intended; how the goals are perceived by contractors; and, ultimately, whether the program will be manageable as it is implemented.
2. **Request for Proposals (RFP) Process.** With surveys, the bidding process was reviewed to determine the level of competitiveness and whether program goals were achieved in the selection process. The reasons for nonresponse were used to determine if the utility had an appropriate list of potential bidders and whether any features of the program or RFP presented barriers to response.
3. **Delivery Mechanism.** The major issues to assess were the costs borne by the contractors and customers, the relationship between the utility and the contractors, the relationship between the contractors and utility customers, and the ability of program contractors to deliver the DSM savings.
4. **Administrative Mechanism.** The administrative components of the program included development of the RFP, the contracting process between the utility and the ESCOs, the administrative costs borne by different parties, the cost- and quality-control procedures used by the utility, and the savings-verification process.
5. **Customer Response.** Customer response and satisfaction are key ingredients to program success. However, the utility was not a direct party to the customer's experience because the program was delivered by ESCOs. The evaluation obtained the customers' view of the utility's role in the program, their response to marketing materials, and their view of the ESCO role. In addition, the evaluation addresses the effects of franchising on customers, contractors, and the utility.
6. **Implementation.** This phase estimated energy and demand savings and program costs. These data were then used to assess cost-effectiveness.
7. **Measurement Plans.** The validity of the cost-effectiveness analysis rests on the accuracy of the cost and savings data for the installed measures. The cost data were obtained during the evaluation of the administrative and delivery mechanisms. Collecting actual savings data requires detailed estimates of the impact of the installed measures on consumption and demand. Each contractor was required to develop a measurement plan. This phase of the evaluation reviewed the plans and determined if additional data were required to assure the validity and reliability of the savings data.

Planning and Design

Process evaluations can address program-design issues before the program is in the field. Evaluations conducted during this stage can lead to early program enhancement and more effective long-term implementation (Peters, Gustafson, and Vowles 1987; Peters, Barry, Horowitz, and Gordon 1990). The most critical issue for investigation during this period is to determine whether the goals of the program are explicit and whether they are being met. Lack of clear articulation of goals can have severe implications for future evaluations of program performance (Lerman and Bronfman 1986a). Process evaluations are designed to determine if programs are achieving their goals. Frequently, program goals change over time. These changes should be documented, and program design should be reviewed to determine whether it supports the new goals.

Assessing the planning and design process requires early review and feedback. The evaluator seeks to determine whether the design process successfully addressed all parties' concerns and whether the resulting program has sufficient staff resources and commitment for effective implementation. Experience at other utilities with similar programs is usually brought to bear to review the proposed program structure, delivery mechanism, and administration.

Implementation

Program implementation includes administration, program-delivery mechanism, ongoing program activities, and customer response to a program. Most process evaluations focus on this phase, which can be repeated at annual intervals to track the program over time. For example, Lerman and Bronfman (1984); Lerman, Bronfman, and Tonn (1983) and Synergic Resources Corporation (1991) conducted process evaluations of BPA's Residential Weatherization Program as it developed from a pilot to an interim to a long-term program. In another example, the early adoption of model conservation standards in Tacoma, Washington, led to the implementation of a kick-off evaluation as the program was starting, and a revisit to examine implementation as the program matured (Lerman and Bronfman 1986b). Several process evaluations that tracked the evolution of an industrial retrofit program are chronicled in Evans and Peters (1989).

Administration mechanisms range from contractual agreements between the utility and customers or contractors, to payment procedures and quality- and cost-control procedures. Issues about contractual agreements are only addressed when contracts must be signed, while payment and quality- and cost-control procedures should be monitored throughout the program to assure that they are functioning well.

Customer response to a program determines whether it is effective or not. Customer response includes such issues as free riders, take back, free drivers, market effects, and market segmentation. If customers are dissatisfied with the program or if they find participation difficult, the program may require redesign.

How the program is organized and how the utility manages a program are critical aspects of program process. For example, if the utility permits districts to implement the program in their own way, several programs will be in the field, and a variety of explanations will exist for program effects (Fig. 17) (Lerman, Bronfman, and Tonn 1983). Reviews of how utility staff communicate within a program can reveal interesting information about implementation. For example, Lerman and Bronfman (1984) showed that the delay in resolving technical issues and communicating those results to field staff led to ad hoc decision making in the field and infrequent use of the required technical-issue-resolution process.

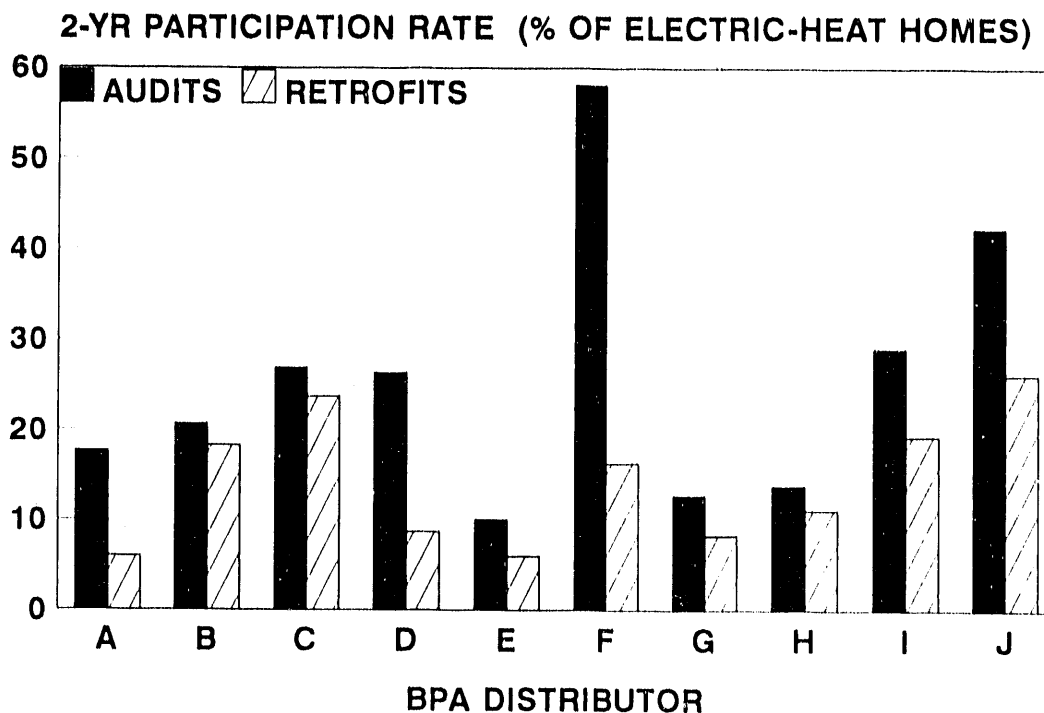


Fig. 17. Customer participation rates across the ten utilities participating in the BPA pilot Residential Weatherization Program. Because these utilities each implemented the program differently, customer response to the energy audit and financing of retrofits differed substantially.

A critical question for process evaluations is, "How much of the target market participated in the program?" Market segmentation helps to identify appropriate groups of customers to target for a particular program. The most common approach to market segmentation is to link attitudinal and demographic information obtained from surveys or geographically with specialized databases (Van Liere and McKinnell 1989; Feldman 1988). Process evaluations seek to determine whether the expected segments participated in the program, whether the segmentation strategy was effective, and how it can be improved. For example, one evaluation found that the structure of investment decision making among manufacturers and their place in the annual budget cycle was much more important in determining program participation than was company size or type.

Barriers to participation are the factors that prevent customers from participating in a program. A difficult sign-up process or confusion about program requirements or benefits can limit participation. Two evaluations examined the low response rates to RFPs for a third-party financing commercial program. These evaluations showed that the low response rates were, in large part, caused by the many firms receiving the RFP who were either not qualified to respond or had no intention to respond.

Process evaluations address nonparticipating customers as well as participating customers to identify the program features that are least attractive or that require modification to improve participation rates. Two evaluations examined the barriers to participation in commercial retrofit and new-construction programs (Freeman Research Resources 1991; Charles Rivers Associates 1991). Both studies linked reasons for participation and nonparticipation to specific program features.

Utilities often want to estimate the number of free riders (Chapter 8). Identification of free riders is often accomplished through surveys of participants and nonparticipants (Charles Rivers Associates 1991; Freeman Research Services 1991).

Customers must be aware of the program before they can participate. Measurement of program awareness focuses on nonparticipants in the target market. Awareness can mean that they have heard of the program or that they are aware the utility conducts energy-efficiency programs in general. Awareness following media campaigns can be expected to rise. Determining how much rise occurs after different types of campaigns can help utilities select the most appropriate strategies for future marketing efforts (Birnbaum and Davis 1989; Columbia Information Systems 1988).

Customer satisfaction includes general satisfaction with the utility and satisfaction with specific programs. Changes in customer satisfaction can be measured by repeated studies over time. Satisfaction with DSM programs is critical for understanding the benefits of the program and making modifications to improve customer response. If the utility conducts general-satisfaction studies as well, participants can be compared to nonparticipants (Peters, Haeri, and Seratt 1990).

Utilities frequently examine attitudes and beliefs to determine how they affect participation and how participation affects attitudes. For instance, it was found that attitudes toward the utility can be improved and customers' sense of control over energy use can be increased through participation in residential low-income programs.

During the lifetime of a program, the particular technologies promoted by the program can advance beyond the levels included in the program, or the energy-efficient technology can become the dominant option offered in the marketplace. Measuring these changes is an important part of process evaluations. Studies of retailer and distributor inventories, mystery-shopper activities, and discussions with equipment vendors and dealers can help determine whether a program has outlived its usefulness or whether the installation of new equipment should be encouraged through changes in program design. Two process

evaluations showed, for example, that more than 75% of eligible street lighting had been replaced in BPA's Street and Area Lighting Program (Peters and Bronfman 1986) and that the high-efficiency water heaters were the only options available in Washington and Oregon (Lerman 1987).

Program costs include incentives, utility staff, and contractor expenses. In addition, customers may have costs associated with the program. Surveys can be used to identify costs that are not recorded and to estimate costs. For example, Peters, Haeri and Gustafson (1988) estimated administrative costs for three commercial programs with a combination of utility records, surveys, and in-person interviews.

DATA COLLECTION AND ANALYSIS

Process evaluation relies on traditional social-science methods to collect and analyze data. Data sources include utility staff, trade allies, program participants, nonparticipants, implementation contractors, program reports, program databases, and data from other utilities conducting similar programs. Data-collection methods include structured in-person interviews, group interviews (e.g., focus groups), mail and telephone surveys, direct observation, and literature and document reviews. Analytic techniques range from narrative description and anecdotal reporting to simple and multivariate statistical analyses.

Data

Identification of the appropriate data sources requires a sound understanding of the program's design. The evaluation staff must identify the key people involved in the program within the utility. The following questions (see also Exhibit 11) can be used to identify these key people.

- Who designed the program?
- Who approved the program design?
- Who delivers the program to customers?
- Who administers contracts and rebates?
- Who interprets questions about the program?

Once the correct contacts are identified, program documents and communications among these people should be located. Utility contacts can assist evaluation staff to identify the role of contractors and trade allies in implementing the program. The program database and monitoring system provide additional information.

A process evaluation of an all-source bidding program identified six target populations (Environmental and Energy Services Company and Pacific Energy Associates 1990):

Exhibit 11. Key objectives of process evaluations for Boston Edison Company

Boston Edison (1991) prepared a series of questions to address in evaluating each of its DSM programs. The key objectives for evaluation of a residential program are:

- Document the history and progress of the program
- Assess the promotion and delivery system for use in designing future programs
- Assess installation-contractor effectiveness and comprehensiveness in program delivery
- Assess building-owner and occupant satisfaction and attitudes toward the program and its products and services
- Identify barriers to program penetration
- Assess the market-driven versus program-motivated efficiency actions of customers to estimate the numbers of free drivers and free riders
- Evaluate product removal and retention to estimate persistence of savings
- Identify and compare the characteristics of owner and occupant participants and nonparticipants
- Assess the remaining market potential
- Examine the roles played by contractors, community organizations, building owners, occupants, and Boston Edison staff in the design, marketing, and implementation of the program
- Assess the educational components of the program and its effect on participating customers

- Regulators in New York and New Jersey involved in planning the program
- Utility staff involved in the design, development, and implementation of the program
- ESCOs and utility customers submitting proposals and signing contracts to deliver DSM resources
- Nonresponding ESCOs and customers (those on the RFP mailing list who did not submit a proposal)
- Participating customers who received program services from ESCOs
- Nonparticipating customers who considered participation but then decided against it

Selection of the appropriate number of contacts and samples for surveys requires understanding the difference between sampling for information and sampling for statistical reliability. Sampling for information provides a wide enough range of experience with the program to cover all points of view. This is the approach generally taken with key-contact sampling for the utility staff, implementation contractors, and small numbers of participants.

Sampling for statistical reliability assures that statistical analyses yield the desired precision in results (Chapters 3 and 4). Statistical reliability is especially important in impact evaluations. Samples should be constructed so that the results can be generalized to the population as a whole. While general guidelines suggest that a sample of 200 to 500 is sufficient in most cases, some situations require larger samples to ensure that every market segment is included in the analysis.

If the program is new to the utility, drawing on the experience of other utilities to identify issues and approaches for solving problems is appropriate. Literature reviews and surveys of other utilities' experiences can also be valuable when addressing difficult issues, such as free ridership or measure persistence.

Process evaluations can occur throughout the life of a program, including the first few months of program implementation. Such evaluations are used to obtain an early assessment of the program process. They are conducted over a short time and rely on interviews as the primary data-collection method.

Process evaluations conducted during the first few years of the program are generally comprehensive. These evaluations focus on both planning and implementation, and are often coordinated with a preliminary impact evaluation. Such evaluations require extensive interviews and surveys with many people involved in the program.

Evaluations that occur in the middle and later periods of the program might focus on specific issues of implementation or whether specific recommendations have been adopted. Process evaluations might collect data required for the impact evaluation or to determine if the market has changed so that the program is no longer required. These special evaluations require targeted interviews and surveys of appropriate populations.

The final stage of a program may require no evaluation at all. If only limited studies have been conducted, however, a final, comprehensive evaluation may be undertaken. Such an evaluation can help interpret impact data and provide a resource for utility staff who might wish to build upon the experience of the program in the future. These evaluations would also require extensive interviews with many contacts and surveys of large samples of participants and nonparticipants.

The depth of the analysis required will also influence the selection of a data-collection approach. In general, the greatest depth is gained from in-person interviews, which can provide detailed and candid accounts of the program process. Surveys, focus groups, and document reviews provide less detail but permit access to a greater number of contacts. In

programs with many participants, depth must generally be discounted to achieve breadth of coverage for participant and nonparticipant views.

The number of contacts is an important factor in determining the breadth of coverage required in collecting data. If there are many potential contacts, those with unique and comprehensive familiarity with the program should be identified. These contacts will be the most appropriate for in-person interviews. Field staff or trade allies with comprehensive implementation experience can be effectively interviewed in groups. Surveys are most appropriate when breadth is required, such as with large groups of participants, nonparticipants, and trade allies.

Analysis

Analysis techniques depend on the types of data collected and on the questions asked by the process evaluation. The fundamental requirement of a process evaluation is to explain not only the who, what, and where of the program, but also why things happened as they did.

Quantitative techniques are generally applied to surveys and program records. These techniques include simple descriptive statistics, such as counts, tabulations, and measures of central tendency. Statistical tests, such as t-tests, chi-square tests, and multivariate analyses (e.g., conjoint analysis, regression analysis, factor and cluster analysis, and discrete-choice analysis), are also used.

Simple descriptive statistics are applicable to most survey and program data and even to some interview data. In fact, most who, what, and where questions are answered with counts and tabulations. Statistical tests and multivariate analyses, however, require careful construction of survey instruments and sample-selection procedures to assure that the appropriate data are available for analysis.

Qualitative techniques are used in all process evaluations to analyze interviews, focus groups, and program documents and to integrate quantitative data with these data sources. Qualitative analysis requires a structured and detailed organizational framework. Data are derived from a variety of sources requiring the analyst to systematically organize and classify the data into the analysis framework.

Miles and Huberman (1984) provide a detailed sourcebook on qualitative data-analysis methods; they describe three components of the analysis: data reduction, data display, and conclusion drawing and verification.

Data reduction transforms raw data into usable material. This reduction includes, for example, reviewing interviews to summarize and synthesize key issues and to formulate new questions. This process transforms the data into a presentation format from which conclusions can be drawn. For example, responses to a series of questions can be examined for key contact groups, or a matrix can be developed to tabulate responses across a variety of issues.

The final process is conclusion drawing and verification. Each of the identified issues must be checked through the data-collection and -analysis process and verified by more than one source. Eventually, the analyst has confidence that the conclusions are well grounded and documented and can be considered reliable and valid. Patton (1982b, 1990) describes qualitative analysis as a process of triangulation in which data from multiple sources are compared and contrasted. Each finding and conclusion is substantiated by a variety of sources, providing confidence that the findings are valid and reliable.

REPORTING AND USING RESULTS

After the data have been collected and analyzed, the findings should be reported to the utility. Evaluation results can be presented formally or informally (Chapter 5). Informal evaluations may result in memo reports or presentations and often provide information for implementation of recommendations prior to the release of a formal report. Formal evaluations always result in written reports. The reports and presentations transfer evaluation findings to those who can use the results (Evans and Peters 1989). These include program managers, program planners, forecasters, utility managers, and nonutility parties.

A commitment to use results is required for the evaluation to be effective. Process evaluations can be viewed either as a report card or as strategic information. As a report card, staff may fear the findings and try to manipulate the results to give a favorable view of the program. As strategic information, the evaluation is seen as providing insights to assist in improving the program's operation and effectiveness. This second point of view encourages evaluators to conduct an insightful and penetrating analysis of the program.

LINKS BETWEEN PROCESS AND IMPACT EVALUATIONS

Process and impact evaluation are intimately linked. While certain activities are peculiar to each (impact evaluation measures actual savings, and process evaluation assesses organizational efficiency), they often deal with the same issues and populations. Often, information gathered through one activity is critical for another (Fig. 18). For example, a major evaluation of DOE's low-income WAP is using process-evaluation methods to collect and analyze data on the determinants of program energy savings and cost-effectiveness. These factors include regional differences, dwelling-unit characteristics, occupant characteristics, types of measures installed, service-delivery differences, methods of client selection and education, and use of diagnostic procedures (Beschen and Brown 1991).

In conducting process evaluations, program implementation issues are frequently discovered that might affect how the impact evaluation is done. As illustrations, program implementation approaches might differ across utility districts (Fig. 17), marketing approaches might differ, or the types of customers might differ. Early identification of these variations allows the impact evaluation to better identify appropriate control groups or to interpret analytical results more intelligently.

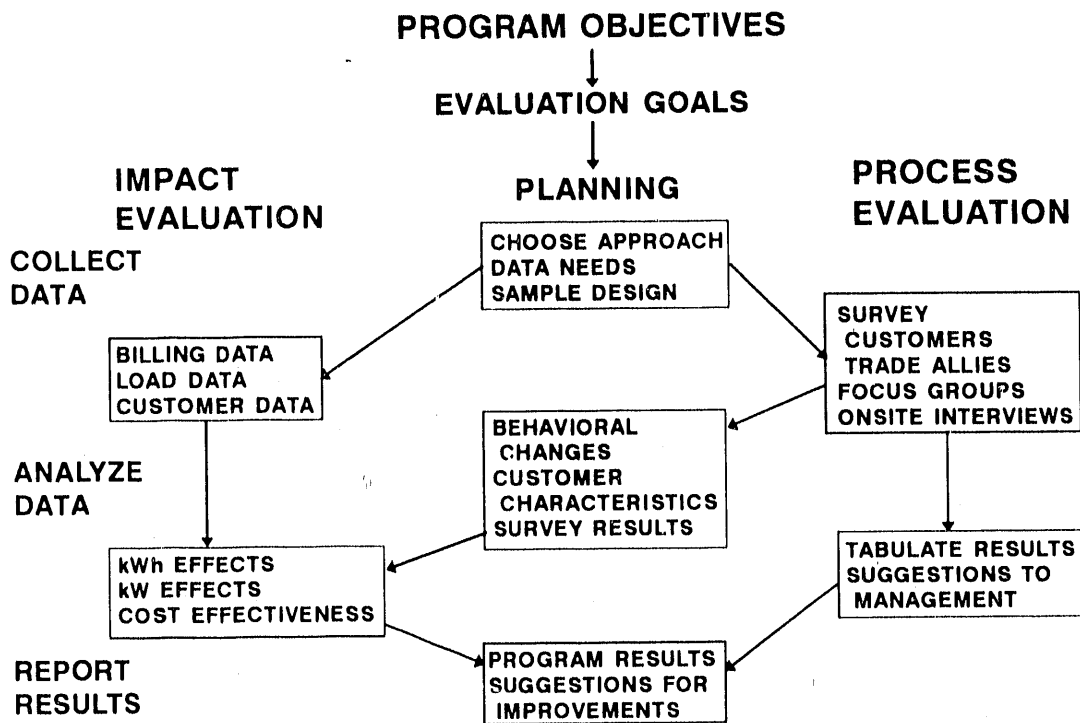


Fig. 18. Although process and impact evaluations differ in their underlying purpose, they both deal with the behavioral changes that explain the observed outcomes and offer suggestions for program improvement.

Because process evaluation personnel are "on the ground" early in the program, issues of data quality and availability can be addressed. Will the data needed for an impact evaluation be available in a timely manner? Evaluators assume that basic data (such as monthly electricity bills) are available for utility customers in recognizable form. This assumption is often incorrect. It is not unusual, for example, to find no data on customer contribution to the cost of DSM measures, to discover that entire classes of customer have meters read sporadically (with electricity use estimated the remainder of the time), to find billing histories archived in obscure formats, or to find that only a fraction of the data is digitized (and the rest exists only on paper at widely scattered district offices).

Often, no corporate memory of prior evaluations or data systems exists. Usable data are often discovered through the process evaluation. Early identification of problems and opportunities regarding data systems allows corrections to be made early enough to be useful in impact evaluation. Or it allows the impact evaluator enough time to change the evaluation strategy. For example, BPA's Commercial Audit Program evolved from a resource-acquisition program to a data-gathering/resource-assessment program to a field test of alternative auditing methods. Little documentation of these changes was made. The resulting database was limited in its usefulness when impact evaluations were implemented (Lerman and Bronfman 1986a; Cambridge Systematics and ERCE 1989).

Process and impact evaluations often address the same participant and nonparticipant customers. Impact evaluators may survey samples of participants and nonparticipants regarding decisions to participate, nonprogrammatic actions, building energy-use patterns, and equipment/appliance portfolios. Process evaluations may also deal with decisions to participate, participant and nonparticipant characteristics, attitudes, satisfaction with services, and intentions to take action. Many different purposes can be combined into one survey instrument. Again, with process evaluations in the field early, additional data can be gathered through follow-up surveys or site visits to make the data more current or to add information on new issues, such as measure life.

A few residential retrofit programs have been in operation for six to eight years. In situations like this, where year-to-year savings have been stable, impact evaluations may not be performed on an annual basis. Rather, periodic process evaluations are conducted, focusing on specific issues, such as market acceptance of changing incentives or measures offered. These "maintenance" evaluations and special projects can trigger new impact evaluations if results are substantially different from expectations.

Process evaluation has assumed a larger role in utility program evaluation, especially in settings where integrated resource planning and incentive regulation are in force. Utilities often do not have the time to pilot-test programs, to wait several years until the first impact results are available, or to conduct a priori market assessments or DSM-resource assessments. By looking at the goals of program evaluation, rather than the methods used to answer specific questions, we find that process evaluation is becoming a more universal tool in addressing utility DSM concerns.

Table 28 shows the key issues related to evaluations of DSM programs. The two main goals of evaluation are to estimate program savings and to optimize program design and delivery. Under each of these headings are the issues that utilities might examine. While several issues are addressed by one evaluation approach, others can be addressed by both process- and impact-evaluation methods. While energy-savings and net-savings analyses are addressed with only impact-evaluation methods, other issues (like persistence of savings and measure life) require both.

CONCLUSIONS

Process evaluations can help program planners, designers, and implementors in determining appropriate ways to improve DSM programs. Process evaluations also provide valuable information for deciding when a program is no longer required and when program efforts should be enhanced. Perhaps most importantly, such evaluations provide a vital tool for interpreting the results of impact evaluation.

Process evaluation, at its most basic, is the assessment of program implementation for the purpose of improving program delivery. With the increasing importance of DSM programs and the growing scrutiny by utility staff, regulators, and outside parties, process evaluations

are becoming valuable tools because of their timing, flexibility, and easy integration into program operations.

Table 28. DSM-program-evaluation components and methods

	Impact evaluation	Process evaluation
Program savings		
Load-shape impacts	✓	
Total savings	✓	
Net savings	✓	
Attribution of effects	✓	
Persistence of savings	✓	✓
Cost-effectiveness	✓	✓
Program Optimization		
Infrastructure support		✓
Measure life	✓	✓
Program implementation		✓
Market assessment		✓
Technology assessment	✓	✓
Customer preference and behavior		✓

ACKNOWLEDGMENTS

Sam Swanson, in 1988, suggested a project that ultimately resulted in this handbook. David Wolcott, in late 1990, organized a committee, consisting of Martin Cummings, William Saxonis, William Vinhage, David Wooley, and himself, which developed the outline and suggested authors for this handbook. David Wolcott and William Vinhage organized a conference, held in March 1991, that brought together all the authors of this volume. Several people, including Frank Congedo, Martin Cummings, William Eastlake, Tamarica Eichelburg, Deborah Ross, and Sam Swanson reviewed outlines and drafts of the chapters. Fred O'Hara edited the final handbook. And Ethel Schorn prepared the final manuscript and ensured that all the details (especially the lengthy list of references) were complete and in the proper format. Finally, we thank David Wolcott and Diane Pirkey, the project managers, for their support throughout development of this handbook.

REFERENCES

Applied Energy Group, Inc., *Massachusetts Joint Utility End-Use Monitoring Project*, Final Report, Woodbury, NY, August, 1988.

Applied Energy Group, Inc. "Commercial Dollars and Sense Program Evaluation," *Impact Evaluation of Long Island Lighting Company's Electric Conservation and Load-Management Plan*, Woodbury, NY, 1989.

Applied Management Sciences, Inc. *Central Hudson Commercial/Industrial Incentives Program - Final Evaluation*, November, 1990a.

Applied Management Sciences, Inc. *Process Evaluation of Worcester Energy Fitness Program*, prepared for New England Electric System, Westborough, MA, March, 1990b.

Arkin, H. *Handbook of Sampling for Auditing and Accounting*, 2nd ed., McGraw-Hill, 1974.

Association of Demand-Side Management Professionals. "Are You Ready for DEEP Data?", *Strategies* 2(3), Berkeley, CA, Summer, 1991, p. 16.

Berry, L. *The Administrative Costs of Energy Conservation Programs*, ORNL/CON-294, Oak Ridge, TN: Oak Ridge National Laboratory, November, 1989.

Berry, L. *The Market Penetration of Energy-Efficiency Programs*, ORNL/CON-299, Oak Ridge, TN: Oak Ridge National Laboratory, April, 1990.

Beschen, D. A. and M. A. Brown. *Evaluation Plan for the Weatherization Assistance Program*, ORNL/TM-11668/V1, Oak Ridge, TN: Oak Ridge National Laboratory, August, 1991.

Birnbaum, I. and T. Davis. "Using Commercials to Sell DSM Programs: A Measurement of Customer Response," *Proceedings, Energy Program Evaluation: Conservation and Resource Management*, Argonne, IL: Argonne National Laboratory, 1989, pp. 195-200.

Blevins, R. P. and B. A. Miller. *1987 Survey of Commercial-Sector Demand-Side Management Programs*, EPRI CU-6294, Palo Alto, CA: Electric Power Research Institute, March, 1989a.

Blevins, R. P. and B. A. Miller. *1988 Survey of Residential-Sector Demand-Side Management Programs*, EPRI CU-6546, Palo Alto, CA: Electric Power Research Institute, October, 1989b.

Bonneville Power Administration (BPA). *Guidelines for Proposing an Energy Savings Plan Efficiency Project*, Portland, OR: Bonneville Power Administration, July, 1988.

Bonneville Power Administration (BPA). *Evaluation of the Commercial Incentives Pilot Program Report*, Portland, OR: Cambridge Systematics, 1990.

Boston Edison. *1991 Demand-Side Management Monitoring and Evaluation Plan*, Boston, MA, July, 1991.

Bradford, P. Remarks at Workshop on Demand-Side Management Program Evaluation, sponsored by New York State Energy Research and Development Authority and New York State Electric & Gas Corp., Saratoga Springs, NY, March, 1991.

Brian Gard William Lesh Inc. *Report on Market Research and Program Recommendations, Bonneville Power Administration Regionwide Promotion of Energy Efficient Appliances*, January, 1986.

Brooker, R. and R. Fichtner. "Measuring Program Impact in a Commercial and Industrial Lighting Program: The Case of Wisconsin Electric's Smart Money Program," *Proceedings of the 1991 International Energy Evaluation Conference*, Argonne, IL: Argonne National Laboratory, August, 1991, pp. 94-99.

Brown, M.A., J. O. Kolb, D. Baylon, M. H. Haeri, and D. L. White, *The Impact of Bonneville's Model Conservation Standards on the Energy Efficiency of New Home Construction*, ORNL/CON-310, Oak Ridge, TN: Oak Ridge National Laboratory, August, 1991.

Brugger, D. and R. Brooker. "Evaluation of Wisconsin Electric's Dealer Incentive Plan," *Proceedings of the 1991 International Energy Evaluation Conference*, Argonne, IL: Argonne National Laboratory, August, 1991, pp. 290-293.

Calder, B. "Cognitive Consistency and Consumer Behavior," *Perspectives in Consumer Behavior*, Harold H. Kassarian and Thomas S. Robertson (eds.), Glenview, Il: Scott, Foresman and Company, 1973.

California Public Utilities Commission and California Energy Commission, *Standard Practice Manual for Cost-Benefit Analysis of Demand-Side Management Programs*, San Francisco, CA and Sacramento, CA, December, 1987.

Cambridge Systematic and ERCE. *Commercial Incentive Pilot Program (CIPP) Process Evaluation*, Berkeley, CA, 1989.

Campbell, D. T. and L. C. Stanley. *Experimental and Quasi-Experimental Designs for Research*, Skokie, IL: Rand McNally, 1966.

Central Hudson Gas and Electric Company. *Comprehensive Process and Impact Evaluation of the Dollar Savers Program for Commercial Lighting, Air Conditioning, Motors, and Anything Goes Subprograms*, Boulder, CO, December, 1991.

Central Maine Power Company (CMP). *Evaluation of the Energy Savings Resulting from Central Maine Power Company's Good Cents Home Program*, Augusta, ME, 1990.

Charles Rivers Associates Incorporated. *New England Electric Service's Design 2000 Program: A Process Evaluation*, CRA Report No. 701.00, Cambridge, MA, 1991.

Columbia Information Systems. *Third Homebuilder Survey, Super Good Cents Evaluation, A Report on Findings From a Follow-Up Survey and a Comparison to Previous Surveys*, Portland, OR, 1988.

Consolidated Edison Company, *Process and Impact Evaluation of the Commercial and Industrial DSM Rebate Program*, New York, NY, November 1991.

Converse, J. M., and S. Presser. *Survey Questions, Handcrafting the Standardized Questionnaire*, Newbury Park, CA: Sage Publications, 1986.

Cook, T. D. and D. T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago, IL: Rand McNally, 1979.

Cummings, M. *Program Evaluation Task Force – Two Year Work Plan*, New York State Public Service Commission, Albany, NY, June, 1991.

Dillman, D. A. *Mail and Telephone Surveys, The Total Design Method*, New York, NY: John Wiley and Sons, 1978.

District of Columbia Public Service Commission. *Formal Case #834, Phase III, Order #9714*, Washington, DC, May, 1991.

Dubin, J. and D. McFadden. "An Econometric Analysis of Residential Electric Appliance Holdings and Consumption," *Econometrica*, Vol. 52, March, 1984.

Electric Power Research Institute (EPRI). *Information and Automation Technology: Serving Electric Utility Customers in the 1990s*, EPRI Report CU-6400, Palo Alto, CA: Electric Power Research Institute, June, 1989.

Electric Power Research Institute. *Impact Evaluation of Demand-Side Management Programs, Vol. 1: A Guide to Current Practice*, EPRI CU-7179, Palo Alto, CA, February, 1991a.

Electric Power Research Institute. *Impact Evaluation of Demand-Side Management Programs, Vol. 2: Case Studies and Applications*, EPRI CU-7179, Vol. 2, Palo Alto, CA, September 1991b.

Empire State Electric Energy Research Corp. *Demand Side Management Program Evaluation Scoping Study*, ESEERCO Project EP90-34, November, 1990.

Energy Program Evaluation Conference. "Energy Program Evaluation: Uses, Methods and Results," *Proceedings 1991 International Energy Program Evaluation Conference*, CONF-910807, Argonne, IL: Argonne National Laboratory, August, 1991.

Environmental and Energy Services Company and Pacific Energy Associates. *Evaluation Plan for the Orange and Rockland Utilities Demand Side Bidding Programs*, Portland, OR, 1989 (Revised 1990).

Evans, P. and Peters. J. "Integration of Process Evaluation and Program Design," *Proceedings, Energy Program Evaluation: Conservation and Resource Management*, Argonne, IL: Argonne National Laboratory, 1989, pp. 115-119.

Faruqui, A. et al. *Impact of Demand-Side Management on Future Customer Electricity Demand: An Update*, EPRI CU-6953, Palo Alto, CA: Electric Power Research Institute, September, 1990.

Feldman, S. "Forging the Link: Needs Segmentation and Geodemographics," *Proceedings, 1988 EPRI/EUMRC Market Research Symposium*, Scottsdale, AZ, November, 1988.

Fels, M. "The Princeton Scorekeeping Method: An Introduction," *Energy and Buildings*, 9, Feb./May 1986, pp. 5-18.

Fels, M. and C. Reynolds. "Toward Standardizing the Measurement of Whole-Building Energy Savings in DSM Programs," *Proceedings of the 1991 International Energy Program Evaluation Conference*, Chicago, IL, August 1991.

Franklin, J. and J. Thrasher. *An Introduction to Program Evaluation*, New York: John Wiley and Sons, 1976.

Freeman Research Resources. *A Process Evaluation of ENERGY INITIATIVE*, Vol. I, Monterey, MA, May, 1991.

Gallagher, J. T. *DSM Incentives in New York State: A Critique of Initial Utility Methods*, Albany, NY: New York Department of Public Service, March, 1991.

Gordon, F. M., M. McRae, M. Rufo, and D. Baylon. "Use of Commercial Energy Efficiency Measure Service Life Estimates in Program and Resource Planning," *Proceedings, IEEE 1988 Summer Study on Energy Efficiency in Buildings*, Washington, DC: American Council for an Energy Efficient Economy, Vol. 3, August, 1988, pp. 3.84 – 3.97.

Haeri, M. H. *Electricity Savings Three Years after Participation in the BPA Regionwide Weatherization Program*, prepared for the Bonneville Power Administration, Portland, OR, September, 1988.

Hanser, P. Electric Power Research Institute, personal communication, 1991.

Heckman, J. J. "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica* 46(6), 1978, pp. 931-959.

Heckman, J. J. and R. Robb. "Alternative Methods for Evaluating the Impact of Interventions," *Longitudinal Analysis of Labor Market Data*, J. J. Heckman and B. Singer, (eds), New York: Cambridge University Press, 1985, pp. 156-246.

Hickman, C. and T. Steele. "Site Visits to Enhance Commercial Program Evaluation," *Proceedings of 1991 International Energy Program Evaluation Conference*, Argonne, Il: Argonne National Laboratory, August, 1991, pp. 174-180.

Hicks, E. "Conservation for Profit and Program Evaluation in a Collaborative Planning Process," *Proceedings from the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Washington, DC: American Council for an Energy Efficient Economy, 1990, pp. 6.47.

Hirst, E. "Evaluating Demand-Side Management Programs," *Electric Perspectives* 14(6), Nov./Dec., 1990a, pp. 24-30.

Hirst, E. *Balancing the Scales: Data on Electricity Supply and Demand*, ORNL/CON-306, Oak Ridge, TN: Oak Ridge National Laboratory, September, 1990b.

Hirst, E. *Possible Effects of Utility DSM Programs, 1990 to 2010*, ORNL/CON-312, Oak Ridge, TN: Oak Ridge National Laboratory, January, 1991.

Hirst, E. and K. M. Keating. "Dynamics of Energy Savings Due to Conservation Programs," *Energy Systems and Policy* 10(3), 1987, pp. 257-273.

Hirst, E. and C. Sabo. *Electric-Utility DSM-Program Terminology*, ORNL/CON-337, Oak Ridge, TN: Oak Ridge National Laboratory, October, 1991.

Hirst, E., D. L. White, and R. Goeltz. *Three Years after Participation: Electricity Savings Due to the BPA Residential Weatherization Pilot Program*, ORNL/CON-166, Oak Ridge, TN: Oak Ridge National Laboratory, January, 1985.

Horowitz, M. J., L. K. Ecker, and P. Degens. *Long Term Impacts of the Interim Residential Weatherization Program on Household Energy Savings*, (ERCE/DSM-65), prepared for the Bonneville Power Administration, Portland, OR, May, 1991.

Joint Committee on Standards for Educational Evaluation. *Standards for Evaluation of Educational Programs, Projects, and Methods*, New York: McGraw-Hill, 1981.

Kaplan, M., B. Jones, and J. Jansen. "DOE2.1C Model Calibration with Monitored End-Use Data," *Proceedings of the ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Washington, DC: American Council for an Energy Efficient Economy, August, 1990, pp. 10.115 – 10.125.

Keating, K. M. "Self-Selection Bias: Are We Beating a Dead Horse?" *Evaluation and Program Planning*, 12(2), 1989, pp. 137-142.

Keating, K. "Issue Paper on Free Riders," *1990 Conservation and Load-Management Evaluation Plan*, New England Electric System, June, 1990.

Keating, K. Personal communication, Bonneville Power Administration, Portland, OR, September, 1991.

Keating, K. and S. Blachman. "In Search of an Impact: An Evaluation of an Institutional Buildings Program," *Proceedings, Energy Conservation Program Evaluation, Practical Methods, Useful Results, 1987 Conference*, Vol. 1, Argonne, IL: Argonne National Laboratory, August, 1987, pp. 107-116.

Keating, K. and E. Hicks, (eds). "Program Evaluation," *Proceedings, ACEEE 1990 Summer Study on Energy Efficiency in Buildings*, Vol. 6, Washington, DC: American Council for an Energy Efficient Economy, August, 1990.

Keating, K. M. and E. Hirst. "Advantages and Limits of Longitudinal Evaluation Research in Energy Conservation," *Proceedings, Evaluation and Program Planning*, Vol. 9, Washington, DC: American Council for an Energy Efficient Economy, August, 1986, pp. 113-120.

Keating, K. M. and T. V. Oliver. *Evaluating Commercial Conservation Programs: A Difficult Business*, Bonneville Power Administration, Portland, OR (draft), 1991.

King, M. and W. Gavelis. "Informed Conditional Demand Analysis: An Integrating Methodology for Estimating Energy Impacts," *Proceedings, American Council for an Energy Efficient Economy, Washington, DC*, August, 1990, pp. 10.127 - 10.134.

Komor, P., W. Kempton, and J. Haberl. *Energy Use, Information, and Behavior in Small Commercial Buildings*, Princeton, NJ: Center for Energy and Environmental Studies, Princeton University, July, 1989.

Krause, F. and J. Eto. *Least-Cost Utility Planning Handbook for Public Utility Commissioners, The Demand Side: Conceptual and Methodological Issues*, Vol. 2, Washington, DC: National Association of Regulatory Utility Commissioners, December, 1988.

Kreitler, V. "Market Issues in Free Rider Estimation," *Proceedings of the Electric Utility Marketing Research Council EUMRC/EPRI Conference*, EPRI CU-7010, Bala Cynwyd, PA: Synergic Resources Corporation, November, 1990, pp. 10-28, 33, 34.

Kreitler, V. "On Customer Choice and Free Ridership in Utility Programs," *Proceedings of the 1991 International Energy Evaluation Conference*, Argonne, IL: Argonne National Laboratory, August, 1991, pp. 303.

Kunkle, R. *Case Study Analysis of the Institutional Buildings Program in Washington State*, Olympia, WA: Washington State Energy Office, October, 1990.

Kunkle, R. and C. Johnson. "Building Changes and Conservation Measure Problems: Institutional Buildings Program Case Studies," prepared for the Bonneville Power Administration, Portland, OR, May, 1991.

Kushler, M. G. "Use of Evaluation to Improve Energy Conservation Programs: A Review and Case Study," *Journal of Social Issues* 45(1), 1989, pp. 153-168.

Lerman, D. *Regional Study of Residential Water Heating Equipment*, IEAL/PO-24, Portland, OR: Bonneville Power Administration, 1987.

Lerman, D. and B. Bronfman. *Process Evaluation of the Bonneville Power Administration Long Term Residential Conservation Program*, ERC/PO-5, Portland, OR: Bonneville Power Administration, 1984.

Lerman, D. and B. Bronfman. *Process Evaluation of the BPA Commercial Audit Program*, IEAL/PO-4, Portland, OR: Bonneville Power Administration, 1986a.

Lerman, D. and B. Bronfman. *Process Evaluation of the Tacoma MCS Adoption: Part II*, IEAL/PO-9, Portland, OR: Bonneville Power Administration, 1986b.

Lerman, D., B. Bronfman, and B. Tonn. *Process Evaluation of the BPA Residential Weatherization Pilot Program*, ORNL/CON-138, Oak Ridge, TN: Oak Ridge National Laboratory, 1983.

Long Island Lighting Company. *Electric Class of Customer Study*, Hicksville, NY, 1972.

Long Island Lighting Company and Applied Energy Group, Inc. *Impact Evaluation of Long Island Lighting Company's 1990 Electric Conservation and Load-Management Plan*, Hicksville, NY, 1991.

Lui, D. and J. Fang. *Free Rider Ratios in Conservation programs: Estimates and Issues*, staff topic paper prepared for the Illinois Commerce Commission, Springfield, IL, 1989.

Lui, D. and J. Fang. "Issues in Free Rider Estimations," *Proceedings of the 1990 ACEEE Summer Study on Energy Efficiency in Buildings*, Vol. 6, Washington, DC: American Council for an Energy Efficient Economy, August, 1990, pp. 6.96.

MacDonald, S. L. *Residential Peak Load-Reduction Project, Shed LMG Test Analysis for 1986 Operating Season*, San Francisco, CA: Pacific Gas & Electric Company, July, 1987.

Malcolm, W. H. "Ten Years After: Lessons Learned from the Residential Direct Control Program," *Proceedings: Demand-Side Management Strategies for the 90s, Fourth National Conference on Utility DSM Programs*, EPRI CU-6367, Palo Alto, CA: Electric Power Research Institute, April, 1989, pp. 30-1-8.

Massachusetts Electric. *1990 DSM Performance Measurement Report*, Westborough, MA: New England Electric System, June, 1991.

McRae, M. "Don't Shoot the Messenger: Conducting DSM Process Evaluations in (Inevitably) Political Environments," *Proceedings, ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Program Evaluation*, Vol. 6, Washington, DC: American Council for an Energy Efficient Economy, August, 1990, pp. 6.99-6.107.

Meier, A. *Supply Curves of Conserved Energy*, Ph.D. dissertation, Berkeley, CA: Energy and Resources Group, University of California, May, 1982.

Miles, M. and A. Huberman. *Qualitative Data Analysis: A Sourcebook of New Methods*, Newbury Park, CA: Sage Publications, 1984.

Nadel, S. *Lessons Learned: A Review of Utility Experience with Conservation and Load-Management Programs for Commercial and Industrial Customers*, Albany, NY: New York State Energy Research and Development Authority, April, 1990.

Nadel, S. "Use of Simple Performance Indices to Help Guide Review of DSM Program Performance," *Proceedings, National Conference on Integrated Resource Planning*, Washington, DC: National Association of Regulatory Utility Commissioners, April, 1991, pp. 116-130.

Nadel, S., K. Cress, and M. Ticknor. "The NEES Stand-By Generation Program: Design, Results, and Lessons Learned," *Demand-Side Management Strategies for the 90s, Proceedings, Fourth National Conference on Utility DSM Programs*, CU-6367, Vol. 2, Palo Alto, CA: Electric Power Research Institute, April, 1989, pp. 65-1-14.

Nadel, S. M. and K. M. Keating. "Engineering Estimates vs. Impact Evaluation Results: How Do They Compare and Why?" *Proceedings of 1991 International Energy Program Evaluation Conference, CONF-910807*, Argonne, Il: Argonne National Laboratory, August, 1991, pp. 24-33.

Nadel, S. and H. Tress. *The Achievable Conservation Potential in New York State from Utility Demand-Side Management Programs*, Report 90-18, Albany, NY: New York State Energy Research and Development Authority, November, 1990.

New England Electric System (NEES). *Six Month Evaluation of New England Electric System Partners in Energy Planning Programs*, Westborough, MA: New England Electric System, April, 1988.

New England Power Service Company, *1991 DSM Performance Measurement Report*, Westboro, Ma, June, 1991.

New Jersey Conservation Assessment Team. *Final Contractor's Report to the New Jersey Conservation Analysis Team*," Boulder, CO: RCG/Hagler, Bailly, Inc., August, 1990.

New York Power Pool and the Empire State Electric Energy Research Corporation, *Report of Member Electric Systems pursuant to Article VIII, Section 149-b of the Public Service Law*, 1976, Vol. 1, pp. 96-171.

New York Public Service Commission. *Order Concerning 1991 and 1992 Demand Side Management Plans*, Case 28223, Albany, NY, November, 1990.

New York State Energy Office. *Program Evaluation of the New York State Energy Efficient Appliance Rebate Demonstration Program*, Albany, NY, May, 1989.

Nichols, D., et al. *Savings from the Smart Money Program, An Audit of the 125 Megawatt Demand Reduction*, Boston, MA: The Tellus Institute, May, 1990.

Northeast Region Demand-Side Management Data Exchange. *Program Data Summary Report*, Washington, DC: Edison Electric Institute, June, 1989.

Northeast Utilities. *ESLR Short-Duration Monitoring Interim Report*, Cambridge, MA: RLW Analytics, Inc., 1991a.

Northeast Utilities. *Integrating Information for DSM Evaluation Using the Engineering Calibration Approach*, Cambridge, MA: RLW Analytics, Inc., 1991b.

Okumo, D. "Multifamily Retrofit Electricity Savings: the Seattle City Light Experience," *Proceedings of the ACEEE 1990 Summer Study on Efficiency in Buildings*, Washington, DC: American Council for an Energy Efficient Economy, 1990, p. 6.119.

Ozog, M. *Free Ridership in WP&L Buysmart Residential DSM Programs*, Draft report prepared for Wisconsin Power and Light, July, 1991.

Pacific Consulting Group, Inc. *Evaluation of PG&E's 1984-85 Customized Rebate Program Final Report*, Vol. I, June, 1986.

Parti, C. and M. Parti. "The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector," *Bell Journal of Economics*, Spring, 1980, pp. 309-321.

Patton, M. Q. *Practical Evaluation*, Beverly Hills, CA: Sage Publications, 1982a.

Patton, M. Q. *Creative Evaluation*, Newbury Park, CA: Sage Publications, 1982b.

Patton, M. Q. *How to Use Qualitative Methods in Evaluation*. Newbury Park, CA: Sage Publications, 1987, (Fifth printing 1990).

Peters, J., D. Barry, M. Horowitz, and F. Gordon. *Evaluation of the Orange and Rockland Utilities Competitive Bidding Program for Demand Side Resources*, Interim Report, ERCE/DSM-66, Portland, OR: Bonneville Power Administration, 1990.

Peters, J. and B. Bronfman. *Street and Area Lighting Program Evaluation Update*, IEAL/PO-6, Portland, OR: Bonneville Power Administration, 1986.

Peters, J. and G. Gustafson. *Process Evaluation of the Sponsor-Designed Site Specific Program*, Portland, OR: Bonneville Power Administration, March, 1987.

Peters, J., G. Gustafson and M. Vowles. *Process Evaluation of the BPA Purchase of Energy Savings Pilot Program*, IEAL/PO-20, Portland, OR: Bonneville Power Administration, 1987.

Peters, J., H. Haeri, and G. Gustafson. *Comparative Analysis of Commercial Sector Financing Mechanisms: CIPP, PES Pilot and Puget Incentive Programs*, ERC/PO-35, Portland, OR: Bonneville Power Administration, 1988.

Peters, J., H. Haeri, and P. Seratt. *Evaluation of the PG&E Home Energy Survey Programs*, ERCE/PO-54, San Francisco, CA: Pacific Gas & Electric Corp., 1990.

Peters, J., K. Oswald, and M. Horowitz. *Evaluation of the Northeast Utilities EnergyCHECK Program*, Hartford, CT: Northeast Utilities, December, 1990.

Petersen, F. J. "Remodel and Tenancy Changes: Threats to the Reliability of Commercial Conservation Savings," *Proceedings, 1990 ACEEE Summer Study on Energy Efficiency in Buildings*, Vol. 3, Washington, DC: American Council for an Energy Efficient Economy, August, 1990, pp. 3.165 - 3.172.

Petersen, F. J. and S. Sandler. "Institutional Buildings in Flux," *Proceedings of 1991 International Energy Program Evaluation Conference*, Argonne, Il: Argonne National Laboratory, August, 1991, pp. 41-48.

Pollard, W. E. "Bayesian Statistics for Evaluation Research: An Introduction," *Contemporary Evaluation Research Series No. 8*, Newbury Park, CA: Sage Publications, 1986.

Prahl, R. "Evaluation for PUC's," *Proceedings of the 1988 Summer Study on Energy Efficiency in Buildings*, Washington, DC: American Council for an Energy Efficient Economy, 1988, p. 9.126.

Puget Energy Services, Inc. Final Report for the *Commercial Retrofit End-Use Study*, prepared for the Bonneville Power Administration, Portland, OR, March, 1991.

Regional Economic Research, Inc. *SDG&E Non-Residential Audit Evaluation*, October, 1990.

Regional Economic Research, Inc. *An Evaluation of Niagara Mohawk's Low Cost Measures Program*, June, 1991.

Rochester Gas & Electric and Applied Energy Group, Inc. *Evaluation of Rochester Gas & Electric's 1990 Demand Side Management Programs*, 1991.

Rogers, E. "Evaluation of a Residential Appliance Rebate Program Using Billing Record Analyses," *Proceedings, 1989 Energy Program Evaluation Conference*, Argonne, IL: Argonne National Laboratory, 1989.

Rossi, P. and F. Freeman. *Evaluation – A Systematic Approach*, Beverly Hills, CA: Sage Publications, 1982.

San Diego Gas & Electric Company. *SDG&E Non-Residential Audit Evaluation*, San Diego, CA: Regional Economic Research, Inc., 1991.

Saxonis, W. "Program Evaluation of the New York State Energy Efficient Appliance Rebate Demonstration Program: The Free Rider Factor," *Proceedings of the 1989 International Energy Program Evaluation Conference: Conservation and Resource Management*, Argonne, IL: Argonne National Laboratory, August, 1989, pp. 237-240.

Saxonis, W. "Measuring Free Riders: Does the Economic Climate Make a Difference?" *Proceedings of the 1991 International Energy Evaluation Conference*, Argonne, IL: Argonne National Laboratory, August, 1991, pp. 286.

Schmitt, S.A. *Measuring Uncertainty: An Elementary Introduction to Bayesian Statistics*, Reading, MA: Addison-Wesley, 1969.

Schoch-McDaniel, K. *Third Year Analysis of Electricity Use and Savings for the Hood River Conservation Project*, prepared for the Bonneville Power Administration, Portland, OR, October, 1990.

Schön, A. and K. Hamilton. "Commercial Sector Alternatives to End-Use Metering," *Proceedings of End-Use Load Information and Its Role in DSM Conference*, Irvine, CA, 1990.

Skumatz, L. A., K. M. Lorberau, R. J. Moe, R. J. Bordner, and R. D. Chandler. *Bonneville Measure Life Study: Effect of Commercial Building Changes on Energy Using Equipment*, prepared for the Bonneville Power Administration, Portland, OR, 1991.

Stout, T. and W. Gilmore. "Motor Incentive Programs: Promoting Premium Efficiency Motors," *Demand-Side Management: Partnerships in Planning, Proceedings ECNE National Conference on Utility DSM Programs*, EPRI CU-6598, Bedford, MA: Electric Council of New England, November, 1989, pp. 27-1-11.

Stucky, L., M. Khawaja, S. French, C. Collette, and H. G. Peach. *Impact Evaluation of the Low-Income Segment of the Weather Shield and Attic Attack Program*, Augusta, ME: Central Maine Power, 1990.

Sumi, D. H. and B. Coates. *Longitudinal Evaluation of Energy Savings from Seattle City Light's Home Energy Loan Program*, Seattle WA: Seattle City Light, December, 1988.

Synergic Resources Corporation. *Analysis of the NYSEG Power Pincher High Efficiency Refrigerator Program*, Bala Cynwyd, PA: Synergic Resources Corporation, November, 1987.

Synergic Resources Corporation. *Process Evaluation of the Bonneville Long Term Weatherization Program: 1987-1989*, SRC Report 7582-R3, Bala Cynwyd, PA: Synergic Resources Corporation, 1991.

TechPlan Associates. *Evaluation of the Easy Savers Program*, Poughkeepsie, NY: Central Hudson Gas and Electric Corporation, March, 1990.

Tolkin, E. *Energy Fitness Resurvey Analysis*, memo report to New England Electric System, Westborough, MA, October, 1990.

Townsley, M. and R. Wright. "Measuring DSM Impacts: End-Use Metering and the Engineering Calibration Approach," *Proceedings of 2nd Conference on End-Use Load Information and Its Role in DSM Association of DSM Professionals*, Irvine, CA, 1990.

Train, K., P. Ignelzi, and Kumm. "Evaluation of a Conservation Program for Commercial and Industrial Customers," *Energy*, 10(10), 1985, pp. 1079-88.

Train, K. E. *Correcting Self-Selection Bias in the Estimation of Audit Program Impacts*, Berkeley, CA: Cambridge Systematics, 1988.

Train, K. and P. Ignelzi. "The Economic Value of Energy-Saving Investments by Commercial and Industrial Firms," *Energy*, 12(7), 1987.

Van Liere, K. and T. McKinnell. "Comparison of Alternative Segmentation Approaches for Residential Customers," *Proceedings Energy Program Evaluation: Conservation and Resource Management*, Argonne, IL: Argonne National Laboratory, 1989, pp. 457-464.

Vine, E., O. De Buen, and C. Goldman. *Stimulating Utilities to Promote Energy Efficiency: Process Evaluation of Madison Gas and Electric's Competition Pilot Program*, Report LBL-29642, Berkeley, CA: Lawrence Berkeley Laboratory, December, 1990.

Violette, D. "Use of Metered Data in DSM Evaluation," *Proceedings of End-Use Load Information and Its Role in DSM*, sponsored by Aloha Systems, Inc., Irvine, CA: The Fleming Group and the Association of DSM Professionals, June 1990.

Violette, D. and M. Ozog. "Correction for Self-Selection Bias: Theory and Application," *Energy Program Evaluation: Conservation and Resource Management Conference*, Argonne, IL: Argonne National Laboratory, August, 1989, pp. 241-250.

Violette, D., M. Ozog, and G. Wear. "Measuring Free Ridership: Do Some Experimental Designs Control Twice for Free Ridership?" *Proceedings of the 1991 International Energy Evaluation Conference*, Argonne, IL: Argonne National Laboratory, August, 1991, pp. 296.

Violette, D., M. Ozog, M. Keneipp, and F. Stern. *Impact Evaluation of Demand-Side Management Programs, Vol. 1: A Guide to Current Practice*, CU-7179, Palo Alto, CA: Electric Power Research Institute, February, 1991.

Walpole, R. E. and R. H. Myers. *Probability and Statistics for Engineers and Scientists*, 3rd ed, McMillan Publications, 1985.

Weiss, C. H. *Evaluation Research: Methods of Assessing Program Effectiveness*, Englewood Cliffs, NJ: Prentice Hall, 1972. P. 11.

White, D. L. and M. A. Brown. *Electricity Savings Among Participants Three Years after Weatherization in Bonneville's 1986 Residential Weatherization Program*, ORNL/CON-305, Oak Ridge, TN: Oak Ridge National Laboratory, September, 1990.

White, D. L., T. K. Stovall, and B. E. Tonn. *Persistence of the Impact of the Hood River Conservation Project on Typical and Peak Loads: Three Years after Weatherization*, ORNL/CON-321, Oak Ridge, TN: Oak Ridge National Laboratory, June, 1991.

Wiggins, D. San Diego Gas and Electric, personal communication, February, 1991.

Wisconsin Electric. *1990 Annual Evaluation Report for Wisconsin Electric's Demand-Side Management Programs*, March, 1991.

Wisconsin Public Service Corporation. *Good Cents Home Program: Impacts and Evaluation*, Boulder, CO: RCG/Hagler, Bailly, Inc., March, 1989.

XENERGY, Inc. *Demand-side Management Program Evaluation Scoping Study*, ESEERCO Project EP90-34, New York, NY: Empire State Energy Research Corporation, November, 1990.

XENERGY, Inc. *Joint Utility Baseline New Construction Practices Study*, Burlington, MA. 1991.

ANNOTATED BIBLIOGRAPHY

John Reed

INTRODUCTION

This bibliography directs readers to works that are seminal to evaluation and energy evaluation. Many of the items are infrequently cited in the energy literature but provide important foundations on which the field of energy evaluation is built. No attempt has been made to make an exhaustive list. For example, many of the high-quality studies from the ACEEE Summer Studies series, the Chicago Evaluation Conference Proceedings, and EPRI conference proceedings are underrepresented. We urge the reader to explore those sources for examples of high-quality evaluations. The brief annotations are designed to help readers locate those books and articles that will meet their needs. In addition to being chosen for their importance, these works were also chosen for accessibility. Most of these items should be available in university libraries or through interlibrary loan. The list is presented alphabetically according to the name of the first author under three headings: General Evaluation Methods, Analytical Methods, and Energy Evaluation.

GENERAL EVALUATION METHODS

Cook, T. D., and D. T. Campbell. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Chicago, IL: Rand McNally, 1979.

Energy-program evaluators are seldom able to conduct evaluations with classical experimental design (i.e., measurements taken before and after program implementation for a control group and an experimental group with random assignment of subjects to the two groups). Most frequently, energy researchers are faced with situations in which they are asked to perform an evaluation after a program has been implemented and for which no control group is available. This volume, which is the successor to an earlier volume by Stanley and Campbell, *Experimental and Quasi-Experimental Designs*, discusses just such problems. The book begins with an excellent discussion of validity and then moves on to discussions of quasi-experimental designs, such as the one-group post-program design, the post-program only with nonequivalent groups design, and the one-group pretest-posttest design. Other designs are also discussed. The book portrays a variety of statistical techniques that may be used to reduce uncertainty about causality given the use of quasi-experimental designs. Beginners in evaluation and even the mature researcher can benefit from reviewing this book from time to time.

Dillman, D. A. *Mail and Telephone Surveys, The Total Design Method*, New York, NY: John Wiley and Sons, 1978.

The development, administration, and analysis of surveys is often poorly done because evaluators underestimate the difficulties and complexities of doing a quality job. Want to know how to word a question? Want to reduce the complexity of a question? Want to know how to improve the response rate for a mailed survey? Want to know how to reduce the cost of processing data once a questionnaire is returned? This volume provides answers to these and other questions that may confront an evaluator. Highly recommended even for the seasoned veteran of survey research, who can benefit from perusing this volume to see if all the bases have been covered before that questionnaire is mailed or the telephone survey begun.

Patton, M. Q. *Creative Evaluation*, Newbury Park, CA: Sage Publications, 1982.

This is not your typical cookbook providing a recipe for how to do evaluation. Rather, Patton says that the book contains "ideas, techniques, and approaches" that he "never taught ... in formal courses" and that still are probably not a part of most courses in evaluation. What Patton does is provide methods that he uses to creatively work with decision makers and users of evaluation information. The book describes activities like flow charting, picture thinking, story thinking, and simulation games. If you have trouble communicating about evaluation with management and other decision makers, this volume will help. Evaluation reports might be less dry and better utilized if some of these techniques were used.

Patton, M. Q. *How to Use Qualitative Methods in Evaluation*, Newbury Park, CA: Sage Publications, 1987.

In-depth interviews, direct observation, and the analysis of written documents, such as speeches, diaries, program records, and open-ended written items on questionnaires, are the main types of qualitative methods. Decision makers often view the type of information produced by qualitative methods with suspicion. However, if qualitative methods are applied with knowledge and skill, the result can be reliable and valid data that provide insight not available from quantitative methods. As the author puts it, "Systematic and rigorous observation involves far more than just being present and looking around. Skillful interviewing involves much more than just asking questions. Content analysis requires considerably more than just reading to see what's there." This book describes how and when to use qualitative methods and how to conduct observation and in-depth interviews and how to analyze written documents as well as the data produced by observation and interviews. The techniques described here are particularly useful for scoping a problem and for process evaluation. They also may help in interpreting the anomalies that occur in quantitative data.

Patton, M. Q. *Practical Evaluation*, Beverly Hills, CA: Sage Publications, 1982.

This is another volume that is full of practical guidance. Patton's motivation for writing this book was to call attention to the need for excellence in evaluation. The book contains the "thoughtful" chapters, "Thoughtful Questionnaires," "Thoughtful Interviewing," and "Thoughtful Methods Decisions." It also contains a chapter on "Managing Management Information Systems," which discusses the problem of managing data, which is often overlooked, is little discussed, and can lead to severe problems. While management-information science has changed since this volume was written, this chapter is a helpful reminder, and the information is still of value.

The book is very readable. Instead of being prescriptive, the author provides examples that lead the reader to draw his own conclusions. For example, at one point, the author gives examples of three styles of tabular presentation. Reading the text is almost unnecessary to grasp the point the author is making about presenting tabular data. This book contains important reading, but unlike many books on evaluation, this is one that can be read with enjoyment and understanding.

Shadish, W. R., Jr., T. D. Cook, and L. C. Leviton. *Foundations of Program Evaluation: Theories of Practice*, Newbury Park, CA: Sage Publications, 1991.

This new book discusses the work of some of the most respected names in evaluation, Campbell, Weiss, Wholey, Cronbach, and Rossi to name a few. The book has several purposes. First, it provides an intellectual history of evaluation so that the reader can understand how and why evaluation practices have evolved. Second, the book attempts to broaden the perspective of evaluation practitioners by highlighting how the historical circumstances in various sectors (education and mental health, for example), have shaped the form that evaluations in those sectors have taken. This is useful because it helps to enlarge our view of what evaluation is and the options that are available to practitioners. It also helps to sensitize the practitioner to the ways in which the context of a problem may shape the evaluation. Third, the book is targeted to policy-makers, who often make decisions about evaluation without understanding evaluation and without consulting evaluation specialists.

Energy evaluation as a specialty has been somewhat insular from the main stream of evaluation as a discipline. At times, it seems as if energy-evaluation practitioners have been quick to reinvent the wheel and slow to recognize and to solve problems that have already been dealt with in other sectors. The exposure of energy evaluators to this book will partially remedy this problem.

Weiss, C. *Evaluation Research: Methods of Assessing Program Effectiveness*, Englewood Cliffs, NJ: Prentice Hall, 1972.

The focus of this slim volume is the application of evaluation research methods to real-world environments. The author offers guidance on adapting "textbook methods" to the reality of the world in which evaluation must be done. One of the chapters, "Formulating the Question and Measuring the Answer," focuses on the problem of developing the goals of a program and how these relate to measuring program outcomes. Another chapter discusses the changing nature of programs and the relations between the evaluator(s) and program personnel and the impact of these on evaluation. Also, a chapter treats the use of evaluation results. In contrast to the Cook and Campbell volume with its theoretical focus, this volume deals with the practical issues of evaluation. Those new to the field might find this a good introduction and useful first reading before tackling Cook and Campbell.

Wholey, J. S. *Evaluation and Effective Public Management*, Boston, MA: Little Brown and Company, 1983.

The premise that underlies this book is that more managerial and analytic attention needs to be given to program assessment and improvement. Although this book is oriented to government programs and the civil service, the content is of value to those conducting energy evaluations for utilities as well as public service commissions. The book is divided into four parts. Two of these parts are of particular interest. One is the section on getting agreement on results-oriented objectives. Many programs and many evaluations fail because of lack of agreement among participants about objectives. This section includes a chapter discussing the problem of setting realistic and measurable objectives. The section on "Assessing Program Performance and Results" discusses the problems of developing performance measures and evaluation designs. Also in this section is a discussion of outcome monitoring. The book contains numerous examples of results-oriented management. This book is of value to managers wanting to know how to effectively use evaluations to better manage their programs.

STATISTICAL AND ANALYTICAL METHODS

Aigner, D. J., C. Sorooshian, and P. Kerwin. "Conditional Demand Analysis for Estimating Residential End-Use Load Profiles," *The Energy Journal*, 5(3), 1984, pp. 81-97.

One of the real frustrations for energy evaluators is obtaining locale-specific end-use load shapes. Although costs are dropping, end-use load data are still costly to gather and analyze. This article describes the use of conditional-demand analysis (see also the article by Parti and Parti) to statistically extract end-use load profiles from whole-household data. Twenty-four regression equations were fitted with the hourly consumption of electricity averaged over the days of the month regressed on the presence or absence of household appliances. The result was a series of well-defined residential load shapes, although the "load level often seemed questionable." The authors attribute this to the simple model specification and to

the fact that ownership and intensity of use were not incorporated into the model. Nonetheless, the technique is well illustrated. With ownership and intensity of use incorporated into the model, the results should be excellent. Conditional-demand analysis offers a lower-cost alternative to end-use metering, especially if a utility already has remote-meter-reading capability in place.

Fels, M. F. (ed). "Measuring Energy Savings: The Scorekeeping Approach," a special edition of *Energy and Buildings*, 9(1), 1986.

One cannot be in the business of energy evaluation for long without having to compare energy-consumption measurements taken at different times (before and after a program) or places (that are influenced by different weather conditions). Examples of such situations include comparing energy consumption for the same household in two different years and comparing the consumption of different households with different monthly billing cycles. Fortunately, adjusting the data to remove the effects of weather is straightforward. The required data include whole-building meter readings and daily average temperature readings. This volume, which is an excellent treatment of weather normalization, presents a series of papers that use and discuss PRISM. There is an excellent introduction to the subject by the editor. Many of the remaining 15 papers present examples of the use of PRISM for different building types (single-family and multifamily dwellings) and fuel types (electricity, wood, and oil). Additional papers comment on the reliability of PRISM and other uses of this model. If you evaluate programs with aggregate building data, this is an important collection of papers.

Loether, H. J., and D. G. McTavish. *Descriptive and Inferential Statistics*, Boston, MA: Allyn and Bacon, Inc., 1976.

This is one of the best of the general surveys of statistics and one of the easiest to use. The book does a good job of presenting the techniques so that they are easily understood. Although it lacks discussion of some of the most recent advances, it covers most of the descriptive and inferential statistics that an evaluator might use. It has a particularly strong section on the analysis of cross-classification data, which is useful if you are analyzing questionnaire or household data. It also has a section on graphic presentation and analysis that may be helpful if you use a microcomputer graphics software package to analyze or display data. The book also does a good job of helping the user to understand when to use the various statistical techniques. The book contains numerous sidebars that summarize important material. The book is highly recommended if you are new to the use of statistics or if you need a quick refresher.

Mosteller, F., and J. W. Tukey. *Data Analysis and Regression: A Second Course in Statistics*, Reading, MA: Addison-Wesley, 1977.

The key to this book is its subtitle, "A Second Course in Statistics." This is a good follow-on to the Tukey book on Exploratory Data Analysis and the Loether and McTavish volumes described elsewhere in this bibliography. This is a book about regression analysis and not a general statistics book. The book discusses what regression is and what regression coefficients can and cannot tell us. It discusses model fitting and what information can be extracted from residuals. Most of us were introduced to statistics based on the normal distribution. However, the energy researcher might find the beta distribution more useful, especially for modelling the distribution of duty cycle of appliances, such as air conditioners. The first chapter in this book discusses the strengths and weaknesses of the Gaussian (normal) and beta distributions. If you need to use regression or to evaluate studies that used regression, this text should be very useful.

Parti, M., and C. Parti. "The Total and Appliance-Specific Conditional Demand for Electricity in the Household Sector," *Bell Journal of Economics*, 2(2), Spring, 1980, pp. 309-321.

Faced with the need to determine appliance-specific energy consumption, the evaluator can use engineering estimates, install end-use monitoring devices, or use conditional-demand analysis. With appropriate models, engineering estimates are straightforward although often not as accurate as one would like, and the estimates tend not reflect to regional differences in appliances and usage patterns. End-use monitoring provides excellent data but is expensive. An alternative is conditional-demand analysis, especially if your utility already has the data required for the analysis. Using information about household-specific appliance stocks, one can disaggregate total household load into estimates of energy consumption for each appliance. The term "conditional" comes from the notion that you either have an appliance or do not. Regression techniques, in which the availability of appliances are represented as dummy variables, are used to estimate the parameters of the component demand functions. Add demographic variables, information about the size of the building and the season, follow the recipe in this article, and you, too, can estimate energy usage by appliance. For a slightly different use of this approach, see the Aigner article above.

Tukey, J. W. *Exploratory Data Analysis*, Reading, MA: Addison-Wesley, 1977.

This is the classic work on Exploratory Data Analysis (EDA). EDA is a series of techniques that allows one to look effectively at data. The purpose of EDA is to gain insight into data when the patterns in the data or the meaning of data may be unclear. This volume describes techniques that allow the user to look at data in new ways. Most of the techniques (for example, leaf-and-stem analysis) can be applied with pencil and paper, although statistical packages for microcomputers will now present the same data.

Violette, D., M. Ozog, M. Keneipp, and F. Stern. *Impact Evaluation of Demand-Side Management Programs: A Guide to Current Practice*, Vol. 1, CU-7179, Palo Alto, CA: EPRI, 1991.

As its title implies, this volume focuses on impact evaluation. While it discusses many general evaluation issues, it is dedicated to presenting the state of the art in impact analysis. Important chapters on data collection, engineering methods, statistical methods, and combining data from multiple sources are included. Each chapter discusses specific techniques in detail as well as providing examples. The strengths and weaknesses of the various techniques are also discussed. If you need detailed information on specific impact-analysis techniques, this volume is the place to start. (The reader should be forewarned that this is an EPRI document and that its use is governed by a license agreement. This may limit its usefulness to organizations that are affiliated with EPRI.) Most of the material in this volume is available in other formats, but access is not nearly so convenient.

ENERGY EVALUATIONS

ACEEE 1990 Summer Study on Energy Efficiency in Buildings, Washington, DC: American Council for an Energy Efficient Economy, 1990.

These are the proceedings for the ACEEE 1990 Summer Study. The general foci of these proceedings are buildings and energy technologies, but many other issues (such as organizational behavior and decision making, evaluations of incentive programs, social stratification and appliance stratification, and efficiency and greenhouse gases) are discussed. In particular, one volume of the proceedings deals exclusively with the evaluation of DSM programs. This conference alternates years with the Chicago Evaluation Conference. If you want to know the state of the art in energy-efficient technologies, this is the place to start. Proceedings for 1988 and 1986 are also available.

Hirst, E. "The Hood River Conservation Project: An Evaluators Dream," *Evaluation Review*, 12(3), June, 1988, pp. 310-325.

The Hood River Conservation Project was a 5-year, \$20-million research and demonstration project to install cost-effective conservation measures in electrically heated homes in Hood River, Oregon. The measures were aimed at the building shell to reduce electricity use for SH and at WH retrofits. This article summarizes the remarkable and largely successful efforts to design a comprehensive evaluation to address energy-policy issues important to the Pacific Northwest. Evaluation results are presented concerning the number of eligible households that participated, the number of recommended conservation actions that were adopted, and the actual electricity use and savings that were achieved.

Energy Program Evaluation: Uses, Methods, and Results: Proceedings of the 1991 International Energy Program Evaluation Conference, CONF-910807, Chicago, IL: National Energy Program Evaluation Conference, August, 1991.

The International Energy Program Evaluation Conference is a biannual conference that alternates with the ACEEE Summer Study. The papers in this proceedings include evaluations of energy programs, discussions of evaluation methods, and discussions of the practical problems of evaluating energy programs. This proceedings and its predecessor volumes present the state of the art in energy program evaluation.

Nadel, S. M., and K. M. Keating. "Engineering Estimates vs. Impact Evaluation Results: How Do They Compare and Why?" in *Energy Program Evaluation: Uses, Methods, and Results: Proceedings of the 1991 International Energy Program Evaluation Conference*, CONF-910807. Chicago, IL: National Energy Program Evaluation Conference, August, 1991, pp. 24-33.

How good are engineering estimates compared to the results of impact evaluations based on billing or metered data? This paper examines the results of 42 studies in an attempt to answer this question. Impact-evaluation results are lower than engineering estimates for residential retrofit programs, commercial and residential lighting programs, and low-flow showerhead programs; and they are about equivalent for residential appliance programs, new construction programs, and multiple-measure C&I programs. The reasons for the discrepancies are erroneous assumptions in the engineering calculations, complex interactions among measures, quality-control problems with the measures, and greater-than-expected adoption of measures by nonparticipants. This work is an example of a metastudy that can help guide better program designs and evaluations.

Nadel, S. M., and M. Ticknor. "Electricity Savings from a Small Commercial & Industrial Lighting Retrofit Program: Approaches and Results," in *Energy Program Evaluation: Conservation and Resource Management: Proceedings of the 1989 International Evaluation Conference*, Chicago, IL: National Energy Program Evaluation Conference, 1989, pp. 107-111.

This short piece describes an impact evaluation of a pilot lighting retrofit program for small C&I customers offered by the Massachusetts Electric Company during 1985 to 1987. This study is unusual because four methods were employed: engineering calculations, comparison of pre- and postprogram daily kWh use with a control group, comparison of pre- and postprogram daily kWh use with a survey on changes in participants' energy-use patterns, and conditional-demand analysis. The first and third methods were of questionable accuracy, while the second worked well, and the fourth performed adequately but not without some difficulties. This paper is a good comparison of the problems and promises of various methods.

Sumi, D. H., and B. Coates. "Persistence of Energy Savings in Seattle City Light's Residential Weatherization Program," *Proceedings of the 1989 International Energy Program Evaluation Conference: Conservation and Resource Management*, Chicago, IL: National Energy Program Evaluation Conference, 1989, pp. 311-316.

Because energy evaluation and DSM programs are relatively young, few studies deal with long-term impacts of energy programs. This study is one of the few examples. This paper is about the persistence of net impacts and not the persistence of savings from an installed measure. The durability of net savings over the study years 1982-1987 was about 73%. The study is well done, and the authors note some important limitations to their work, including the lack of data on household characteristics that might have changed the estimates of net impacts. If you are interested in the long-term impact of savings, you should study this paper.

Train, K., P. Ignelzi, and M. Kumm. "Evaluation of a Conservation Program for Commercial and Industrial Customers," *Energy*, 10(10), pp. 1079-88.

This work studies the savings resulting from a conservation program directed at C&I customers in the Southern California Edison service area. After an energy audit, rebates were offered to customers for installing items like time clocks, photocells, load controllers, lighting-system changes, and skylights. The evaluation found that the cost per kWh of savings was less than the average of fuel costs of generation, indicating that the measures were cost-effective. This study is important because it is one of the early studies of C&I conservation and because of the techniques used to conduct the evaluation. Program participation was estimated with logit models. Savings estimates were based on regression techniques that took into account nonweather factors influencing load (hours of operation, square footage, etc.), weather-sensitive loads, and conservation-induced effects.

AUTHOR BIOGRAPHIES

Benson H. Bronfman is President of Bronfman & Associates, a consulting practice specializing in DSM program planning and evaluation. He is an expert in evaluation planning, process-evaluation theory and methods, and the application of research design and statistical approaches to analyzing DSM program impacts. His experience includes evaluations of DSM programs in all sectors and at all stages of development. He previously headed the DSM evaluation practice at ERC International and the Social Impact Analysis Group at Oak Ridge National Laboratory. He holds a Ph.D. in Political Science from the University of Oregon.

George L. Fitzpatrick is the President of Applied Energy Group, Inc., a technical consulting and software-development firm serving the electric-utility industry, primarily in DSM evaluation and implementation, load research, strategic planning, power-plant cost and performance analysis, utility-performance-standard development, and comparative economic analysis of utility investments. Prior to founding Applied Energy Group, Inc. in 1981, he was Vice President at Stone & Webster Management Consultants, Inc. He was also Manager of the Load- Research, Costing, and Forecast Division at Long Island Lighting Company. He holds an M.B.A. and a B.A. in Economics, both from St. John's University.

Elizabeth G. Hicks is the Director of Demand Planning at New England Power Service Company. She oversees the planning, design, and evaluation of all DSM programs, and the research and development on new DSM technologies. Previously, she was the Manager of Conservation and Load-Management Planning and Evaluation, responsible for the development of DSM programs, participation in collaborative planning processes, and evaluation of DSM programs. She has a B.A. in Economics and Mathematics from the College of William and Mary and an M.S. in Resource Systems and Policy Design from Dartmouth College.

Eric Hirst is a Corporate Fellow at Oak Ridge National Laboratory. He received a Ph.D. in Mechanical Engineering from Stanford University in 1968. Since 1970, Hirst has been at ORNL. He directs ORNL's work on integrated resource planning, which includes projects on uncertainty, modeling, collaboratives, and evaluation of DSM programs.

Marc G. Hoffman, Vice President at XENERGY, is a resource economist specializing in utility planning and evaluation with 11 years of experience in supply- and demand-side planning. For several years, he has developed leading applications of end-use analysis to utilities' load-forecasting and demand-side-planning needs. Most recently, he is leading the development of XENERGY's evaluation efforts. He holds an M.S. in Agricultural Economics from the University of Connecticut and a B.A. from Brandeis University.

Kenneth Keating began his career in energy program evaluation at the Bonneville Power Administration two weeks after receiving his Ph.D. from Washington State University in 1982. Previously, he had worked on evaluations of traffic programs and solar energy. During his 10 years at Bonneville, he has been responsible for more than 140 evaluation projects. He has been active on the planning committees for both the ACEEE Summer Study and the five International Conferences on Energy Program Evaluation, which, in 1991, honored him for his work in the development of the field.

Harvey Michaels, Senior Vice President at XENERGY, is responsible for the company's R&D in end-use methods and software as well as for corporate operational management. In his 13 years at XENERGY, he has led a number of significant projects in end-use forecasting and conservation analysis, including the development of industrial energy-management guidebooks for 15 trade-association energy programs, development of an energy-auditing system, and an end-use research system. He has an M.S. in Economics and Policy and a B.S. in Civil Engineering from Massachusetts Institute of Technology.

Steven Nadei is a Senior Associate with the ACEEE. Nadel's areas of emphasis are programs, policies, and technologies to reduce electricity demand. This work focuses on utility DSM efforts and on the efficiency of appliances, lighting, and motor systems. Prior to joining ACEEE, he spent two years planning and evaluating electricity conservation programs for the New England Electric System. He has a B.S. in Government, an M.A. in Environmental Studies from Wesleyan University, and an M.S. in Energy Management from New York Institute of Technology.

Jane S. Peters is a Project Director with Barakat & Chamberlin, Inc., and co-manager of the Portland, Oregon, office. She specializes in both impact and process evaluation of DSM programs. She is an expert in process and market evaluations focusing on customer response and on the effectiveness of program delivery and administrative mechanisms. She received her Ph.D. in Urban Studies from Portland State University and her A.B. in Psychology from Occidental College.

John Reed is group leader of the Energy and Environmental Applications Group, Energy Division, Oak Ridge National Laboratory. He has done considerable work in the evaluation of energy-conservation programs. He helped design and conduct a series of workshops on program evaluation for state energy offices. He worked on the load-management portion of the Athens Automation and Control experiment in Athens, Tennessee. Most recently, he has been involved in the design of computer-based decision-support systems. He received a Ph.D. in sociology from Cornell University.

William Saxonis is Manager of the New York State Energy Office's Evaluation Unit, which is responsible for evaluating a wide range of energy-conservation programs. He is the author of several articles dealing with free-rider measurement and evaluation planning. He currently serves on the planning committee of the International Energy Program Evaluation Conference and leads a project team to evaluate the long-term effectiveness of energy retrofits in multifamily buildings. He earned an M.A. in Political Science from the Rockefeller College of Public Affairs and Policy.

Andrew Schön, Vice President of DSM Engineering at XENERGY since 1979, is the architect of a hybrid engineering and statistical methodology to develop end-use load shapes. He has served as project director for numerous end-use research, evaluation, and engineering studies of DSM technologies. Additionally, he oversees and directs all engineering design and monitoring services. Mr. Schön has a B.S. in Mechanical Engineering from the University of Massachusetts and an M.S. in Technology and Policy from the Massachusetts Institute of Technology.

Daniel M. Violette is Senior Vice President of RCG/Hagler, Bailly, Inc. He has extensive experience in the evaluation, design, testing, and implementation of DSM programs. He was the lead author of *Impact Evaluation of DSM Programs* and has managed a number of large evaluations of DSM program impacts. He has developed and used econometric, engineering, and hybrid models to best use limited and uncertain information. He is a founding member of the Association of Demand-Side Management Professionals, and he currently chairs the Association's committee on DSM program monitoring and evaluation. He has a Ph.D. in Economics from the University of Colorado.

INTERNAL DISTRIBUTION

- | | |
|---------------------|---------------------------------|
| 1. V. D. Baxter | 24. P. Leiby |
| 2. L. Berry | 25. J. M. MacDonald |
| 3. D. S. Bjornstad | 26. L. N. McCold |
| 4. R. B. Braid | 27. V. C. Mei |
| 5. M. A. Brown | 28. W. R. Mixon |
| 6. R. S. Carlsmith | 29. S. Purucker |
| 7. R. Cantor | 30. S. Rayner |
| 8. F. C. Chen | 31. J. H. Reed |
| 9. J. Christian | 32. D. E. Reichle |
| 10. G. Courville | 33. D. T. Rizey |
| 11. S. J. Dale | 34. M. Schweitzer |
| 12. P. D. Fairchild | 35. R. B. Shelton |
| 13. W. Fulkerson | 36. G. G. Stevenson |
| 14. M. B. Gettings | 37. J. N. Stone |
| 15. C. W. Hagan | 38. J. Van Dyke |
| 16. L. J. Hill | 39. J. M. Veigel (ORAU) |
| 17. E. Hillsman | 40. D. B. Waddle |
| 18. E. Hirst | 41. D. L. White |
| 19. P. J. Hughes | 42. T. J. Wilbanks |
| 20. C. R. Kerley | 43. ORNL Patent Office |
| 21. J. O. Kolb | 44. Central Research Library |
| 22. M. A. Kuliasha | 45. Document Reference Section |
| 23. R. Lee | 46. Laboratory Records (RC) |
| | 47-49. Laboratory Records Dept. |

EXTERNAL DISTRIBUTION

- | | |
|-----------|--|
| 50. | Bruce G. Buchanan, Computer Science Department, University of Pittsburgh, 206 Mineral Industries Building, Pittsburgh, PA 15260 |
| 51. | Allan Hirsch, Vice President, Environmental Sciences, and Director, Washington Operations, 5109 Leesburg Pike, Suite 414, Falls Church, VA 22041 |
| 52. | Martin Williams, Professor, Department of Economics, Northern Illinois University, DeKalb, IL 60115 |
| 53. | Helen Ingram, Director, Udall Center for Studies in Public Policy, University of Arizona, 803/811 East First Street, Tucson, AZ 85719 |
| 54. | Calvin MacCracken, President, Calmac Manufacturing Corporation, 101 West Sheffield Avenue, P. O. Box 710, Englewood, NJ 07631 |
| 55.-65. | OSTI, U. S. Department of Energy, P. O. Box 62, Oak Ridge, Tennessee 37831 |
| 66. | Office of Assistant Manager for Energy Research and Development, DOE/ORO, P. O. Box 2001 Oak Ridge, TN 37831-8600 |
| 67.-1050. | External Energy Efficiency and Renewables Section Distribution Mailing List and extra copies to E.M. Schorn, 4500N, H-19A |

END

**DATE
FILMED**

3 / 26 / 92

