LBL-37158
UC-1321

Lawrence Berkeley Laboratory

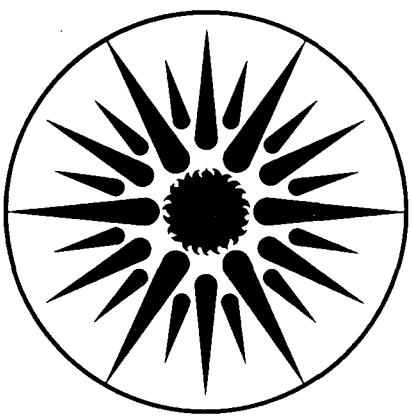
UNIVERSITY OF CALIFORNIA

ENERGY & ENVIRONMENT DIVISION

A Framework for Improving the Cost-Effectiveness of DSM Program Evaluations

R. Sonnenblick and J. Eto

September 1995



ENERGY
AND ENVIRONMENT
DIVISION

Prepared for the U.S. Department of Energy under Contract Number DE-AC03-76SF00098

DISCLAIMER

LBL-37158
UC-1321

A Report from the
Database on Energy Efficiency Programs (DEEP) Project

A Framework for Improving the Cost-Effectiveness of DSM Program Evaluations

Richard Sonnenblick and Joseph Eto

September 1995

Energy Analysis Program
Energy and Environment Division
Lawrence Berkeley Laboratory
Building 90-4000
Berkeley, California 94720

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Utility Technologies, of the U.S. Department of Energy under Contract No. DE-AC03-76F00098. The DEEP Project has also received funding from the New York State Energy Research and Development Authority, the Bonneville Power Administration, the Rockefeller Family and Associates, the Electric Power Research Institute, and in-kind support from the Energy Foundation.

Table of Contents

List of Figures

List of Tables

Acknowledgments

We would like to thank the following individuals for helping us gather information for the report: Michael Blasnik, Ellen Franconi, Marvin Horowitz, Joe Huang, Bruce Nordman, Michael Ozog, Mary Ann Piette, Rick Ridge, Peter Shaw, and Ken Train.

We would also like to thank the following individuals for their help in reviewing and providing us with comments on our draft report: John Amalfi, Les Baxter, Meg Fels, Miriam Goldberg, Chuck Goldman, Liz Hicks, Eric Hirst, Adrienne Kandel, Ken Keating, Suzie Kito, Les Owashi, H. Gil Peach, Diane Pirkey, Rick Ridge, Greg Rodd, Jeff Schlegel, Leslie Shown, Ed Vine, Dan Violette, and Roger Wright.

Without their help, the report could not have been written. However, the authors alone assume all responsibility for the opinions, errors, and omissions contained in the report.

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Utility Technologies, of the U.S. Department of Energy under Contract No. DE-AC03-76F00098. The DEEP Project has also received funding from the New York State Energy Research and Development Authority, the Bonneville Power Administration, the Rockefeller Family and Associates, the Electric Power Research Institute, and in-kind support from the Energy Foundation.

Abstract

The prudence of utility demand-side management (DSM) investments hinges on their performance, yet evaluating performance is complicated because the energy saved by DSM programs can never be observed directly but only inferred. This study frames and begins to answer the following questions: (1) How well do current evaluation methods perform in improving our confidence in the measurement of energy savings produced by DSM programs? (2) In view of this performance, how can we best allocate limited evaluation resources to maximize the value of the information they provide. We review three major classes of methods for estimating annual energy savings: tracking database (sometimes called engineering estimates), end-use metering, and billing analysis and examine them in light of the uncertainties in current estimates of DSM program measure lifetimes. We assess the accuracy and precision of each method and construct trade-off curves to examine the costs of increases in accuracy or precision. We demonstrate several approaches for improving evaluations for the purpose of assessing program cost effectiveness. The methods can be easily generalized to other evaluation objectives, such as shared savings incentive payments.

Executive Summary

American utilities spent nearly three billion dollars on demand-side management (DSM) programs in 1994.[1] The prudence of these investments hinges on their performance, yet evaluating performance is complicated because the energy saved by DSM programs can never be observed directly but only inferred. Utilities currently rely on a variety of methods, drawn from a variety of academic disciplines, including engineering, statistics, social psychology, and economics. Given the relative newness of utility DSM programs, it is not surprising that no consensus has emerged on a single best evaluation method. There are significant unanswered questions regarding how much evaluation, and what types, are appropriate in view of the expected benefits and costs of the programs.

The objective of our study is to frame and begin to answer the following questions: (1) How well do current evaluation methods perform in improving our confidence in the measurement of energy savings produced by DSM programs? (2) In view of this performance, how can we best allocate limited evaluation resources to maximize the value of the information they provide? We approach the subject humbly and do not presume that there is a single best method for conducting a DSM evaluation; we acknowledge that all evaluation methods provide some information. The quantity and types of information one needs depend on the intended use of the evaluation results. Therefore, how much one should spend acquiring DSM evaluation information depends on how much the information is worth.

Our study examines current practices in the evaluation of DSM programs that target lighting in the commercial sector, both because of their significance as major elements of most utility's DSM program portfolios and because they have been the subject of extensive evaluations. We examine different evaluation methods from the particular objective of improving our knowledge regarding the cost effectiveness of these programs. Establishing cost effectiveness is not the only objective of an evaluation; establishing shareholder incentives paid to a utility for running a DSM program is another. The methods we develop are general and can be readily extended to these and other evaluation objectives.

Although ours is not the first study to recognize that the value of information and the cost of acquiring it should be important inputs into decisions about evaluation methods, we believe ours is the first comprehensive application of this insight to the practice of DSM program evaluations. Moreover, in developing the information required to allocate evaluation resources cost-effectively, we have uncovered substantial new information on the strengths and limitations of current evaluation methods.

[1] EIA. "Annual Energy Outlook 1994." Energy Information Administration, Washington, 1994.

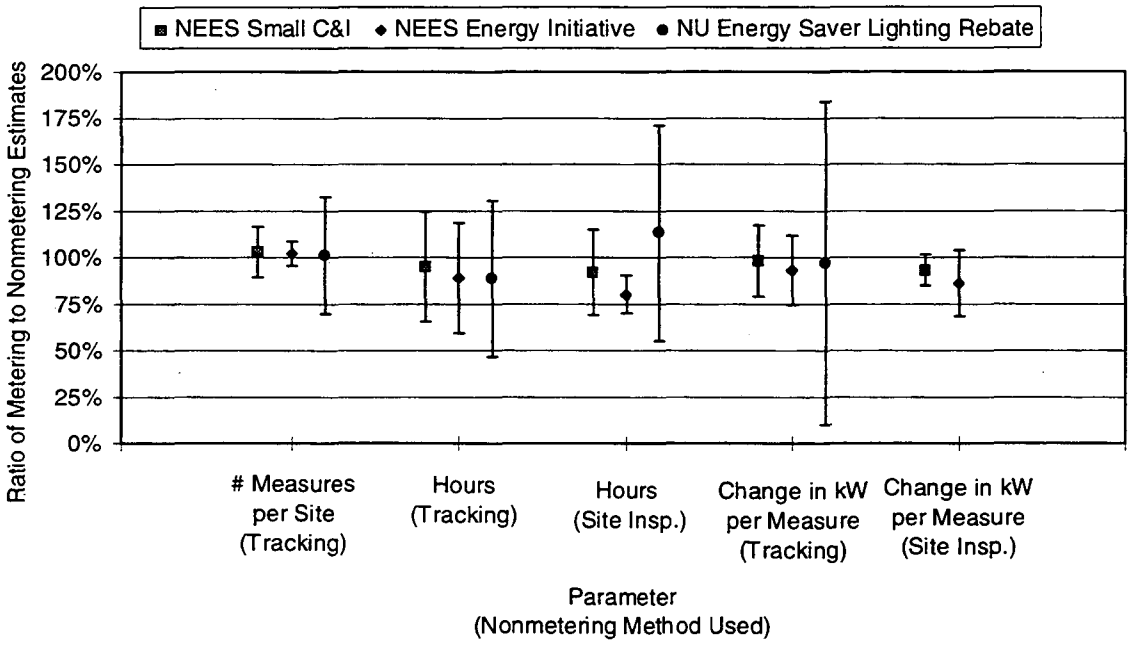
EX.1 The Performance of Current DSM Program Evaluation Methods

A major contribution of our study is a detailed assessment of the performance of current evaluation methods. We reviewed three major approaches for estimating annual energy savings: tracking database (sometimes called engineering estimates), end-use metering, and billing analysis. We also examined current estimates of DSM program measure lifetimes. The objective of our assessment was two-fold: First, we attempted to systematically characterize what is known about the accuracy and precision of current methods, based on reviews of recent evaluation studies and on our own analyses. Second, we constructed "trade-off" curves to examine the costs of increases in accuracy or precision. What follows is a summary of key findings on the performance of these methods.

EX.1.1 Tracking Database Estimates of Savings

Although "engineering estimate" is traditionally defined as a method that does not rely on measured consumption data (such as load metering or bills), we believe the term does not adequately describe the range of current methods. Moreover, the pejorative implications of the term are inappropriate given the often substantial after-the-fact performance data, such as site inspections and spot metering, that are routinely incorporated by these approaches.

Figure EX-1. Comparison of Accuracy and Precision of Tracking Database Estimates of Savings. Realization rates for individual components of savings (number of measures, hours of operation, changes in installed capacity or delta kW) from tracking databases and site inspections are compared to more accurate end-use metering estimates of the same quantities.



Nevertheless, for a small sample of studies in which we could directly compare tracking database estimates of savings to end-use metering, we found considerable variation in bias and precision (see Figure EX-1). Because an evaluator, without additional evaluation information, has no means of estimating the bias and precision of his/her tracking database estimate, we conclude that tracking database estimates alone are not reliable. Among the computational elements used in tracking databases (i.e., number installed, change in load per measure, and hours of operation), we found hours of operation were the largest contributor to bias and imprecision in annual savings estimates. If future studies with a larger sample of programs can confirm these findings, it would suggest additional attention should be given to inexpensive and accurate methods for improving estimates of hours of operation.

EX.1.2 End-Use Metering Estimates of Savings

Although end-use metering offers the promise of being the most accurate method for estimating lighting energy savings,[2] we find that contemporary end-use metering studies are often limited. These limitations stem ultimately from the high cost of end-use metering, which, because of its cost, is generally implemented for only a subset of program participants, for a subset of affected circuits, and for only a few weeks at a time. Clearly, rationing these high costs to maximize the value of the information produced by this method is an important evaluation objective.

The imprecision of limited-duration metering and the effects of HVAC/lighting interactions were not addressed in the majority of studies we reviewed. We estimate that these uncertainties reduce the precision of end-use metering estimates by approximately 20%. This reduction could be tempered by: (1) longer-duration metering, or, (2) a better understanding of interaction effects coupled with detailed information about each customer's HVAC system.

Our sample of office building lighting hours of operation data suggests that hours vary seasonally. On average, hours of operation are half an hour longer in the winter and half an hour shorter in the summer than during the shoulder months. Neglecting to account for the season in which metering is performed could bias the estimate of hours of operation and the resulting estimate of annual savings.

HVAC/lighting interaction effects increase program electricity savings for most non-electrically-heated office buildings. Omitting this effect from consideration can result in a 5-15% downward bias in annual savings estimates, depending on the climate and particular HVAC equipment used.

Small sample metering studies depend heavily on the representativeness of the metered sample. Most evaluators already stratify the population to select a representative sample of participants and then select representative equipment within each facility. Detailed metering results from evaluations would allow an

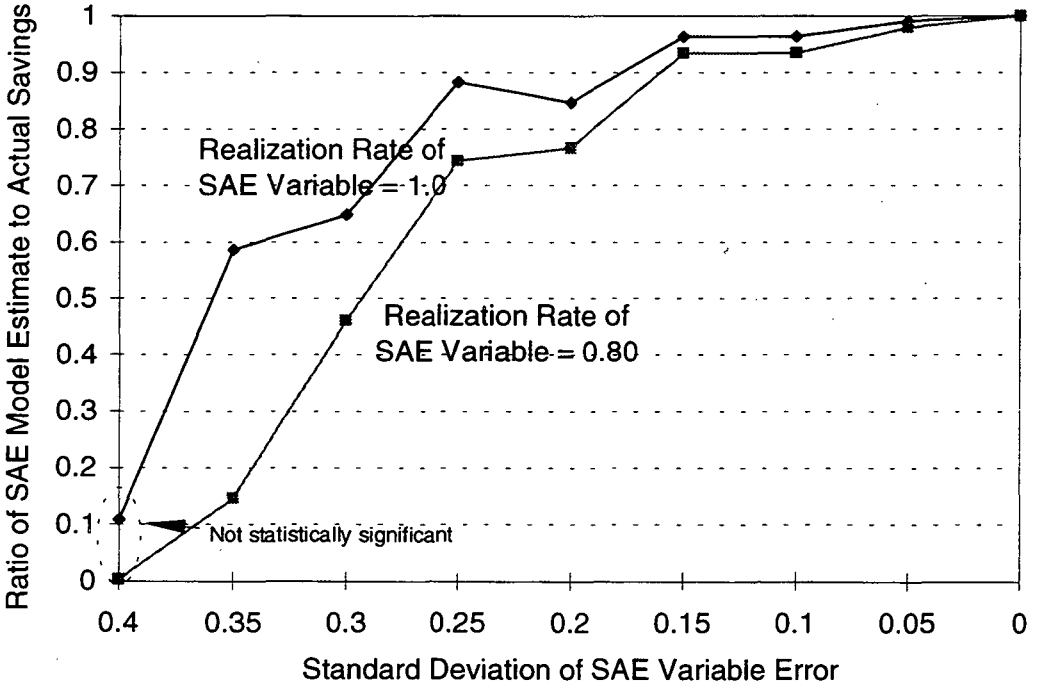
[2] It can potentially be most accurate because it measures a quantity that most closely resembles actual energy savings: the actual consumption of individual pieces of equipment before and after a efficiency-enhancing retrofit.

assessment of differences in equipment operation across participants, facility types, and facility zones. Until such detailed reporting is commonplace, which will enable analyses of equipment operating differences to be performed, the representativeness of current metered samples will remain uncertain.

EX.1.3 Billing Analysis

Regression-based analyses of customer billing information are perhaps the most widely used post-program evaluation method. We examined a range of methods using a simulated data set of 500 buildings where we could precisely control the level of savings, influence of weather, and changes in building hours of operation. In evaluating the popular Statistically Adjusted Engineering or SAE model, which introduces site-specific engineering estimates of savings, we confirmed the magnitude of a widely-recognized but underappreciated limitation of the method, namely, that its reliability depends strongly on the quality of the initial engineering estimate of savings (see Figure EX-2). Based on our earlier findings regarding typical levels of imprecision and bias in these estimates, we found that the SAE model did not perform as well as simpler time-series regression methods. We believe this is a major finding and, if confirmed by subsequent application of our methods to a wider range of situations, represents a particularly sobering conclusion for the evaluation community.

Figure EX-2. SAE Model Realization Rate Bias Dependent on Engineering Estimates. The realization rate for an SAE model depends on both the accuracy and precision of the underlying engineering estimate of savings. Even a completely unbiased estimate leads to an erroneous realization rate of less than 1.0 when the standard deviation of the estimate is greater than, say, 0.25 (see upper curve).



We also found that inclusion of comparison groups in time-series regression can greatly improve the precision of annual savings estimates, at moderate costs. When the DSM program reduces customer consumption by a small amount (4% in our simulation), incorporating nonparticipant data improves the precision of savings estimates by a factor of three. For programs that save a larger proportion of customer electricity consumption, the improvement is smaller but still significant.

Table EX-1. Summary of Annual Savings Evaluation Methods Examined.

Method To Estimate Annual Savings	Effects Treated/ Accounted For	Primary Accuracy Limitations	Potential Bias In Annual Savings Estimate	Potential Imprecision In Annual Savings Estimate
Tracking database (engineering estimate)		Baseline equipment, usage patterns, equipment installations not verified, efficiencies from mfr. Specifications, requires gross assumptions regarding consistent customer behavior	Over/underestimation of baseline and program equipment efficiencies, hours of operation	Precision not estimated
Site inspection	Baseline equipment (with pre-installation inspections) and efficient equipment specification errors in tracking database, hours of operation (from auditor/ customer survey)	Still simplifies equipment usage patterns, does not verify equipment energy consumption at customer sites	Over/underestimation of operating hours or equipment efficiencies by auditors/ in customer surveys	Precision not estimated
End-use metering	Variations in equipment usage, baseline usage (if pre/post metering)	Metered sample may not accurately represent population, metering of limited time duration, no comparison group	Seasonal variations in equipment usage, hvac/lighting interaction effects, unrepresentative sample of customers/equipment/ building zones metered	Limited duration metering, extrapolation from sample to population
Customer bill-based econometric models	Changes in equipment usage, changes in weather, changes in baseline energy use (with comparison group)	Provides little understanding of program strengths/ weaknesses or justification for its savings estimate, requires one year of post-program data	Non-normality of data/error term, improper model specification, improper comparison group, inadequate variability in data, low signal/noise ratio	Improper model specification, inadequate variability in data, low signal/noise ratio

EX.1.4 From Annual Savings to Lifetime Savings: Economic Measure Lifetime and Its Influence on the Cost of Conserved Energy

The value of DSM programs depends on both annual savings and the economic lifetime of the measures. We caution that the current practice of simply estimating equipment measure lifetimes based on expert judgment may be highly unreliable.

We demonstrate that measure lifetimes represent a significant source of uncertainty for estimates of energy savings (see Table EX-2). The importance of uncertainties in measure lifetime for the cost of energy savings depends on the effect size, and the method chosen to estimate annual energy savings. With the exception of methods involving time-series analyses of a small effect size, measure lifetime is the dominant contributor to uncertainty (i.e., the rank correlation for measure lifetime is greater than that for annual savings). In every case, the contribution of measure lifetime to uncertainty is comparable to that of the annual savings estimates.

Table EX- 2. Importance of Uncertainty in Cost of Conserved Energy Inputs.

		Rank Correlations (1 indicates maximum importance)	
Top-Down Method	**Effect Size (savings per participant)**	**Annual Savings Estimate**	**Measure Lifetime**
Time-Series	***Small***	0.88	0.49
	Medium	0.61	0.74
	Large	0.46	0.87
Time-Series Cross Section	***Small***	0.50	0.82
	Medium	0.39	0.90
	Large	0.37	0.92
Time-Series Cross Section w/Lagged Dependent Variable	***Small***	0.44	0.90
	Medium	0.23	0.98
	Large	0.22	0.99

A comparison of the results of recent equipment lifetime studies (we located only two complete studies) with measure lifetime estimates from evaluations of 20 commercial lighting DSM programs suggests that the lifetime estimates commonly used today by utilities could be biased upwards, resulting in estimates of the cost of conserved energy that are biased downwards. Now that DSM is maturing as an energy resource, it is time for additional studies that verify, through surveys of participants, estimates of measure lifetime.

EX.2 Efficient Allocation of Evaluation Resources

Our objective in examining the performance of current evaluation methods is to offer recommendations on how to improve the current practice of conducting evaluations by explicitly recognizing the tradeoffs involved in evaluation method

performance, costs, and the value of evaluation information. We demonstrate several approaches from the particular evaluation objective of assessing program cost effectiveness. Our methods can be easily generalized to other evaluation objectives, such as shared savings incentive payments.

We begin by comparing the likely impact of improved evaluation methods on the cost-effectiveness findings for 20 recent commercial lighting programs. We find relative precisions in the range of 90/50 (±50% at a 90% confidence interval) are sufficient to confirm the cost effectiveness of the majority of programs from this sample of 20. Table EX-3 illustrates, for differing initial levels of program cost effectiveness, the effects of different levels of evaluation method precision on ultimate program cost effectiveness. Confidence that programs are cost effective decreases as the initial TRC ratio approaches 1.0 and as evaluation method precision decreases. Thus, we conclude that the 90/10 precision standard often required of evaluations may only rarely be cost-justified from the standpoint of confirming program cost effectiveness.

Table EX-3. Fraction of Distributions Representing Non-Cost-Effective Programs.

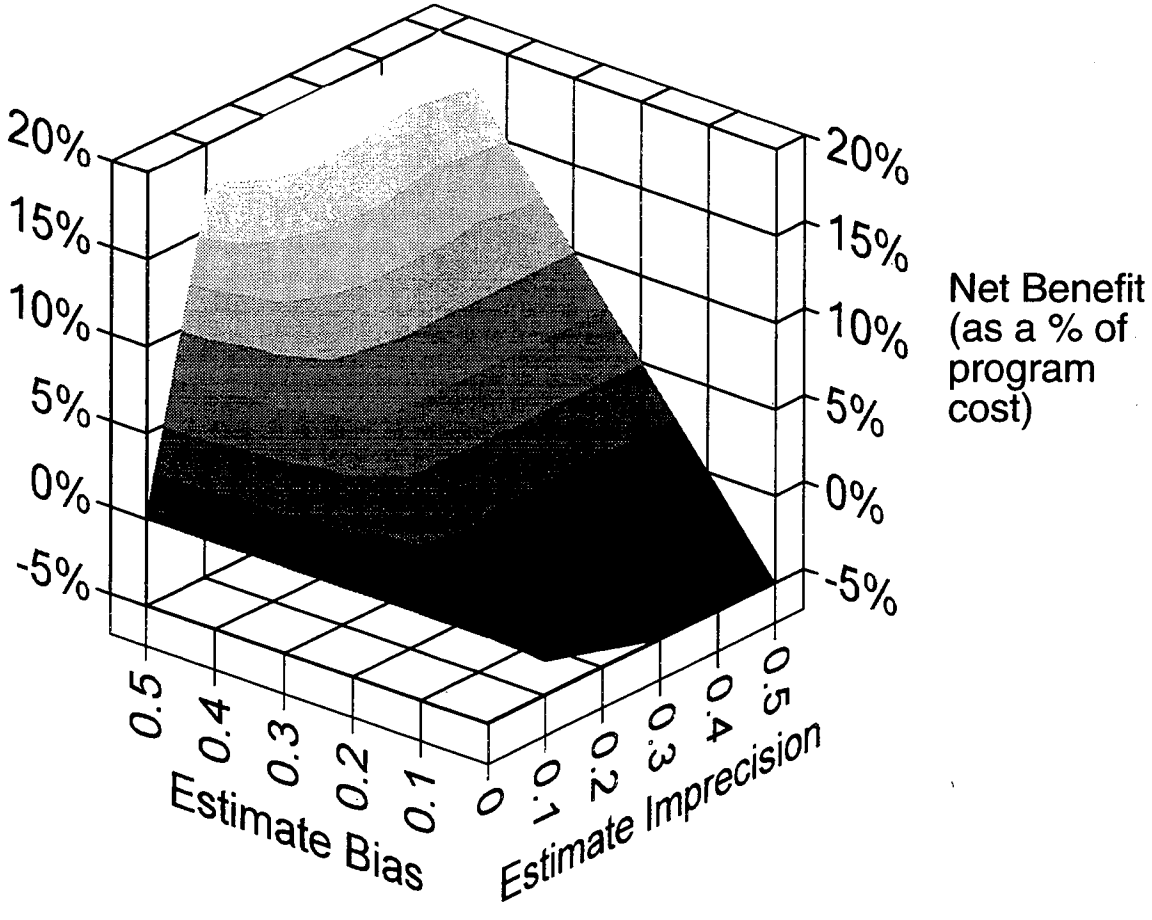
Mean TRC Test Ratio	Savings Estimation Method	Precision	Percent of Distribution Less Than 1.0
Low (1.1)	End-Use Metering	Low (±50%)	40%
		Medium (25%)	29%
		High (10%)	11%
	Econometric	Low (15%)	19%
		Medium (10%)	11%
		High (5%)	3%
Medium (1.8)	End-Use Metering	Low	7%
		Medium	—
		High	—
	Econometric	Low	—
		Medium	—
		High	—
High (4.2)	End-Use Metering	Low	1%
		Medium	—
		High	—
	Econometric	Low	—
		Medium	—
		High	—

However, biases in measure lifetime estimates can also cause misstatements of cost effectiveness. Coupling such biases with imprecise estimates of measure costs and annual energy savings can further decrease evaluator confidence in program cost effectiveness. The biases we identify in current practice of calculating annual savings and measure lifetime estimation are sufficient to mislabel non-cost-effective programs as cost effective, even when estimated Total Resource Cost Ratios are as high as 2.0. Unfortunately, although we can the identify sources and likely magnitudes of bias in current methods, we cannot offer definitive guidance on the

bias likely to be present in any particular application of methods. A notable exception is our earlier characterization of the bias introduced by the SAE method in the presence of bias and imprecision in initial engineering estimates of savings.

Finally, we consider the issue of evaluation resource allocation directly. For a decision to continue funding a program based on cost effectiveness, this requires: (1) a subjective estimate of the chances that the program is actually not cost-effective, in the face of any evaluation results, and (2) an estimate of the resources that could be misallocated to the program in the following year. We represent the decision to fund as being based on (a) a mean evaluation estimate of cost effectiveness, or (b) an estimate of cost effectiveness that includes imprecision. The difference between (a) and (b) is the value of including uncertainty in the program screening decision. The product of (1) and (2) is the expected value of future misallocated resources. The results for a hypothetical program, with value expressed as a percentage of total program cost, are plotted in Figure EX-3.

Figure EX-3. Expected value of including uncertainty: TRC estimates in the low (mean total resource cost ratio =1.1) range.



The value of including a measure of evaluation imprecision depends on the mean estimate of cost effectiveness and its bias and imprecision. When the imprecision is zero, there is no [intuitive] value in considering an estimate of imprecision. As imprecision increases, but more importantly as bias increases, the value of taking imprecision into account increases. Our analysis leads us to conclude that taking estimate imprecision into account can mitigate the effects of estimate bias and imprecision, when evaluation information is used to screen ongoing DSM programs. Moreover, including imprecision in program screening decision making is more valuable when mean program cost-effectiveness ratios are close to one.

EX.3. Concluding Thoughts

The introduction of competitive forces in the industry is creating substantial pressures for utilities to control costs. Formal decision-analytic approaches to ration DSM program evaluation resources offer the potential to guide cost control decisions in a systematic and defensible fashion that maximizes the value of evaluation expenditures. Application of these approaches, however, requires detailed information on the performance of evaluation methods. This information is not yet widely available. Hence, we recommend increased effort by future evaluation efforts to report intermediate findings, especially on precision, so that a more comprehensive and reliable base of information upon which to ground these decisions can be developed.

Introduction and Overview

American utilities spent nearly three billion dollars on demand-side management (DSM) programs in 1994.[1] The prudence of these investments hinges on their performance, yet evaluating performance is complicated because the energy saved by DSM programs can never be observed directly, but only inferred. Given the relative newness of utility DSM programs, it is unsurprising that consensus has not emerged on a single best evaluation method. Utilities currently rely on a variety of methods, which in turn are drawn from a variety of academic disciplines, including engineering, statistics, social psychology, and economics. There are significant unanswered questions of how much evaluation, and what types, are appropriate in view of the expected benefits and costs of the programs.

The objective of our study is to frame and begin to answer the following questions: (1) How well do current evaluation methods perform in improving our confidence in the energy savings produced by DSM programs? (2) In view of this performance, how can we best allocate limited evaluation resources to maximize the value of the information they provide? We approach the subject humbly in that we do not presume there is a single best method for conducting a DSM evaluation. Instead, we start by acknowledging that all evaluation methods provide some form of information. The quantity and types of information one needs depends on the intended purpose of the evaluation result. Therefore, how much one should spend acquiring this information depends on how much the information is worth, in view of the cost of obtaining it.

Our study examines current practices in the evaluation of commercial sector lighting energy efficiency DSM rebate programs, both in view of their significance as major elements of most utility's DSM program portfolios and because, as a result, they have been the subject of extensive evaluations. We examine the value of different evaluation methods from the particular objective of improving our knowledge regarding the cost-effectiveness of these programs. This, of course, is not the only objective of an evaluation; establishing shareholder incentives paid to a utility for running a DSM program is another. The methods we develop are general in nature and can be readily extended to consider this and other objectives.

While the method is general, it is important to recognize the constraints we have placed on the scope of our investigation. As indicated, we apply our methods to a study of only one particular, albeit popular, type of DSM program, namely, those that offer rebates to commercial sector customers to retrofit or replace existing lighting systems. For these programs, we are concerned only with methods for estimating the direct annual energy savings attributable to them. We do not examine evaluation methods that attempt to measure the level of free-ridership or spillover from these

[1] EIA. "Annual Energy Outlook 1994". Energy Information Administration, Washington, 1994.

programs. Finally, while we explicitly examine the uncertainty in estimates of energy savings, we do not consider uncertainty in the cost of these programs or in the estimates of the utility supply costs avoided by these programs.

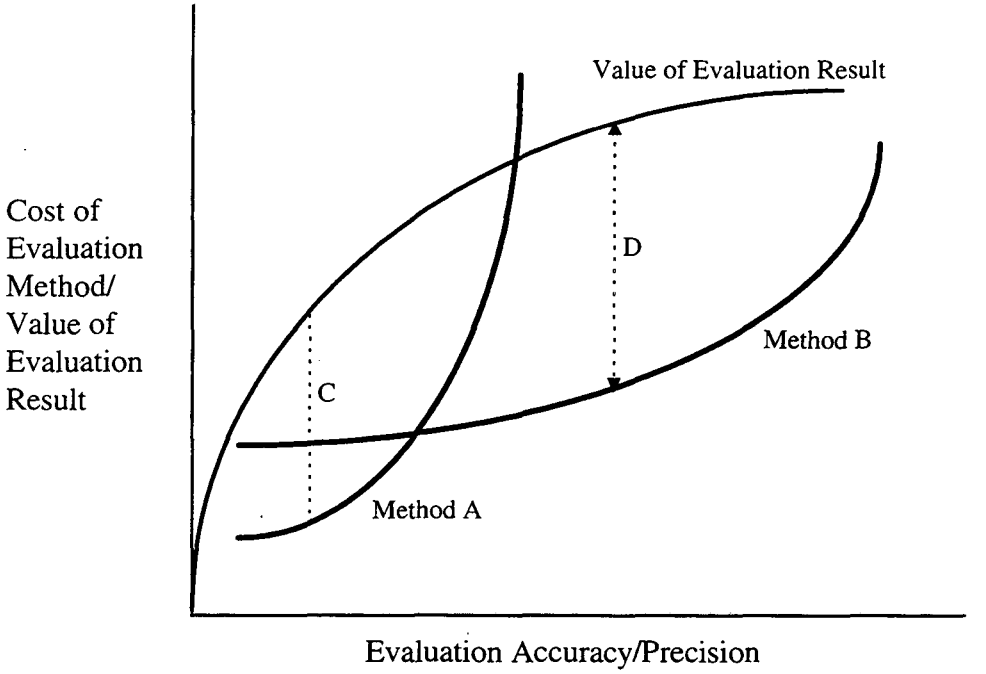
The remainder of this chapter motivates our study. First, we expand the above conception of the DSM program evaluation problem, which illustrates the notion of trading-off the cost of conducting one or another evaluation method against the anticipated benefits associated with each method. Next, we identify a particular set of very important evaluation objectives as one basis for making these trade-offs: the cost of conserved energy, and the related, total resource cost net benefit of DSM programs. We then differentiate between bias and precision, which are two interrelated, yet very distinct features of evaluation results. We then describe a particular type of DSM program, commercial sector, lighting energy efficiency rebate programs, which we use to illustrate these trade-offs. These programs are of special interest because they often account for the largest part of a utility's DSM program portfolio (and, consequently, evaluation spending). Finally, we provide a detailed overview of the following chapters of the report.

1.1. Integrating Cost Information with Bias and Precision of Methods

The basic premise of our study is that a comparison of evaluation methods is of little practical use unless the costs of the evaluation methods are also compared. Integration of cost information with evaluation method results allows trade-offs between cost and the bias/precision of each method. In theory, one could construct a curve which explicitly described the tradeoff between evaluation cost and bias/precision, with each evaluation method represented by a point (or a range) on the curve. A sample of such a graph is given in Figure 1-1. If one also had a measure of the value of increases in bias/precision, then one could decide not only which evaluation method to choose (i.e., which curve to be on), but also at which point on the curve the difference between cost and value is maximized.

Each curve represents a group of similar evaluation methods (for example, similar methods which incorporate data which is more accurate but increasingly more expensive to collect). For a given level of accuracy, several methods may be available to provide similar results at different costs, as indicated in Figure 1-1 by line C. As increased unbiasedness or precision is required, evaluation methods with gentle curves would be favored over ones with steeper curves. In order to precisely determine the appropriate *level* of evaluation, information on the use of the resulting savings estimates and requirements for bias and precision must also be incorporated into the analysis. These are represented hypothetically as a value curve, which decreases in marginal value as bias or precision is increased. The optimal level of evaluation is found at the point where the distance between the two curves (and the net benefit of evaluation) is maximized, as indicated in Figure 1-1 by line D.

Figure 1-1. Hypothetical cost vs. accuracy curves



1.2. Evaluation Objective: The Cost of Conserved Energy

DSM program evaluations provide valuable information regarding program administration, management, costs, and benefits. The appropriate evaluation technique is dependent on the evaluation objective. Appendix A provides a summary of the most often cited objectives of DSM program evaluation, and the evaluation requirements of each objective.

This study focuses on assessing the ability of evaluation to provide accurate and precise estimates of the kilowatt-hour savings of DSM programs and resulting costs to society and to the utility. We express the program cost to society using the Cost of Conserved Energy (CCE) as a metric. The CCE can be used to express the levelized cost (over the life of program equipment) of a DSM program per kilowatt-hour of program savings attained. The equation for calculating the CCE is:

$$\text{Cost of Conserved Energy (¢ / kWh)} = \frac{\textit{Program Cost} \times \dfrac{i(1+i)^n}{(1+i)^n - 1}}{\textit{Annual Savings}}$$

Using a capital recovery factor with discount rate i, the term in the numerator levelizes the total program cost over the number of years n the program equipment is expected to operate. Because the CCE allows one to compare results of DSM programs with different costs, savings, and lifetimes, and because it enables the comparison of DSM programs with supply-side options in an integrated resource plan, the CCE is a quantity of interest to utilities, regulators, and DSM program planners. To evaluate cost effectiveness, the cost of conserved energy is often compared to the supply-side

costs DSM programs allow the utility to avoid. The ratio (known as the total resource cost ratio) of avoided costs to the cost of conserved energy provides one metric of a program's cost effectiveness: programs with a ratio greater than one are considered cost-effective. [2]

1.3. Assessing evaluation methods: Bias and Precision

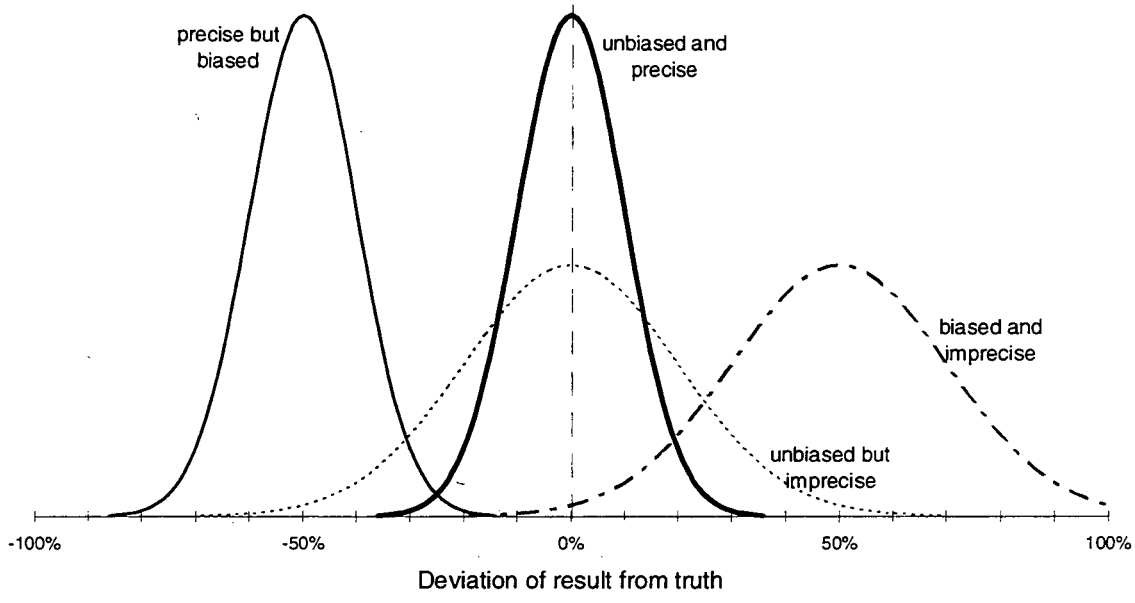
We use two different metrics to assess how well evaluation methods reveal a program's actual energy savings: bias and precision. A biased estimate systematically deviates from the true value, under or over estimating savings. For example, if a method consistently overestimated actual savings by 20%, that method would be considered biased.

The issue of precision is more esoteric. Many program evaluations omit all discussion of estimate precision, and report savings estimates as single values. But because of the difficulties associated with calculating program savings, any estimate of program savings is subject to some uncertainty. It is this uncertainty that one attempts to encapsulate in an expression of precision. An estimate which omits an estimate of precision is incomplete and can be misleading. For example, an estimate of annual savings of 5,000 kilowatt-hours (kWh) with a standard deviation of ±300 kWh is very different from an estimate of 5,000 ± 3,000 kWh. The latter estimate is of less use as a gauge of program savings, because it suggests that the actual savings could be considerably above or below the mean estimate of 5,000 kWh, while the former estimate is more precise, satisfying what is known as a 90/10 criterion; ±10% relative precision at a 90% confidence interval. Thus, figures reported without an estimate of that uncertainty are not as informative as those which include an estimate of uncertainty.

It is important to consider the relative importance of precision and bias. A precise but biased estimate is worth little, unless the magnitude of the bias is known. On the other hand, an unbiased but imprecise estimate can still be useful because, on average, it provides the correct value. Figure 1-2 illustrates the relationship between bias and precision.

[2] When the cost of the program to only the utility, and not society, is included, the ratio is called the utility cost test ratio.

Figure 1-2. Bias and precision in savings estimates



Biased, i.e., under- or over-estimates of savings, have important implications on several levels: For the utility, biased estimates of savings misinform about program cost-effectiveness. Biased over-estimates of savings may cause utilities to retain DSM programs which are not, in reality, cost effective. At the state regulatory level, overestimates of savings will result in utility overcompensation for lost revenues (for lost revenues which, in fact, were never lost) and allowed recovery of excessive shared savings incentives. Thus, the utility is allowed to collect additional, unjustified revenue from ratepayers. At the national level, plans to reduce national dependence on fossil fuels or reduce power plant emissions using DSM activities may fall short of desired goals if plans are based on studies which exaggerate potential savings.

An *imprecise* estimate of savings has some slightly different implications: Imprecision in annual savings or measure lifetimes can affect the mean cost of conserved energy estimate, because of the asymmetric nature of the cost of conserved energy distribution. (As will be shown in Chapter 6, however, the imprecision must be very large in order to significantly bias savings estimates.) Most of the concern regarding precision involves a fundamental desire for a precise estimate, but this desire is not necessarily based in the requirements of any particular use of the estimate. Regulatory agencies in California, among other states, require that precision of evaluation study results strive to reach 90/10 (or 80/20) criteria: evaluations should strive to attain ±10% relative precision using a 90% confidence interval (or ±20 relative precision using an 80% confidence interval).[3]

[3] Hanser, P., Violette, D., "DSM program evaluation precision: What can you expect? What do you want?", *Proceedings of NARUC's 4th annual national conference on IRP*, pp. 299-313, 1992.

In most cases the 90/10 criteria is applied to estimates of annual savings, without a similarly rigorous criteria being required for lifetime savings or for the resulting estimates of the cost of conserved energy. We demonstrate the difficulty of meeting a 90/10 criteria for the cost of conserved energy, which is usually of more importance to regulators than the annual savings estimate itself. We also show that in many cases such a criteria may be excessive because the cost-effectiveness of the program is assured at a much lower level of precision.

In this report, we use simulation techniques and data from past programs and evaluations to explore the extent of current methods' biases. For those methods where there is currently insufficient information to assess bias, we outline a framework which could be implemented with additional evaluation information. We also assess the precision associated with annual savings estimates obtained with different evaluation methods, and with the resulting cost of conserved energy estimates.

1.4. Focus on Commercial Buildings and Efficient Lighting

Rather than attempt to describe appropriate evaluation methods for all types of demand-side management programs, this report focuses on efficient lighting retrofits in commercial buildings. The analytic approach we use can be used to assess the evaluation methods for other sectors (residential, agricultural, and industrial), and for other efficient equipment types (heating/cooling equipment, process improvements, water heating, etc.). However, we have chosen to focus on commercial lighting because of the pervasiveness of utility sponsored commercial lighting programs, and the magnitude of electricity consumed by commercial lighting applications. The commercial building sector is responsible for about 10% of U.S. energy consumption.[4] Interior lighting is an important use of energy in commercial buildings, representing 40% of electricity use and 15% of total energy use in the commercial sector.[5] Interior lighting is also widely believed to be among the most cost-effective conservation opportunities available. Studies on the theoretical potential (sometimes called technical potential) for energy conservation have estimated that an additional 40-70% of lighting energy use could be cost-effectively saved in the commercial sector. In response to these studies, the vast majority of utilities engaged in DSM activities include programs which promote use of efficient lighting equipment in commercial buildings.[6]

---

[4] EIA. "Commercial Buildings Energy Consumption and Expenditures 1989". Energy Information Administration, Washington, 1992.

[5] EIA. "Annual Energy Outlook 1994". Energy Information Administration, Washington, 1994.

[6] Eto, J., Vine, E., Shown, L., Sonnenblick, R., Payne, C., "The Cost and Performance of Utility Commercial Lighting Programs", Lawrence Berkeley Laboratory, Berkeley, CA, LBL-34967, May 1994.

1.5. Summary of Report Organization

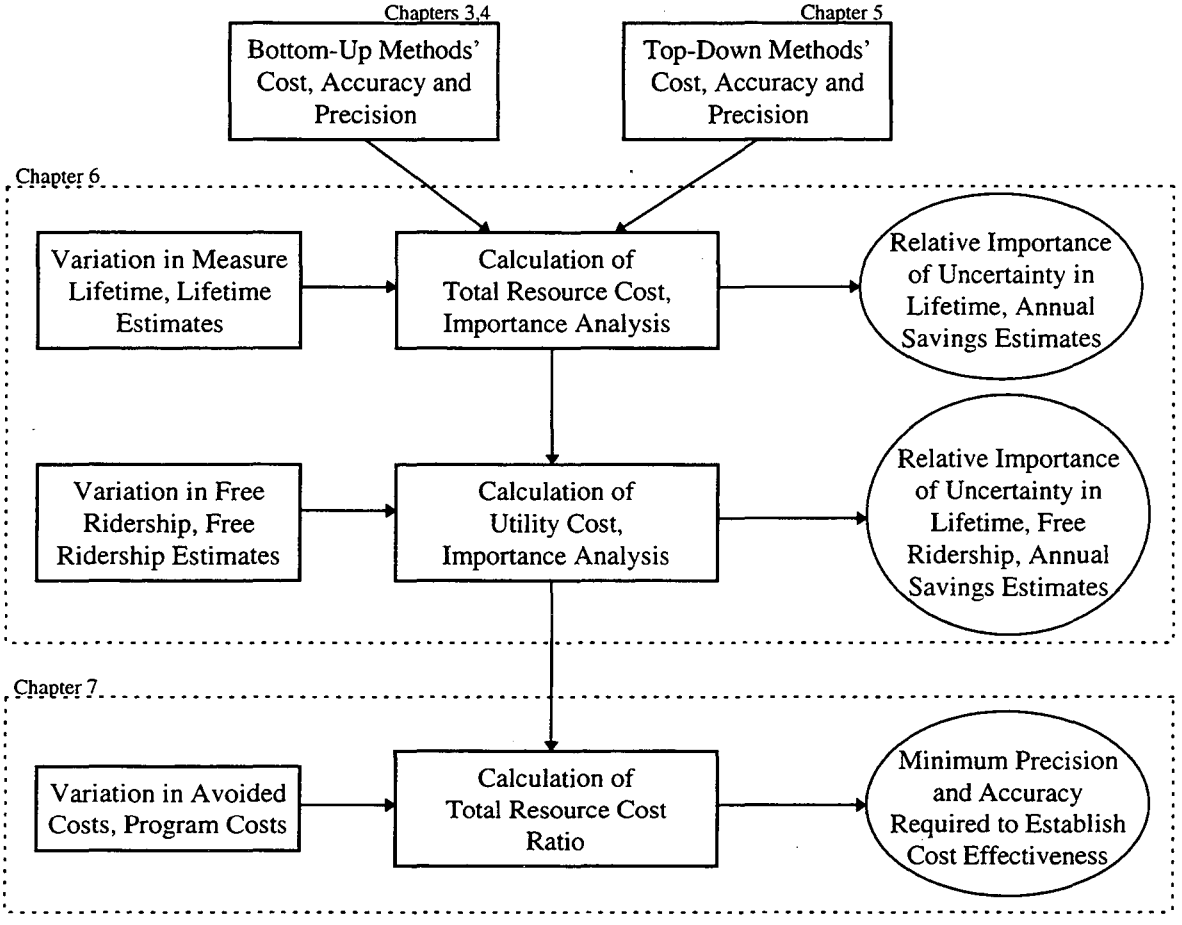
This report consists of seven chapters following this introduction and three appendices. A diagram of the report analysis is shown in Figure 1-3.

Chapter Two contains a detailed description of the evaluation methods examined in this report. It is based on a recent examination of evaluation methods used to estimate savings for twenty commercial lighting programs. The chapter uses the examination to develop formal expressions for each individual savings evaluation method. We distinguish between bottom-up and top-down evaluation methods. Bottom-up evaluation methods calculate savings from tracking databases (information collected by the utility on program participants), from participant site-inspections, and from end-use metering. We call these methods 'bottom-up' because they estimate, collect and measure consumption information at the individual equipment level, and require extrapolation and/or aggregation to construct site and program-wide savings estimates. We call methods which rely on customer billing data and econometric models 'top-down' methods because they examine monthly or annual consumption data, the most aggregate form of customer consumption data, and infer savings from comparatively smaller (compared to the size of the bills) but systematic changes in participant consumption patterns. We also summarize the limitations of the different evaluations which we quantitatively investigate in subsequent chapters. For those well-versed in current DSM practice, this chapter may be skipped.

The assessment of bottom-up evaluation methods is conducted in two separate chapters. Chapter Three uses end-use metering data from a handful of studies where highly disaggregated data were reported to evaluate the accuracy of savings estimates developed from tracking database estimates of savings and site inspection estimates of savings. We also describe a research plan to systematically assess bias and precision of tracking database and site inspection estimates of savings, as the data become available.

In Chapter Four, we use disaggregate long and short term metered data to assess bias and precision of end-use metering estimates of annual savings. To estimate the uncertainty of metered results, we incorporate data on changes in hours of operation over time, and differences in hours of operation across different areas of a building. These data provide us with enough information to discuss the potential error in shorter duration metering studies, and in studies where a sample unrepresentative of the participant population is selected. We combine our estimates of method bias and precision with information on data collection and analysis costs.

Figure 1-3. Overview of Research Plan



In Chapter Five, we examine top-down evaluation methods, which rely upon premise-level energy consumption data. In order to assess the performance of these methods, we construct a set of 250 'participant' and 250 'nonparticipant' commercial buildings, based on building construction data from EIA and several utilities. We then use DOE-2 to generate estimates of monthly energy consumption for these buildings for a year prior to, and a year after, implementation of a lighting retrofit. We apply the top-down evaluation methods to this monthly consumption data. The results of this application are combined with information regarding the costs of each method.

In Chapter Six, we combine annual savings estimates with estimates of measure lifetime, in order to estimate overall program savings. Using Monte Carlo techniques, we estimate the importance of uncertainty in annual savings, measure lifetime, and free ridership, and the overall uncertainty of the program savings estimate. Monte Carlo techniques are also used to estimate the uncertainty in levelized total resource cost.

In Chapter Seven, we describe the results of our analysis of estimate bias, precision, and the cost of conserved energy. For some programs, increases in savings estimate precision can dramatically reduce confidence in the cost effectiveness of the program.

For other programs, increased precision may be unnecessary to conclusively show cost-effectiveness.

In Chapter Eight, we review the implications of our analyses for future evaluation activities.

Appendix A contains descriptions of other objectives of DSM program evaluation. Appendix B contains a detailed description of the data and methods used to simulate the buildings used to assess the relative bias and precision of the top-down evaluation methods. In Appendix C, we compare the cost, precision, and bias of top-down and bottom-up methods used to estimate annual program savings. In addition to method precision and bias, we discuss other important attributes of each method, and discuss hybrid methods which incorporate several methods in a single framework.

Overview of Evaluation Methods

Evaluating the effects of a DSM program on energy consumption is a challenging task. The goal is to measure how much energy would have been consumed by program participants if the program had not occurred. Because energy savings can only be deduced and not directly observed, uncovering savings attributable to a program often utilizes quasi-experimental methods, which utilize information on both program participants and nonparticipants (a comparison group), both before and after program implementation. In this chapter we describe bottom-up and top-down evaluation methods in greater detail. We also describe methods used to estimate the lifetimes of program . The 20 evaluations scrutinized in a recent LBL report on commercial lighting rebate programs provide an opportunity to examine the evaluation methods used in the field to estimate these quantities.[1]

2.1. Annual Savings Methods

As described in the previous chapter, we classify evaluation methods that estimate annual program savings into two categories: bottom-up and top-down methods. Other DSM researchers focus on the distinction between "engineering" and "measured data" evaluation. We find this distinction misleading because all methods of estimating energy savings rely on engineering methods to some extent. For example, even end-use metering relies upon engineering technologies (meters and data loggers). Moreover, measurement does not necessarily imply that the measured value is reality: as we stated above, energy savings can only be deduced, not directly observed. Thus, no method elicits the absolute truth regarding annual program savings.

2.1.1. Examining Bottom-Up Energy Savings Models

For a commercial lighting program, the same basic information is used for all bottom-up evaluation methods: the number of measures of each type installed per site, each measure's kW consumption and the kW consumption of the measures being replaced, and each measure's hours of operation. The basic equation for energy savings which incorporates these terms is:[2]

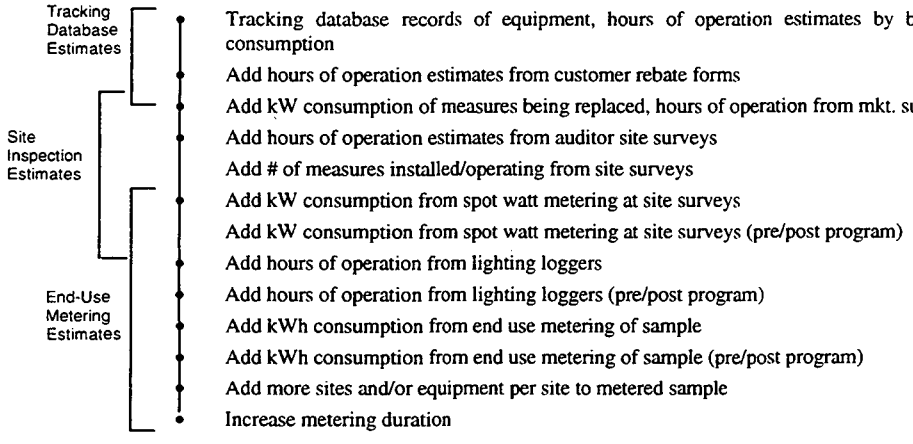
$$Energy\ Savings = \frac{measures}{site} \times \left(\frac{Watts}{measure_{old}} - \frac{Watts}{measure_{new}} \right) \times Annual\ Hours$$

[1] See Eto, J., Vine, E., Shown, L., Sonnenblick, R., Payne, C., "The Cost and Performance of Utility Commercial Lighting Programs", LBL, Berkeley, CA, May 1994, for complete information on the set of evaluations and programs in this sample.

[2] The equation representing savings from other end uses (e.g., heating, cooling) could be much more complex, involving non-linear relationships and greater numbers of parameters.

Bottom-up methods are often divided into three distinct categories: tracking database estimates (sometimes called engineering estimates) which rely upon utility database records of equipment installations and manufacturer-estimated efficiencies and equipment lifetimes; site inspection estimates which employ auditors to verify existence and operation of measures and adjust tracking database estimates based on interviews with customers; and end-use metering methods which rely upon measured consumption data from the efficient equipment installed at customer facilities. However, because these methods all utilize the same basic savings equations, it is more congruous to think of bottom up methods as a continuum, with each successive method utilizing more sophisticated (and costly) techniques to collect increasingly accurate data for the energy savings equation. A pictorial representation of this continuum is given in Figure 2-1.

Figure 2-1. Continuum of Bottom-Up Methods



At the most basic level, simple assumptions are made regarding the hours of operation for all buildings, and the equipment installed (and its efficiency) prior to the program intervention. The measures installed at each site are taken from program tracking database records of participants, often from rebate applications. These estimates are inexpensive to obtain but will probably not provide precise, unbiased estimates of savings, because these quantities are not based on actual data from the participating buildings.

Augmenting information at the basic level with specific information from surveys of program participants regarding their hours of operation and pre-program equipment is more accurate because it incorporates information about the specific participants' sites. However, participants' perception of hours of operation may over- or underestimate actual hours, and their knowledge of the equipment in place prior to participation may also be imperfect. In addition, hours of operation may vary for different parts of a facility in ways that are unbeknownst to the building managers or owners of those facilities.

On-site inspections by utility personnel can provide more consistent, and possibly more accurate, estimates of pre-program equipment efficiencies. On-site inspections also allow utility personnel to ask more detailed questions and visually verify the hours of operation for different areas in a facility. Equipment installed (and still operating) as a result of the program can also be verified. Visual inspection falls short of the accuracy associated

with actual measurement, but is considered more accurate than customer self-reports of operation and equipment description.

On-site inspections can include some measurement, such as instantaneous measurement of kilowatt loads using spot-watt meters. Other metering equipment can be installed during on-site surveys, including lighting loggers, which use photocells to measure hours of operation for lighting equipment, or current transformers connected to data logging equipment, which can be clamped on to circuits of program equipment to meter kilowatt loads over time.

The expense of site inspections and metering usually precludes the use of such methods on every piece of equipment at every participating site. Thus, a sample of sites, and of equipment at each site, is monitored for a limited period of time. The results are then extrapolated to the entire population of participants, over the lifetime of the equipment. The validity of this extrapolation is dependent on the representativeness of the sample and time-period selected for monitoring. Increasing the sample size (both across sites and within each site) and the metering duration improves the robustness of the extrapolation.

With explicit information about the initial uncertainty associated with each component of the tracking database estimate and information about the improvement in accuracy associated with each successive point on the continuum, an evaluator could make explicit, justifiable tradeoffs between the cost of additional data collection and the resulting anticipated increase in evaluation accuracy and precision.

Characterizing the uncertainty and variability of the various components of the bottom-up energy savings estimate requires extensive analysis of actual program data. The more detailed the available data, the more complete the resulting characterization of the variability in savings estimates at each point in the continuum. The difficulty is that the highly disaggregated data required to characterize the variability of savings estimates for each method is scarce. The costs of long term, large scale end-use metering are prohibitive, and not easily justified based on the evaluation needs of the program at hand.[3] Furthermore, most end use metering studies do not report (or even always keep on file) the disaggregated, intermediate-level results which one requires to undertake a characterization of method accuracy.

In the following sections, we discuss the factors which may bias tracking database and site inspection estimates of ΔWatts/measure and hours of operation.

2.1.1.1. Baseline Equipment Efficiency and Program Measure Efficiency

The efficiency of both the program equipment and the equipment being replaced is crucial to the estimate of savings: If equipment being replaced is more efficient than originally thought, savings will be less than predicted. If new equipment does not perform as well as expected, savings will also be reduced.

---

[3] Other reasons for metering, such as gathering customer load data, and verifying demand, as opposed to energy savings, could help justify the added expense.

In many evaluations, program planners make rough, back-of-the envelope estimates of the efficiency of existing equipment in participant facilities. However, estimates of the efficiency of existing lighting equipment are more accurate when based on some market or participant data. In San Diego Gas and Electric's retrofit program, it was originally assumed that equipment being replaced consisted of standard coil-core ballasts and F40 fluorescent lamps. However, site inspections revealed that approximately 50% of all ballasts were efficient coil-core ballasts, and 50% of all lamps were F34 watt Miser lamps. San Diego Gas and Electric revised its savings figures downwards for various measures by 18% to 48% to reflect more efficient base equipment discovered during site surveys.

Program Equipment can also be less efficient than initially thought. Spot watt metering by NU found that HID lamps in their retrofit program were 25% less efficient than originally estimated. NEES found lower than anticipated (85-95% of tracking database estimates) wattage reductions per measure in their commercial lighting programs. Such under or over estimates of equipment efficiency can bias subsequent estimates of program energy savings.

2.1.1.2.Hours of Operation

Tracking database estimates of savings are predicated on consistent use of the equipment. If equipment is used less than originally assumed, installing efficient versions of that same equipment will have a smaller than anticipated effect on energy consumption. Most of the programs that we surveyed required that participants indicate their facilities' hours of operation on the rebate application or audit form. However, more rigorous methods of obtaining hours of operation used by many of the programs demonstrated that participants often over-estimate their own equipment's hours of operation. Table 2-1 lists the results of hours of operation studies performed by the utilities in our sample.

Table 2-1. Summary of Hours of Use Studies in Sample

Utility	*Ratio of More Accurate to Less Accurate Estimate*	*Source of First Estimate*	*Source of Second Estimate*
CMP	0.70	Customer self-reports	189 fixture hours of use metering
BECo	0.73	Customer self-reports	On-site inspections of 18 sites
NEES EI	0.78	Customer self-reports	23 site end-use metering
NEES Sml C/I	1.02	Customer self-reports during on-site survey	21 site end-use metering
NU	0.81	Customer self-reports	30 site end-use metering
PG&E	0.85	Customer self-reports	90 site end-use metering
SDGE	0.93	Assumptions by building type	Customer self-reports
SDGE	1.18	Customer self-reports	88 site hours of use metering

Three methods were used by evaluators to obtain hours of operation information. The most sophisticated evaluations relied on data collected by light-sensitive data loggers (which record ambient light/darkness) or end-use metering (which record electricity consumption on lighting circuits) equipment. Less sophisticated evaluations used program employees to conduct on-site visits and collect information from building managers and employees. Some programs used mail or telephone surveys to obtain hours of operation information from participants.

A systematic bias in customer reports of hours of operation is suggested by the data in our sample. Site inspections, hours of operation metering and end-use metering by CMP, NEES, and PG&E found recorded hours were less than customer self-reported hours. In only two cases, NEES' Small C/I program and SDG&E's Energy Efficient Hardware program, did end-use metering uncover that customer self-reports underestimated equipment operating hours.

Our review also indicates that hours of operation used in tracking database estimates of savings should be disaggregated, at a minimum, by building type. In the six evaluations where hours of operation were logged electronically, annual hours varied by as much as 50% across building types, a much larger variation than is usually found in buildings of the same type (although in two cases, annual hours varied almost as widely across buildings of the same type because of vacancy and usage characteristics).[4] Finally, the differences between customer self-reports and metered estimates of hours of use are fairly large; the additional cost of metering or site inspections may be warranted if the accuracy of savings estimates is a concern.

2.1.1.3.Hours of Operation Changes and Takeback

After an energy efficiency retrofit, consumers may change their behavior so as to vitiate part of the efficiency gain (Hirst 1991). Such "take back" effects can subvert some or all of the energy saved. Consolidated Edison and Central Hudson surveyed program participants; neither utility found any evidence of take back in its commercial lighting retrofit rebate programs. Seattle City Light surveyed program participants and found that operating hours had increased, after measure installation for a small number of participants. But because the increase in operating hours was not due to installation of efficient equipment, take back was not indicated. Our sample suggests that commercial lighting programs have generally not exhibited take back; lighting operation hours are unlikely to change simply because of cheaper operating costs.

2.1.2. Measured Consumption Program Savings Estimates Using End-Use Metering

Electronic meters and data-loggers to monitor energy use are effective means of measuring both energy savings and peak-demand reductions. Metering of equipment is performed both before and after measure installation. For the four programs in our sample that were metered, at NEES, NU, and PG&E, sample sizes ranged from 21 sites

[4] See also Owashi, L.D., Schiffman, D.A., Sickels, A.D., "Lighting hours of operation: Building type versus space use characteristics for the commercial sector", *Proceedings of the 1994 ACEEE Summer Study*, 8:157-162.

to 67 sites. Because all four end-use metering studies were performed by just two contractors, it comes as little surprise that similar methods were used. All four studies used spot-watt metering in tandem with metered hours of operation to determine kWh saved. Demand savings were estimated using data from the metering devices only. All four studies had meters installed for at least two weeks before and two weeks after program measures were installed.

All four metering studies were explicit in their measurement and analysis of distinct program savings parameters. Evaluation reports compared the number of measures per site, annual hours of operation, and watts saved per measure (as described in the tracking database, estimated with site inspections, and measured using end-use metering). By comparing these parameters among evaluation methods, evaluators uncovered important information about the ratio of metered savings estimates to tracking database estimates. For example, in NEES' Energy Initiative Program, on-site estimates of measures installed were 100% of tracking database estimates, metered estimates of hours of operation were 77% of tracking database estimates, and spot-watt metered estimates of the change in watts consumed per measure were 87% of tracking database estimates. Confidence intervals were also calculated around the ratios of these parameters. Parameter level information collected in these kinds of studies can be used to improve future tracking database estimates of savings.

The main drawback of end-use metering is its high cost, which usually precludes metering at every participant site. Metering is labor intensive, with multiple site visits required to install, maintain, and remove the equipment. In none of these programs was every measure sampled at every site, so another potential drawback is the biases that may result from sampling a nonrepresentative set of measures (e.g., those that are easiest to connect to data loggers) at each site.

Metering is also usually performed for a limited amount of time. Because consumption patterns vary with weather and seasons, however, metering over a limited amount of time could result in a biased estimate of savings. And because metering studies omit comparison groups, downturns or upswings in the economy are not correctly recognized as a change in participant baseline consumption, further biasing the savings estimate. Finally, metering only the newly installed lighting equipment does not enable calculation of interaction effects: changes in heating and cooling loads as a result of cooler-operating lighting. We explore the magnitude of such biases in subsequent chapters.

2.1.3. Examining Top-down Models of Annual Savings

The evaluation community uses a wide range of models which incorporate customer billing data. It would be prohibitive to test the simulation datasets on every single model ever used to evaluate an energy efficiency program. Based on the most common econometric models used in our sample of lighting program evaluations, we have selected two different types of models which use billing data to test, comparison models and regression models. Within each type, several popular variations are also tested.

Quasi-experimental designs are used when study and sample characteristics make locating an identical control group difficult. The classic quasi-experimental design types were first explicated by Campbell and Stanley[5]:

a) "One-group pre-test post-test designs" utilize program participant consumption data before and after program intervention.

b) "Static-group comparison designs" utilize program participant and nonparticipant consumption data for the period after program intervention occurred.

c) "Nonequivalent comparison group designs" utilize program participant and nonparticipant consumption data from both pre- and post-program time periods.

The first type of model is a simple comparison model. This model calculates energy savings by taking the difference of pre-program and post-program consumption, or the difference of participant and non participant consumption. The second type of model pools monthly or annual billing data for a group of participants and non participants, and regresses consumption in the current time period against several explanatory variables, including building size, hours of operation, and cooling and heating degree days. The variations we examine are listed in Table 2-2.

Table 2-2. Summary of Comparison and Regression Models Using Billing Data

Model	Pre-program Participant Data?	Post-Program Participant Data?	Pre-Program Non-participant Data?	Post-Program Non-participant Data?	Indicator Variable
Time-series comparison	✘	✘			N/A
Cross-section comparison		✘		✘	N/A
Time-series, cross-section comparison	✘	✘	✘	✘	N/A
Time-series regression	✘	✘			Post-program (0/1)
Cross-section regression		✘		✘	Participant (0/1)
Time-series, cross-section regression	✘	✘	✘	✘	Participant x Post-program (0/1)
SAE regression	✘	✘	✘	✘	Engineering Estimate of Savings

5 We briefly describe the evaluation models here. For more information see the sources of these descriptions: Campbell, D.T., Stanley, J.C., Experimental and Quasi-Experimental Designs for Research, Houghton Mifflin, Palo Alto, 1963., and Impact Evaluation of Demand-Side Management Programs, Electric Power Research Institute, Palo Alto, CA, EPRI-7179, v.1, February 1991.

2.1.3.1.Time-Series Comparison of Consumption Data

This is a one-group pretest-posttest design, requiring collection of program participants' energy consumption data both before and after program participation. As shown below, the post-program (t_1) consumption of participants (Q_P) is subtracted from their pre-program consumption (t_0) to obtain a savings estimate. Such an analysis can be misleading unless the consumption data is normalized for exogenous factors such as weather. Use of a normalization method to adjust for different weather conditions can improve the accuracy of the resulting estimate. However, changes in energy consumption due to price effects and naturally occurring conservation are not controlled.

$$NetSavings = Q_P(t_0) - Q_P(t_1)$$

2.1.3.2.Cross-Section Comparison of Post-Program Consumption

This method is a static-group comparison that compares mean consumption of the participants to the mean consumption of a control group during the post-program period. Collecting post-program consumption data from both participants and non-participants (Q_{NP}) eliminates difficulties associated with weather and price variations (assuming participants and non-participants both experience identical weather and billing conditions). However, this method assumes there were no differences in pre-program consumption between participants and non-participants and that there are no differences in the ways that participants and non-participants respond to changes in weather, fuel prices, and other factors.

$$NetSavings = Q_{NP}(t_1) - Q_P(t_1)$$

2.1.3.3.Time-Series, Cross-Section Comparison of Consumption Data

This is a two-group pretest-posttest design, using participant and nonparticipant consumption data from before and after the program intervention. The method attempts to control for non-program factors which affect energy consumption by including non participant data. This method is usually more accurate than a post-program cross section comparison, because it includes information on changes in consumption over time, and can therefore adjust for trends in consumption. The method, like other cross-section comparison methods, assumes that, aside from program participation, the consumption patterns of participants and non-participants are similar.

$$NetSavings = [Q_P(t_0) - Q_P(t_1)] - [Q_{NP}(t_0) - Q_{NP}(t_1)]$$

2.1.3.4.Time-Series Regression

A variant of the time-series comparison method involves collecting demographic and structural data (such as building square footage *sqft*, hours of operation *hours*, cooling degree days *cdd*, etc.) from participants, and constructing a regression model where a dummy variable (*prepost*) is used to specify the pre-program or post-program time period. This method controls for changes in weather when cooling and/or heating degree days is used as an explanatory variable, but does not control for selection biases or energy price changes.

$$kWh = \alpha + \beta_1 sqft + \beta_2 hours + \beta_3 cdd + \beta_4 prepost + \varepsilon$$

The importance of using a comparison group in an analysis of consumption records is exemplified by the experience of BPA evaluators. The BPA Industrial Lighting Incentive program evaluation included a regression of participant characteristics against pre- and post-program energy consumption. The model was unsuccessful in detecting a program effect, which may have resulted from the model's omission of a comparison group of nonparticipants. Using a comparison group to help identify participants' savings is especially important when the energy impact is expected to be a small proportion of total consumption, as in the case of a lighting program aimed at industrial customers.

2.1.3.5.Cross-Section Regression

A regression model constructed to include cross-sectional data and a dummy variable for participation (*participant*) can control for some differences between participants and non participants if demographic and dwelling data are provided. In most cases, however, this is insufficient to control for free riders, participants who would have installed the measures in the absence of the program.[6] A logit model assessing the probability of adopting the energy conservation measure (among a control group) based on demographic and dwelling data can be incorporated into the model to adjust for free riders. A lagged dependent variable can be added to this model to include pre-program consumption data.

$$kWh = \alpha + \beta_1 sqft + \beta_2 hours + \beta_3 cdd + \beta_4 participant + \varepsilon$$

2.1.3.6.Cross-Section Time-Series Regression

For this method, a variant of the nonequivalent control group design, separate regressions are performed for participants and non-participants before and after program intervention. The means of the resulting estimates of energy consumption are then used to estimate savings due to the program. The model can include customer demographic and socioeconomic data in addition to billing information, so that the analysis can control for non program factors which may affect energy consumption. Alternatively, the evaluator can use a pooled time-series cross-section regression that includes all groups and time periods in one equation, with a dummy variable equal to the product of the time period and participation variables of the previous two models:

$$kWh = \alpha + \beta_1 sqft + \beta_2 hours + \beta_3 cdd + \beta_4 (prepost \times participant) + \varepsilon$$

2.1.3.7.Statistically Adjusted Engineering Analysis

Using the tracking database estimate or some other, more improved estimate of savings in place of a dummy variable has come to be described as a Statistically Adjusted

[6] Train has asserted that a comparison group which properly controls for free ridership among participants is quite difficult to construct. See Train, K.E., "Estimation of net savings from energy-conservation programs", *Energy*, 19(4):423-441, 1994.

Engineering (SAE) analysis.[7] SAE models also include a lagged dependent variable, representing the electricity consumption in the previous period. If nonparticipant data are used to construct a cross-section SAE model, the Savings Estimate variable is zero for nonparticipants. The coefficient on the explanatory variable representing savings can be interpreted as a 'realization rate': the fraction of the savings estimate verified using customer and billing data. A possible SAE model specification is given below.

$$kWh_{t=1} = \alpha + \beta_1 sqft + \beta_2 hours + \beta_3 cdd + \beta_4 (Savings\ Estimate) + \beta_5 kWh_{t=0} + \varepsilon$$

Estimates obtained using SAE models ranged from 0.53 for NEES' Energy Initiative program to 1.05 for ConEdison's C/I Efficient Lighting program. A possible reason for the variation in SAE-obtained ratios of measured consumption savings to tracking database estimates is the differing origins of the elements within the tracking database estimates. For example, NEES used a tracking database estimate based only on rated equipment efficiencies and estimated hours of use. ConEd adjusted its tracking database estimate based on a survey of customers collecting information on hours of operation, take back, and free riders. Differences in sample size, duration of pre/post data used, and other explanatory variables used in each model also have an impact on each model's results.

Table 2-3 summarizes the methods used by the evaluations in our sample along with some characteristics of each model. Neither tracking estimates nor first-year post-program estimates of savings can verify the long-term persistence of program savings over the manufacturer estimates of measure lifetimes. Renovations, building demolition, and equipment failure all reduce the effective measure lifetime. Repeated site visits or billing analyses are required to continually verify savings over the lifetime of the efficient equipment. Not surprisingly, none of the utilities in our sample have performed studies which address the long-term persistence of program savings.[8]

Application-specific considerations may also affect the persistence of savings for reasons that have little to do with the equipment installed. Several recent studies suggest that energy efficiency measures may sometimes be removed from service through remodeling or demolition prior to the end of their useful lives (Skumatz 1993, Petersen 1990, Velcenbach 1993). The probability of premature retirement of equipment is a function of both general economic conditions as well as site-specific considerations (for example, building and business type).

[7] Train, K.E., "An assessment of the accuracy of statistically adjusted engineering (SAE) models of end-use load curves", *Energy*, 17(7), pp. 713-723, 1992.

[8] Utility DSM programs and DSM program evaluation are too young to have long-term studies of persistence; measures from the earliest large-scale DSM programs (from the early 80's) are just reaching the end of their manufacturers' rated lifetimes.

Table 2-3. Summary of Evaluation Methods Based on Billing Data

Utility	*Type of Difference or Regression Model Used*	*Comparison Group*	*Sample Size (total part.)*	*Notes (time-series data used, sample stratification, etc.)*
BECo	$\Delta Consumption_{part.}$ - $\Delta Consumption_{nonpart.}$	Eligible nonparticipants	772 (919) part. 5826 nonpart.	12 mos. pre, 8 mos. post; 10 strata based on size and seasonal usage
CHG&E	SAE, facility type, bldg. characteristics. vars., 2 tracking estimate vars.	Eligible nonparticipants	54 (606) part. 116 nonpart.	4-5 mos. pre, 4-5 mos. post; verified hours w/ customer surveys
Con Edison	SAE, facility type vars.	Eligible nonpart. and soon to be participants	n/a (2,276) part. n/a nonpart.	4 mos. pre, 4 mos. post; verified hours w/ customer surveys
NEES EI	SAE, self-selection var., bldg. characteristics vars, 1 tracking estimate var.	Eligible nonparticipants	369(4,114) part. 611 nonpart.	12 mos. pre, 12 mos. post
NEES Sml C/I	$\Delta Consumption_{part.}$; adjusted for nonparticipants	Eligible nonparticipants	831(2,494) part. 698 nonpart.	12 mos. pre, 12 mos. post
NU	SAE, self-selection var., facility type vars., 1 tracking estimate var.	Eligible nonparticipants	1,123(5,967) part. 1,271 nonpart.	5 mos. pre, 5 mos. post; 7 strata based on size; weather adjusted kWh
PEPCO	Pooled cross-section regression, self-selection var.	Eligible nonparticipants	341 (345) part. 1,452 nonpart.	12 mos. pre, 12 mos. post; 4 strata based on size; weather adjusted kWh
SCL	$\Delta Consumption_{part.}$ - $\Delta Consumption_{nonpart.}$	Eligible nonparticipants	118 (128) part. 229 nonpart.	12 mos. pre, 12-36 mos. post
PG&E	SAE, self-selection var., bldg. characteristics vars., 1 tracking estimate var.	Eligible nonparticipants	724(6,432) part. 370 nonpart.	12 mos. pre, 12 mos. post
SDG&E	CDA, 12 end-use vars.	None	181(789) part.	12 mos. pre, 12 mos. post; adjusted model based on end-use metering results

Notes: *facility type vars*: dummy variables used to indicate the type of facility (office, retail, school, etc.), *building characteristics vars*: variables used to indicate changes in floor space, participation in other DSM, recent renovation, upswing in business, etc., *self-selection var.*: variable obtained from a logit model and used to adjust for self-selection bias, *tracking estimate var.*: variable used to indicate the tracking estimate of savings for each customer, *pre/post:* refers to the numbers of months of billing data compiled before and after program measures were installed.

Current estimates of savings are often based on the assumption that equipment will operate for the duration of the manufacturers' estimates of the equipment's useful life.[9] Assumed measure lifetime varied widely for identical measures from program to program in our sample. In some programs, lifetimes were based only on manufacturers' estimates of product longevity. In a few cases, estimates were adjusted downwards to account some for premature retirement resulting from the predicted frequency of building renovations. Several utilities (CMP, NEES, SCL) used site inspections and bill analyses to estimate savings persistence one, two, and three years after installation; in no cases, however, were measure life estimates based on a complete longitudinal set of data from past program participants. The average measure life used to calculate program savings for each program in our sample is given in Table 2-4. In cases where our original estimate of measure life did not come from the utility, it was subsequently verified by a utility representative.

Examining billing data over several years can provide an estimate of overall savings persistence. NEES evaluators used billing analyses to verify savings persistence over a two-year period. SCL evaluators used comparisons of participant and nonparticipant billing data to estimate savings persistence over a three-year period. While NEES found almost 100% persistence, SCL found a gradual degradation of savings: where approximately 95% and 88% of original savings remained after two and three years, respectively. The cause of such a degradation, however, is not limited to measure removal. Degradation of savings as evidenced by a billing comparison could be the result of increases in nonparticipants' equipment efficiency, poor maintenance of measures, or increased consumption resulting from take-back.

[9] Alternatively, for the ASHRAE or AHAM estimate of measure life.

Table 2-4. Summary of Measure Life Estimates Used to Calculate Lifetime Savings

Utility	*Measure Life Estimate* (years)	*Source of Estimate*[10]
BECo	15.0	IRT report[11]
BHEC	10.0	utility report[12]
BPA	15.0	utility report
CHG&E	10.0	utility contact
CMP	7.0	utility report
Con Edison	11.0	utility contact
GMP Small	14.7	utility report
GMP Large	6.1	utility report
IE	12.0	utility report
NEES EI	18.0	Nordax database[13]
NEES Small C/I	15.0	Nordax database
NMPC	13.0	utility contact
NU	17.0	utility contact
NYSEG	10.0	utility contact
PEPCO	9.5	utility contact
PG&E	15.9	utility report
SCL	12.9	utility report
SCE	16.0	utility report
SDG&E	15.0	IRT report
SMUD	5.0	utility contact

2.2. Summary of Bias and Precision in DSM Evaluation

Table 2-5 summarizes the ways in which each evaluation method can introduce bias or imprecision into an estimate of annual savings for a program distributing commercial lighting equipment. The forthcoming chapters in this report investigate the magnitude of these effects using data from previous evaluation studies as well as simulation techniques. The long-term goal of such an analysis is to improve the characterizations of bias and precision to such an extent that the evaluation needs of all programs are reduced: only those parameters which have been found to induce the worst bias and imprecision are investigated in the course of an evaluation. Because the state of current practice limits our sample of available evaluations to a few handfuls, we conduct as thorough a characterization as the data allows. Ultimately, we hope others will continue these efforts as more evaluations with the requisite data are conducted (and reported on) by utilities.

[10] All measure life estimates, regardless of original source, have been verified with utility representatives.

[11] IRT report: program summary sheet from the Results Center Aspen, CO.

[12] Utility report: evaluation report from utility.

[13] Nordax database: data from the Northeast Region DSM Data Exchange.

Table 2-5. Summary of Annual Savings Evaluation Methods Examined

Method to estimate annual savings	Effects treated/accounted for	Primary accuracy limitations	Potential bias in annual savings estimate due to:	Potential imprecision in annual savings estimate due to:
Tracking database (engineering estimate)		baseline equipment, usage patterns, equipment installations not verified, efficiencies from mfr. specifications, requires gross assumptions regarding consistent customer behavior	over/underestimation of baseline and program equipment efficiencies, hours of operation	Precision not estimated
Site inspection	baseline equipment (with pre-installation inspections) and efficient equipment specification errors in tracking database, hours of operation (from auditor/ customer survey)	still simplifies equipment usage patterns, does not verify equipment energy consumption at customer sites	over/underestimation of operating hours or equipment efficiencies by auditors/ in customer surveys	Precision not estimated
End-use metering	variations in equipment usage, baseline usage (if pre/post metering)	metered sample may not accurately represent population, metering of limited time duration, no comparison group	seasonal variations in equipment usage, HVAC/lighting interaction effects, unrepresentative sample of customers/equipment/building zones metered	limited duration metering, extrapolation from sample to population
Customer bill-based econometric models	changes in equipment usage, changes in weather, changes in baseline energy use (with comparison group)	provides little understanding of program strengths/weaknesses or justification for its savings estimate, requires one year of post-program data	non-normality of data/error term, improper model specification, improper comparison group, inadequate variability in data, low signal/noise ratio	improper model specification, inadequate variability in data, low signal/noise ratio

Using Simulation Techniques to Assess Performance of Tracking Database and Site Inspection Evaluation Methods

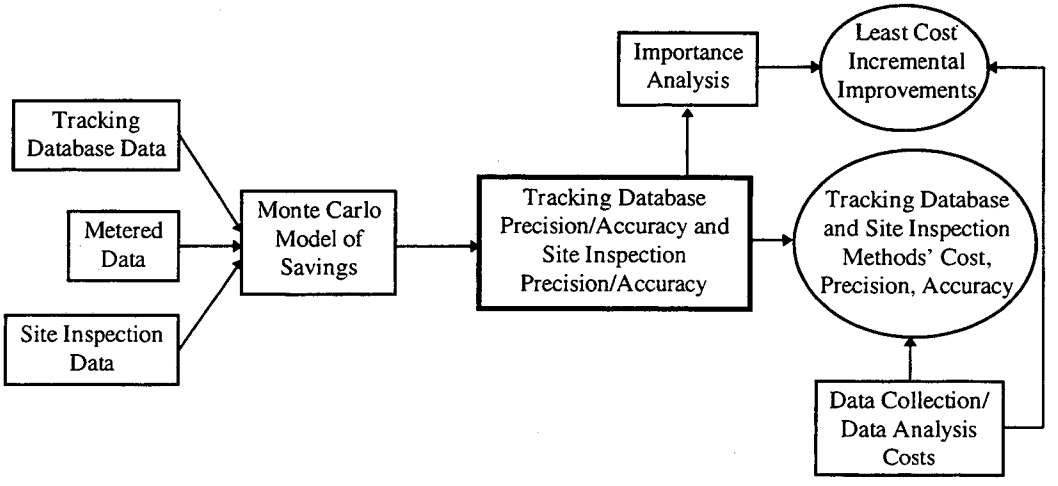
In this chapter we examine 'bottom-up' evaluation methods: tracking database and site inspection methods which utilize information on the exact types and quantity of equipment installed at each participant's site. Using a combination of simulation techniques and actual program and customer data, we investigate the importance of the different parameters used to estimate energy savings, and the accuracy of current methods. In the following chapter, we examine bottom-up methods incorporating end-use metering data.

Bottom-up tracking database and site inspection methods are attractive because they involve minimal additional data collection, utilizing existing data collected during the program's audit and/or application process. However, the correctness of this data, and of additional calculations based on equipment manufacturers' estimates of equipment operation are not well understood. Knowledge of the uncertainties pertinent to bottom-up evaluation will enable evaluators to improve evaluation results, and make better decisions regarding the evaluation method selection.

The chapter begins with an analysis of detailed information from three commercial lighting programs. Visual display of the quantitative information is used to characterize the uncertainty of tracking database and site inspection estimates of savings. Our limited data on tracking database and site inspection methods is combined with estimates of data collection and data analysis costs in order to compare the costs and performance of the different methods.

A flowchart for the analysis described in this chapter is given in Figure 3-1.

Figure 3-1. Analysis of Bottom-up Evaluation Methods



3.1 Tracking Database and Site Inspection Estimates: Unbundling the Realization Rate

For this analysis, we use 75 customer sites of end-use metered data from three commercial lighting retrofit programs: New England Electric System's (NEES) 1991 Energy Initiative and Small C&I programs, and Northeast Utilities' (NU) 1991 Energy Saver Lighting Rebate program.[1] All three programs provided rebates for commercial customers in the Northeastern United States who replaced less efficient interior lighting equipment with more efficient alternatives. Because all three programs were evaluated using end-use metering, and because site inspection and tracking database estimates were also available, we can use the evaluation data to explicitly compare the accuracy of the different methods.

However, the conclusions stemming from an analysis of these three programs' data cannot necessarily be extrapolated to all commercial lighting DSM programs, let alone all DSM programs. Because different populations can have different characteristics which affect variability of energy consumption and the accuracy of different evaluation methods, a much larger sample of programs would be required before more general conclusions could be drawn regarding the accuracy of various evaluation methods. In future research, metering information from a much larger sample of programs could be analyzed using the framework described here. Such an analysis would produce a more generalizable result regarding the accuracy and precision of tracking database, site inspection, and end-use metering estimates of annual energy savings, and could correlate program and tracking database characteristics with method performance.

Some evaluation analysts calculate the ratio of their final savings estimate based on extensive *ex post* evaluation to their tracking database estimates of savings, and refer to this ratio as the "realization rate". In 1991, Nadel and Keating compiled results from more than 40 DSM program evaluations, and pointed out that this ratio often diverges considerably from one, the tracking database estimate of savings usually being larger than the final savings estimate.[2] Some DSM analysts have taken this to mean that engineering estimates of savings are useless, and should be discarded. Such a conclusion is premature. A more thorough characterization of bottom-up evaluation methods is necessary before one can dismiss engineering estimates entirely. The realization rates for the three program evaluations from the Northeast are given in Table 3-1.

---

[1] RLW Analytics, Inc. and The Fleming Group. 1992. Energy Saver Lighting Rebate: Results of the 30-Site Short Duration Monitoring Test, C&LM Department, Northeast Utilities, Westbrook, CT.

RLW Analytics, Inc. and The Fleming Group. 1992. Small C/I Program: Impact Evaluation Using Short-Duration Metering, New England Electric System, Westborough, MA.

RLW Analytics, Inc. and The Fleming Group. 1992. New England Power Service Company Energy Initiative Program: Impact Evaluation Using Short-Duration Metering, New England Electric System, Westborough, MA.

[2] Nadel, S.M. and K.M. Keating. 1991. "Engineering Estimates vs. Impact Evaluation Results: How do they Compare and Why?" Proceedings from the 1991 Energy Program Evaluation Conference, pp. 24-33. Chicago, IL.

Table 3-1. Program realization rates for gross annual savings

Program	Realization Rate
NEES Small C&I	.88
NEES EI	.70
NU ESLR	.87

The data in Table 3-1 indicates that 70-88% of the tracking database estimates of savings were verified by the end-use metering estimate. This is consistent with the assertion that a tracking database usually overestimates actual savings. While previous studies consider realization rates as an end result, these realization rates are the starting point for this more detailed investigation. Forthcoming sections examine the annual savings equation (given in 2.1.1) parameter values' uncertainty in order to understand what the realization rates represent, and how the tracking database estimate can be cost-effectively improved.

In order to compare the results of different methods, we unbundle the realization rates in several dimensions:

- We examine not only the realization rates which represent the differences between metering and tracking database estimates, but also between metering and site inspection estimates of savings.
- We examine the realization rates for three components of the savings estimate:
    1. Measures per site
    2. Hours of operation
    3. Watts saved per measure

Table 3-2 provides estimates of measures installed per site, hours of operation, and watts per measure, obtained using end-use metering and site inspections, for several programs.[3] Multiplying the parameter level realization rates yields the aggregate realization rates presented in Table 3-1. The numbers presented in Table 3-2 are expressed as a ratio of the parameter value obtained using metering to the value in the program's tracking database. For example, the tracking database underestimated the number of measures per site, on average, by 3% for NEES' Small C&I program.

[3] The data in Table 3-2 are not weighted by the number of measures per site because we did not have access to sufficient information to perform such a weighting. Thus, the results we present deviate slightly from the original evaluation studies.

Table 3-2. Comparison of parameter values from different evaluation methods

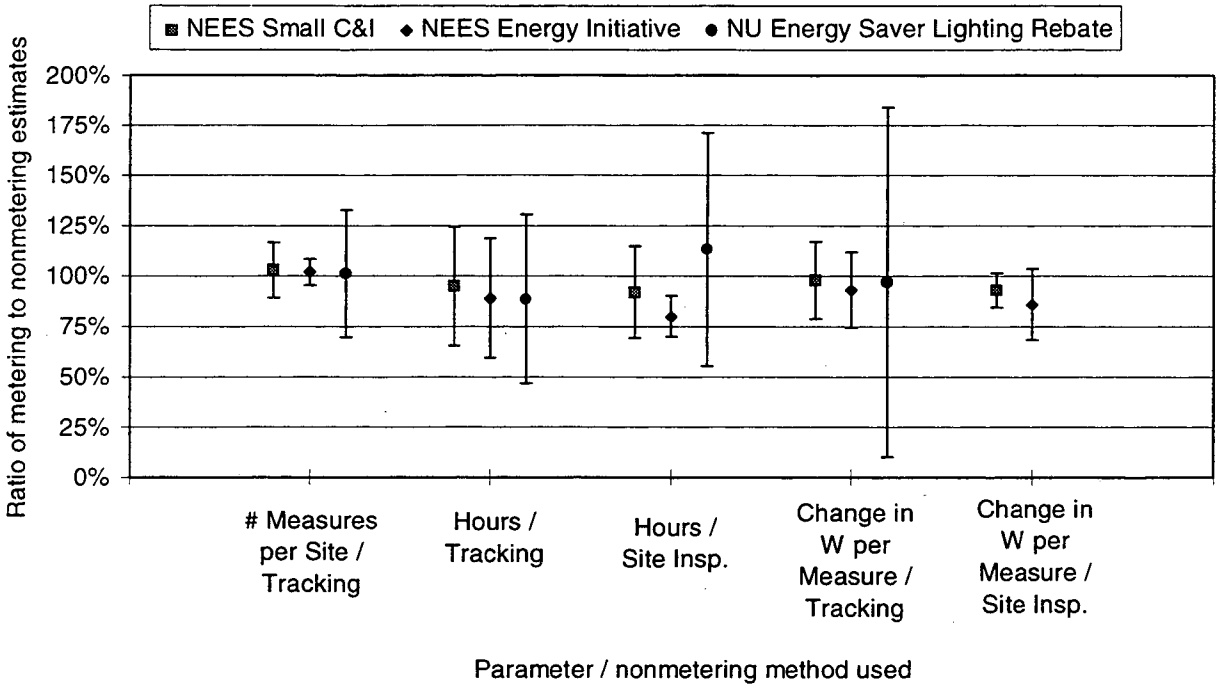
	# Sites Metered	Measures per Site	Hours of Operation		Watts Saved per Measure	
Ratio of Metered Estimate to:		*Tracking*	*Tracking*	*Site Insp*	*Tracking*	*Site Insp*
NEES Small C&I	21	103%	92%	96%	93%	96%
NEES Energy Initiative	23	102%	80%	89%	86%	93%
NU Energy Saver Lighting Rebate	30	101%	89%	113%	97%	91%

While the number of measures installed per site are underestimated slightly in the tracking database, the tracking database overestimates every other parameter. All but one parameter (hours of operation in the Energy Saver Lighting Rebate program) is overestimated through site inspections. The tracking database overestimates the actual savings per site by overestimating the individual parameter values used in the equation to calculate savings. The parameter value ratios in Table 3-2 are more informative than the aggregate realization rates.. A glance at Table 3-2 can inform the analyst that systematic overestimation of hours of operation, and Watts saved/measure, but not measures installed per site, are the largest contributors to inflated tracking database estimate of savings..

Even though the parameter values in Table 3-2 suggest the existence of a systematic bias in the tracking estimates, it is equally important to examine the variability of this bias. This is different than simply examining the variability in a single parameter, such as hours of operation, across sites. Here we are interested in the variability of the *ratio of* tracking database estimates and metered estimates (or site-inspection estimates) for a parameter. A small variability would indicate that a simple adjustment of the parameters in the tracking database could dramatically improve tracking database accuracy and subsequent estimates of savings. But a large variability would suggest that important, extraneous factors could be missing from the parameter values used in the tracking database, requiring more caution than simply using a scalar adjustment to improve the estimate.

Examining the ratio of savings estimates for the NEES and NU programs in our sample reveals significant variability in the realization rates across the three programs' 75 customer sites. We illustrate this variability in Figure 3-2 by plotting the ratio of metered parameter values to tracking parameter values, and of metered parameter values to site inspection parameter values, along with each ratio's standard deviation.

Figure 3-2. Differences between parameter values



For many of the realization rates, the systematic bias described in Table 3-2 is framed by a much more significant stochastic component, as illustrated in Figure 3-2. The large variability in parameter values obtained with different methods may mean that the parameter values used in the tracking database, while fairly accurate on average, are inaccurate for a large number of individual sites and/or measures. The greater the stochastic component, the more difficult it is to generalize from the metered sites to a larger sample of participants, and the more difficult it becomes to systematically correct for error by adjusting tracking database estimates.

The value of expanding the realization rate and presenting the results graphically is especially clear when the results of the NU program are examined. In comparison to the other two programs, the wide variations between site inspection and metered estimates of hours of operation, and between tracking and metered estimates of the change in watts per measure indicate problems with the tracking database. Indeed, evaluators found inaccuracies in the tracking database algorithms used to calculate the change in watts for optical reflector retrofits and metal halide retrofits. These errors in the tracking database calculations explain the large standard deviation for the change in watts parameter: savings from metal halide retrofits were systematically underestimated and savings from optical reflector retrofits were systematically overestimated, creating a wide, bimodal distribution for the hours of operation realization rate. However, the evaluators gave no reason for the discrepancy between hours of operation estimates based on site inspections, and those based on metering.

3.1.1 Uncertainty Propagation: Assessing the Accuracy and Precision of Site Inspection and Tracking Database Estimates of Savings

In the previous section, we described the variability of the parameters which comprise the realization rate. In this section, we use Monte Carlo techniques to estimate the uncertainty of savings estimates based on the uncertainty of the parameters that make up each realization rate. This type of analysis, where information regarding the variability of the inputs is used to estimate variability in the outputs, is known as *uncertainty propagation.*[4]

For this part of the analysis, we use the site inspection and metering data from NEES Small C&I and Energy Initiative programs to construct probability distributions for the number of measures per site, hours of operation, and watts per measure for the model.[5] We construct two sets of input distributions:

1. The first set of input distributions is based on the differences in parameter values obtained using *end-use metering* and those in the *tracking database*. The resulting outcome distribution expresses the extent to which savings estimates obtained using end-use metering differ from estimates in the tracking database. If end-use metering results are assumed to represent actual savings, then the outcome distribution generated here represents the degree to which tracking database savings estimates deviate from this reality.
2. The second set of input distributions is based on the differences in parameter values obtained using *end-use metering* and those obtained with *site inspections*. The outcome distribution estimated using these parameters describes the variation of site inspection estimates of savings from end-use metering estimates, and can be interpreted as the degree to which site inspection estimates of savings deviate from reality.

As a first approximation, parameters in our sample can be approximated with a normal distribution.[6] For example, a histogram of the difference between tracking database estimates of hours of operation and metered estimates of hours of operation from the NEES programs is plotted in Figure 3-3.

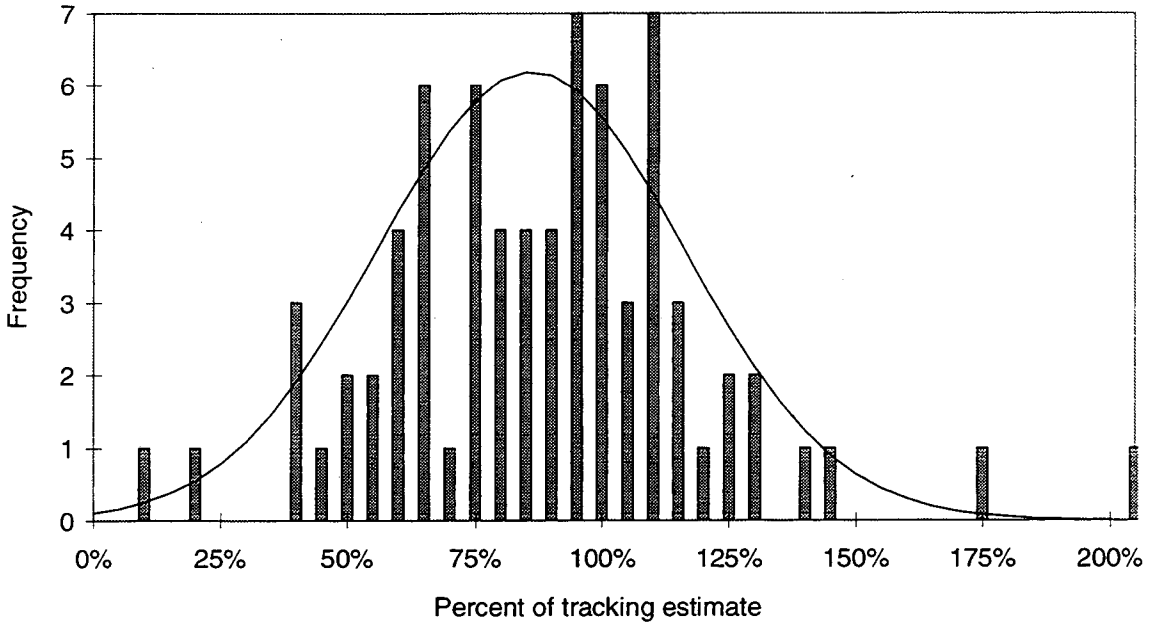
---

[4] Morgan, M.G., Henrion, M., *Uncertainty*, Cambridge University Press, 1991.

[5] We excluded the NU program from the following analysis because of unusually large systematic errors in its tracking database estimates of savings.

[6] Other distributions, such as a beta distribution, were found to have an improved fit, but did not affect significantly the outcome of the analysis.

Figure 3-3. Distribution of hours of operation realization rates



Rank transformations were performed on each set of parameter data to verify the fit of a normal distribution. The Monte Carlo analysis involved random sampling of probability distributions to model the uncertainty of errors in tracking database and site inspection estimates of savings. Latin Hypercube sampling was used to obtain 1000 sample points per set of input distributions. These sample points were then input to the annual savings equations to obtain distributions of annual savings and evaluation method error.

An analysis of the NEES and NU metering study data revealed the correlations described in Table 3-3 between the errors of the components of the tracking database. To represent this data accurately, the sampled points should also approximate errors induced through sampling beta distributions could be adjusted to approximate these correlations. This can be accomplished using the decomposition and rank correlation methods discussed by Iman and Conover.[7] However, the effect of these correlations is minor relative to the errors in tracking database parameters themselves. Thus no correlation is induced in our analysis and we assume the three parameters are uncorrelated.

Table 3-3. Correlations in errors between components of the tracking database

Correlation	Measures per Site	Hours of Use	Watts per Measure
Measures per Site	1.00		
Hours of Use	0.28	1.00	
Watts per Measure	-0.20	-0.01	1.00

[7] Iman, R.L., Conover, W.J., "A distribution-free approach to inducing rank correlation among input variables", *Communications in Statistics: Simulation and Computation*, 11(3), 311-334 (1982).

The average and standard deviations of the outcome distributions from all three sets of input distributions are given in Table 3-4. These results are an estimate of the average accuracy of the tracking database and site inspection estimates of savings for the 44 NEES customer sites.

Table 3-4. Annual Savings realization rates from Monte Carlo models

Ratio of:	End-Use Metering	End-Use Metering
to:	Tracking Estimate	Site Inspection
Average	78%	88%
Standard Dev.	34%	22%

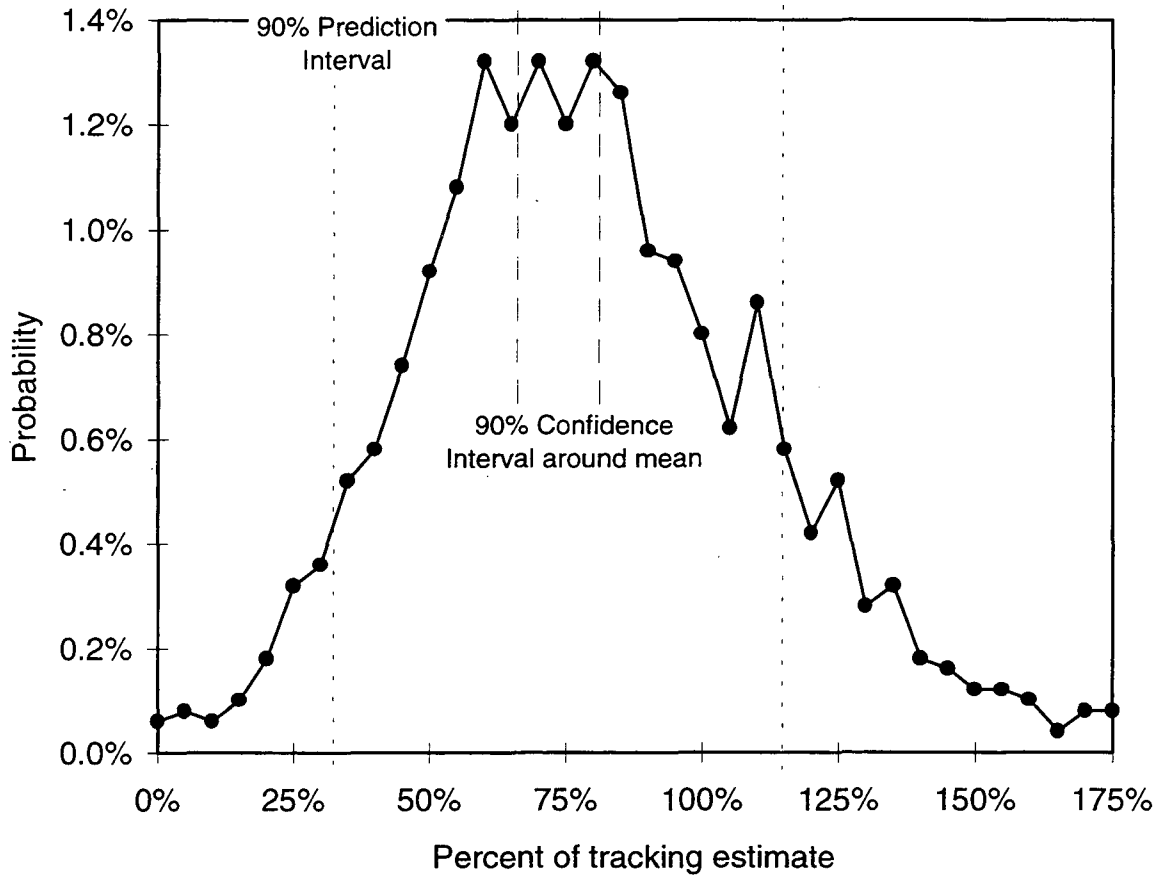
If end-use metering most closely approximates the actual energy savings for the sample, then tracking estimates of savings overestimate energy savings, on average, by approximately 22% and savings estimates based on site inspection data overestimate energy savings by approximately 12%.

While one may be tempted to conclude that the 78% figure in Table 3-4 is a transferable 'realization rate', an examination of the standard deviation associated with this estimate of bias should temper this desire. The standard deviation associated with the model's outcome distribution suggest that the tracking estimate, while biased by only 22% on average, varies considerably from site to site. The distribution of tracking estimate bias across sites, as computed by the Monte Carlo model of annual savings, is given in Figure 3-4.

If the tracking database estimate of savings closely approximated the metered estimate, the distribution shown in Figure 3-4 would be sharp and narrow (and centered near 100%), with a minimum of spread across the x-axis. But as the large standard deviation in Table 3-4 suggests, the distribution of realization rates is subject to a significant amount of uncertainty. If we assume the distribution is roughly normal, the 90% prediction interval for the distribution is between the wide margin of 34% and 122%. If one is interested in only the mean value of savings, the 90% confidence interval around the realization rate point estimate (78%) is between 70% and 86%.[8]

[8] The realization rate and its point estimate can be extrapolated to a larger population only when the sampled population is similar to it in every respect. As a result, applying an average realization rate, gleaned from a subset of participants in one program, to program participants of a subsequent or previous year, should be approached with caution. Even small differences in the characteristics of the sample population and other populations could cause relatively large differences in the average bias of the tracking estimates. Thus, without additional information it is inadvisable to cross-apply realization rates from one program to another, or, in principle, from one program year to another.

Figure 3-4. Distribution of annual savings realization rates



The smaller standard deviation for the outcome variable representing the difference between site inspection estimates of savings and end-use metering estimates indicates that site inspection provides savings estimates which are more precise than the tracking database. The 90% confidence interval around the site inspection point estimate (88%) is between 82% and 93%, considerably narrower than the prediction interval itself.

3.1.2 Uncertainty Analysis: Reducing Uncertainty in Site Inspection and Tracking Database Estimates of Savings

In this section, we compare the importance of the parameter uncertainties in terms of their relative contributions to uncertainty in the savings estimate, i.e., *uncertainty analysis*. This type of analysis reveals which parameters' values must be made more accurate in order to improve the precision of the savings estimate.

We perform uncertainty analysis by computing the rank correlation between input variables and annual energy savings for each model and examining the results. By comparing rank correlations between each input distribution and the outcome distribution, we can determine which input parameters contribute the lion's share of the uncertainty to the outcome distribution. If we then improve a single parameter's precision and compare correlations from different models, we can determine how valuable different evaluation

techniques are in reducing the relative uncertainty of program parameters. This allows program evaluators and planners to trade off evaluation method uncertainty with method cost.

If end-use metering estimates of savings are assumed to best approximate reality, then we can interpret the rank correlations in Table 3-5 to mean that most of the uncertainty in tracking estimates of savings is due to misspecifications of the hours of operation parameter, and the same parameter is responsible for most of the uncertainty in site inspection estimates of savings.

Table 3-5. Correlation of uncertainty in parameters to uncertainty in savings

Importance of Parameter Between:	Tracking Estimate and End-Use Metering	Site Inspection and End-Use Metering
Measures per Site	0.26	—
Watts per Measure	0.48	0.59
Hours of Operation	0.82	0.78

An important issue for evaluation practice involves the question of whether and when to use more rigorous evaluation techniques. In this case, our analysis suggests that for both tracking estimates and site inspection estimates of savings, the estimate of hours of operation are responsible for much of the uncertainty in the final savings estimates. If data loggers, or a similar technique, provides hours of operation parameter estimates that are a significant improvement over those used in tracking estimates and site inspection estimates of savings, then augmenting tracking estimates or site inspection estimates with this improved hours of operation information could result in savings estimates comparable with those obtained using end-use metering, but at a potentially lower cost. Alternatively, disaggregating hours of operation by measure type or by building usage characteristics may improve tracking estimates of hours of operation.

The results would have been dramatically different if we had included NU's Energy Saver Lighting Rebate data in the uncertainty analysis: the systematic errors in NU's tracking algorithms would have skewed the results; most of the uncertainty in tracking database estimates of savings for the three programs would have been due to the change in watts per measure parameter. The small sample (of three programs) which we investigate here does not enable us to determine if systematic errors in tracking databases, such as those uncovered in the Northeast Utilities data, are a common occurrence.

3.2 Comparing Accuracy to the Costs of Data Collection

In this section, we integrate the previous analyses of the chapter to compare estimates of savings from tracking database and site inspection methods with their data collection costs. Our estimates of the precision and accuracy of each method's results are subject to several qualifications:

We only examine a handful of programs in this analysis. Thus, we describe each method's accuracy and precision in the context of these programs; we cannot definitively determine the accuracy and precision of each evaluation method. With a large sample of programs and the use of methods which aggregate information (including Bayesian and meta-analytic methods) one could produce a more definitive estimate of each method's abilities.

Our estimates of each method's precision are based on a finite set of recognized variabilities in the program data and methodological limitations. Other factors may affect the estimate precision which are not covered in this analysis.

Our estimates of the accuracy of tracking database and site inspection estimates of savings are based on a comparison of metered results with the results from tracking database and site inspections. Our assessment of tracking database and site inspection estimates will be affected if metered results suffer from systematic bias due to omission of interaction effects, and overestimate precision due to limited duration metering.

In order to obtain estimates of evaluation data collection and data analysis costs, we reviewed the DSM literature, and we sent a short questionnaire to five DSM evaluation practitioners. Table 3-6 lists the resulting estimates for the cost of bottom-up evaluation methods.

Table 3-6. Estimates of data collection and analysis costs for bottom-up evaluations

Estimate Type	Data Collection Costs/Site	Data Analysis Costs/Site	Economies of Scale
Tracking Database	$0 (collected from program records)	$25	No
Site Inspection	300-750	300-750	Some (1.5% reduction / 10 sites)
Site Inspection with Pre-Post Spot Watt Metering	700-750	700-750	Some (1.5% reduction / 10 sites)

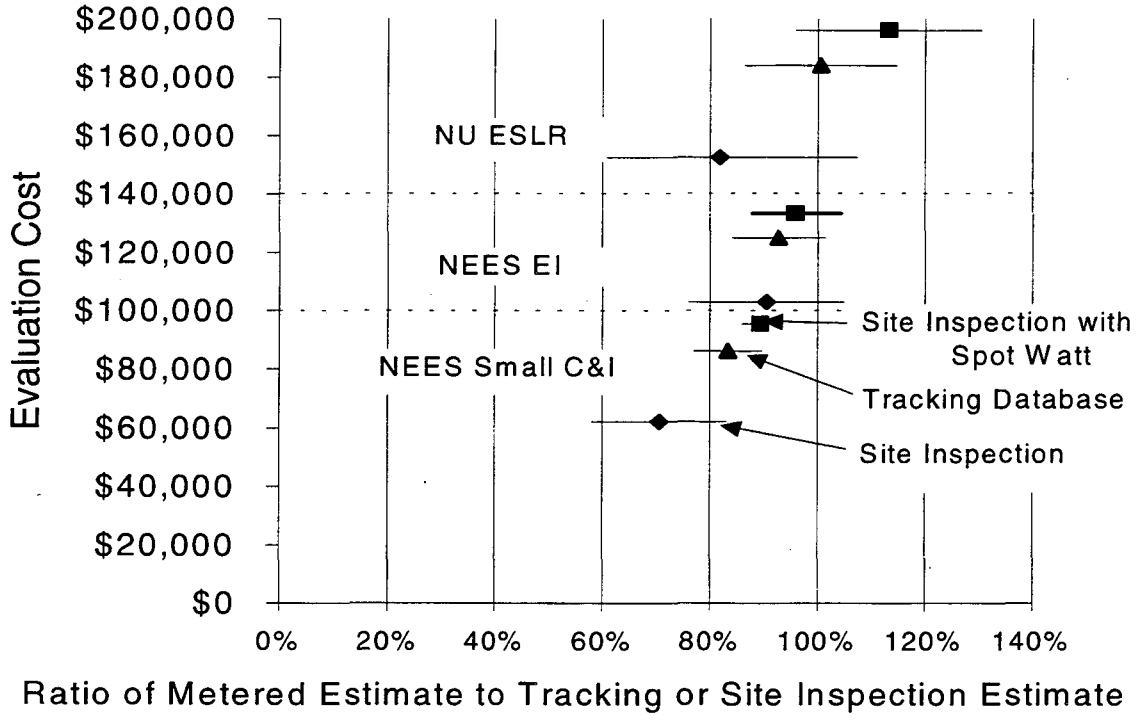
The costs given in Table 3-6 are only rough approximations, based on the judgment of several consultants who regularly conduct these evaluations. The actual costs for a specific evaluation are dependent on the types and sizes of customer buildings, the variety of measures installed by the program, and the specific monitoring equipment used. There are some economies of scale for projects which include large numbers of site visits. Other methods do not provide significant cost reductions with larger sample sizes. For the following cost/precision comparisons, we use the middle value of each range of costs in Table 3-6.

3.2.1 Costs, Accuracy, and Precision of Tracking Database and Site Inspection estimates of Savings

The first comparison of evaluation cost and accuracy we examine is for tracking database and site inspection estimates of savings. In section 3.3 we used the data from NU and NEES to estimate the accuracy (relative to metered estimates of savings) and precision of

these evaluation methods. In Figure 3-5 we display the cost and accuracy/precision of tracking database and site inspection estimates for the programs from NU and NEES. As before, accuracy is expressed as the ratio of each program's metered estimate of savings to the estimate of savings from the tracking database or site inspections (i.e., it is assumed that metering provides the actual savings). Precision is expressed as the 90% confidence interval around the mean estimate of the ratio. The graphs also include a third data point indicating the cost and accuracy/precision of an estimate of savings obtained through site inspections that include spot watt metering to verify consumption of pre-installation and efficient program equipment. Because all evaluations require a tracking database, the cost of the program tracking database is included in each evaluation methods' cost. Cost estimates vary for the three programs because sample sizes and total program size varies.

Figure 3-5. Tracking Database and Site Inspection Cost, Accuracy, and Precision



While the data in Figure 3-5 illustrate that tracking database and site inspection estimates generally do not differ from metered results by more than 40%, an evaluator calculating savings using tracking database or site inspection data has no way of knowing precisely where in this (relatively wide) range their estimate of savings falls. If we had access to a larger sample of results from other end-use metering studies, we could attempt to characterize the uncertainty in tracking database and site inspection estimates more completely, which could aid evaluators in estimating the accuracy and precision of their tracking database estimates without the use of additional evaluation. Without such a

characterization, evaluators must rely upon metered results and top-down estimates of savings to locate their tracking database and site inspection estimates in this range.

Earlier in the chapter we used importance analysis to deduce that uncertainty in hours of operation contributed the largest component to the uncertainty in the tracking database and site inspection estimates of savings. The marginal cost of adding spot-watt metering to site-inspection estimates of savings is a small fraction of the marginal cost of adding run-time data loggers, which require significant additional capital and labor-related expenses. Thus, we estimate the savings associated with adding spot watt meters in this section, and reserve improved hours of operation estimates for the next chapter which focuses broadly on metering cost, accuracy and precision.

The third data point in on each graph signifies the additional cost, accuracy, and precision when site inspection estimates are augmented with improved estimates of watts saved per measure using spot watt metering equipment. Figure 3-5 indicates that for the two NEES programs, such metering appreciably improves the site inspection accuracy and precision. However, the NU program's site inspection estimate of savings did not improve. This was due to a slight negative correlation between the errors in watts saved per measure and hours of operation. In this case, substituting spot watt-verified estimates of watts saved per measure actually increased the imprecision of the savings estimate. For the two NEES programs, we can state that augmenting site inspections with spot watt measures can increase site inspection estimate precision by 10%-50%, with an additional cost of around 10%. However, this small sample does not allow us to make broad characterizations of tracking database and site inspection information value in general.

3.3 Conclusions

Within our small sample, tracking database estimates of savings vary dramatically in their accuracy and precision. We find imprecision in hours of operation to have the largest effect on the uncertainty of the resulting annual savings estimate. We also find, in our small sample, that hours of operation estimates contribute the lion's share of bias to annual savings estimates. If future studies with a larger sample of programs can confirm these findings, it would suggest additional attention should be given to inexpensive and accurate methods for improving tracking database estimates of hours of operation.

Because the precision and bias of tracking database and site inspection estimates of savings seem to vary considerably, and because an evaluator, absent additional evaluation information, has no means of estimating the accuracy and precision of their tracking database estimate, it is dubious to rely upon tracking database estimates of savings alone. A benefit of the type of analysis performed here, and of the detailed site inspection and metering work performed *and reported* in the evaluations of the programs we studied in Chapter Two, is that it allows tracking database accuracy and precision to be assessed and improved by program implementation and evaluation staff. Using the framework outlined in this chapter, analysis of a larger number of metering studies than was available to us would permit a more complete characterization of tracking database and site inspection estimates.

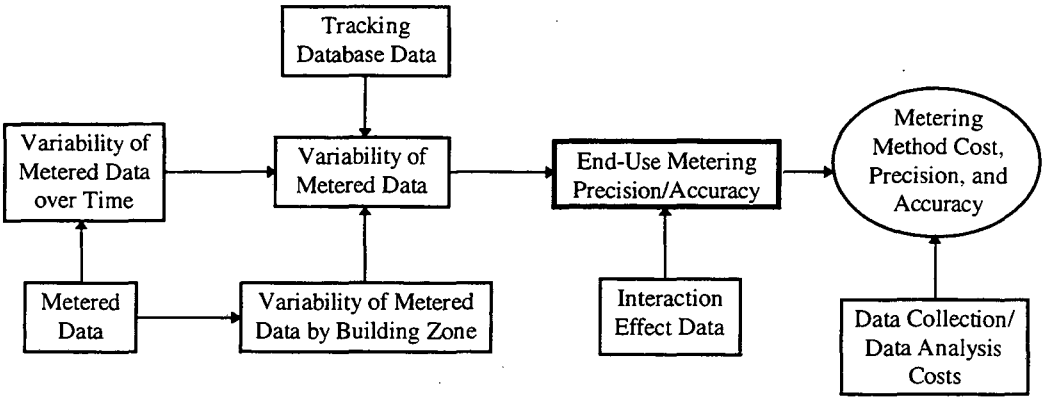
Using Simulation Techniques to Assess Performance of Bottom-Up Evaluation Methods

In this chapter we examine 'bottom-up' evaluation methods: metering methods which utilize measured information on the actual consumption and operation of equipment installed at participant sites. Using a combination of simulation techniques and actual program and customer data, we investigate the importance of the different parameters used to estimate energy savings, and the performance of current methods.

Bottom-up metering methods are useful because they involve detailed data collection on equipment installed at participant sites. However, the costs of bottom-up methods that collect extensive data are usually prohibitive, so that evaluators implement the analysis on a sample of the population. Knowledge of the uncertainties pertinent to bottom-up evaluation will enable evaluators to make better decisions about evaluation method selection.

A flowchart for the analysis described in this chapter is given in Figure 4-1. A combination of metered data, long-term hours of operation data, and simulated commercial building consumption data are used to assess the performance of annual savings estimates obtained from end-use metering data. These data are combined with estimates of data collection and data analysis costs in order to compare the costs and accuracy of the different methods.

Figure 4-1. Analysis of Bottom-up Evaluation Methods



4.1 Estimating Variability in End-Use Metering Estimates

Most metering studies incorporate information from both metering activities and tracking database or site inspection activities. The most common method in current metering studies is to express annual program savings as a ratio of the metered estimate for each site and the tracking database estimate for each site. These ratios can be averaged across

all metered sites, and the resulting average ratio is multiplied by the tracking database estimate of savings for the entire participant population to estimate total program savings.

Integrating the tracking database information with metered results reduces the annual savings estimates' tendency to be biased. Without using tracking database information, the evaluator simply assumes that the metered sample is representative of the entire sample (e.g., by stratifying and randomly selecting sites to meter from the participant population in each stratum) and can therefore apply metered results to the entire population. Instead, the evaluator uses information about the population (the tracking database estimates of savings) and information about the metered sample (the ratio of metered estimates to tracking database estimates of savings) to extrapolate the estimate of savings from the metered sample to the entire population.

A recent EPRI Report describes a related method where the average difference, rather than the ratio, of metered estimates and tracking database estimates is used to adjust the estimates of savings for every program participant.[1] One should expect a ratio approach to be appropriate when a systematic bias in the tracking database over or underestimates actual savings by a certain percentage. A difference approach would be appropriate when the tracking database values are expected to over or underestimate actual savings by a certain value. We focus on the ratio approach in our analysis because the biases we observe in hours of operation, watts saved per measure, and the number of measures installed suggest that tracking databases proportionally overestimate savings; the larger the tracking database estimate of savings, the larger the discrepancy between tracking database estimates and actual annual savings.

4.1.1 How large a population must you sample for a given level of accuracy?

Because of the relatively large per-site expense associated with end-use metering, DSM program evaluations almost never perform end-use metering on all participating customer sites. Thus, the submetering of sites is performed, and sub-subsample of measures at each selected site is metered. If one assumes optimal sample selection and stratification and a normal distribution for the population's savings, the 90% two-sided confidence interval around the mean estimate of the metered sample savings can be used to estimate savings for the entire population. The confidence interval is calculated using:

$$Precision = 1.645 \times \sqrt{\frac{\sigma^2_{sample}}{n_{sample}}}$$

Where σ_{sample} is the standard deviation of savings among metered sites and n_{sample} is the number of sites metered. If one is using an average ratio of metered estimates to tracking database estimates, the standard deviation of the ratio can used in the equation to calculate the precision. If the variation in the sample, or in the ratios, is known in advance, it is possible to back-calculate the sample size required to achieve the desired precision. However, it is difficult to assume a sample variability *a priori*; extensive knowledge of the

[1] RCG/Hagler, Bailly, *Impact Evaluation of Demand-Side Management Programs, Vol 1*, Electric Power Research Institute, CU-7179s, September, 1991.

population and the program is required. Given a large enough sample of σ_{sample}'s from past programs and populations, Bayesian updating, or other techniques incorporating prior information could be used to improve initial estimates of sample variability and estimate appropriate sample sizes.[2]

Another method to extrapolate to the entire program population takes advantage of tracking database estimates of savings for all participants as well as the metered data for a smaller segment. This method calculates the ratio of metered estimates of savings to tracking database estimates for the metered sample of participants, and then extrapolates to the larger participant population using this ratio and the tracking database estimates for all participants.[3] If one assumes that the metered results are unbiased, this method can be used to correct for bias in the tracking database estimates. In the next section, the tendency for bias to exist in the metered results is investigated.

4.1.2 Within-Site Sampling Representativeness Issues

There are profound difficulties in selecting a representative sample of sites, and of equipment at each sampled site. Convenience sampling, where the most accessible sites are selected for metering, and the most accessible equipment at each site are metered, can invalidate the ability to extrapolate from the sample to the program population using the equation presented in preceding section. How much of a threat is convenience sampling to statistical validity? One means of answering this question is to examine differences in equipment usage by building type and by different areas within buildings. A recent study performed by the consulting firm Xenergy for San Diego Gas and Electric presents detailed information on the measures selected for metering.[4]

In the Xenergy study, space in commercial buildings was partitioned into eight space use categories, called zones. Lighting loggers for measuring lighting equipment hours of operation were installed in 3,900 zones in 88 buildings. The results of these measurements, by zone, are given in Figure 4-2. In addition to hours of operation estimates obtained using lighting loggers, Figure 4-2 also includes customer estimates of hours of operation for each building zone.

Two observations can be made regarding the data summarized in Figure 4-2. First, there appears to be a significant range in metered hours of operation, which is at least partially dependent on the location of the equipment in the building. Therefore, an evaluation that only meters equipment installed in accessible locations (such as hallways) will generate an estimate of hours of operation (and consequently savings) that is biased towards equipment installed in that zone. Second, there appears to be a range of systematic biases in customer reports of hours of operation: from equipment installed in halls, where

---

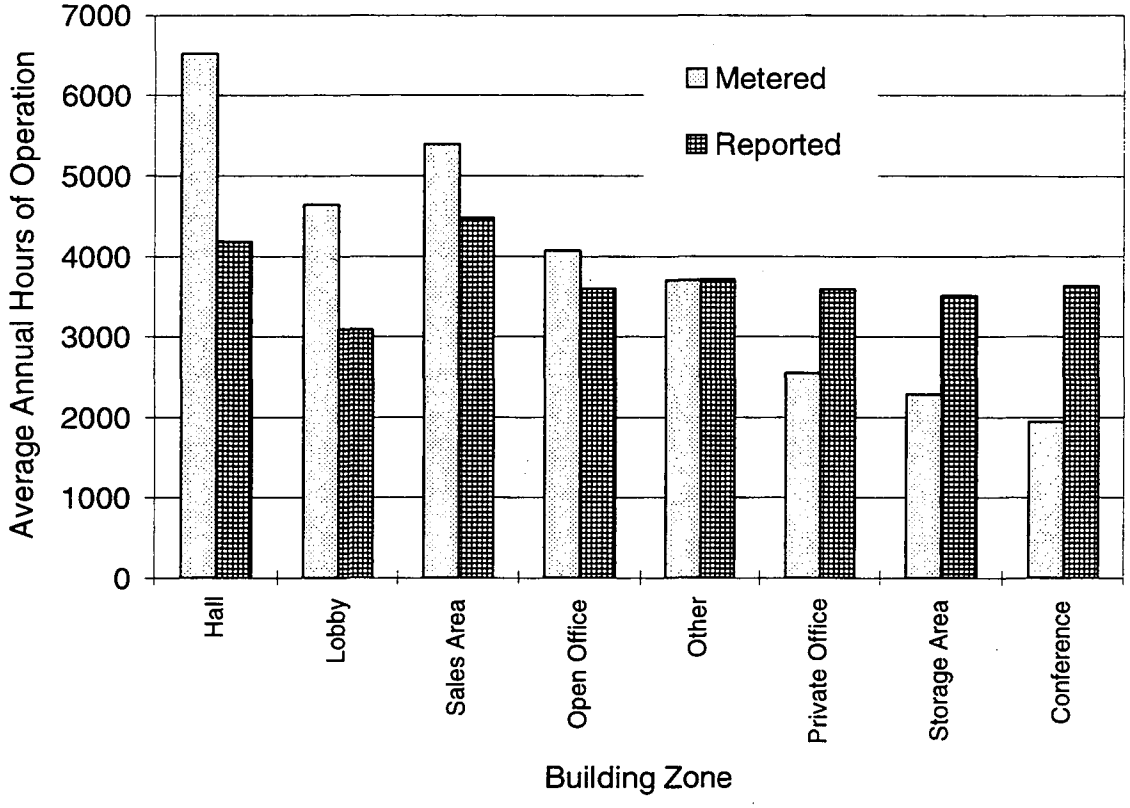
[2] See DeGroot, M.H., *Optimal Statistical Decisions,* McGraw-Hill, New York, 1970.

[3] Such a method was used by RLW Analytics in preparing the evaluation of Pacific Gas and Electric's 1992 Commercial Lighting Express Rebate Program. See PG&E, *Double Ratio Analysis Final Report*, September 1993, CIA-93-X01B.

[4] Owashi, L.D., Schiffman, D.A., Sickels, A.D., "Lighting hours of operation: Building type versus space use characteristics for the commercial sector", *Proceedings of the 1994 ACEEE Summer Study*, 8:157-162.

customers systematically underestimate hours of operation, to equipment installed in conference rooms, where customers systematically overestimate hours of operation. Thus, an evaluation that meters equipment installed in only one or two zones will generate an estimate of the ratio between reported and metered estimates of hours of operation that is probably not accurate for equipment installed in other parts of the building. A final thought regarding customer reports: customers seem to report very similar hours of operation for all zones within a building. Customers, like some evaluators, may not be aware of the differences in hours of operation in different areas of a building.

Figure 4-2. Hours of Operation Estimates by Building Zone



The hours of operation estimates in Figure 4-2 underscore the difficulty of selecting a sample to meter which adequately represents an entire population of program participants and the measures they install in different zones. Conference rooms and hallways are the most deviant, with hallways logging more than 200% more hours of operation than conference rooms. Private offices and storage areas are the most similar, with private offices' hours being approximately 10% longer than hours in storage areas. The sample must be selected to adequately represent a cross-section of participant building types, measures, and locations with buildings of measure installations. Unless the sample is representative of the entire population, the data in Figure 4-2 suggest that the resulting estimate of savings for the population could be biased by between 10% and 200%.

Similar biases could be incurred based on differences between different commercial building types (offices, hospitals, retail, etc.). Some stratification is already performed by evaluators based on building size, building type, and even space surveys of particular buildings. However, these techniques are not widespread.

4.1.3 How long should one sample for a given level of accuracy?

While sample sizes have been discussed at length in the evaluation literature, sample duration has not received similar attention. Yet the issue of how long to sample hours of operation and watts consumed (or saved) per measure should be of similar concern: In the same way that different sites are expected to have different consumption characteristics, necessitating statistical extrapolations from sampled groups to entire populations, energy consumption characteristics can also change over time, necessitating statistical adjustments based on the duration of the sampling for each site. Calculating the accuracy of varying durations of metering requires many assumptions regarding the variability of electricity consumption over time, or actual, long-term, metered data which can be used to characterize the variability in electricity consumption over time.

The Energy Edge project, a research-oriented demonstration of energy efficiency in the Pacific Northwest, provides us with a unique data set with almost all of the required characteristics: 29 commercial buildings, with hourly metering of all lighting fixtures (as well all other energy consuming equipment) for up to four years. From the total dataset, there are five small and three large office buildings. The Energy Edge office buildings for which metered data are available are described briefly in Table 4-1.

Table 4-1. Energy Edge Commercial Office Buildings

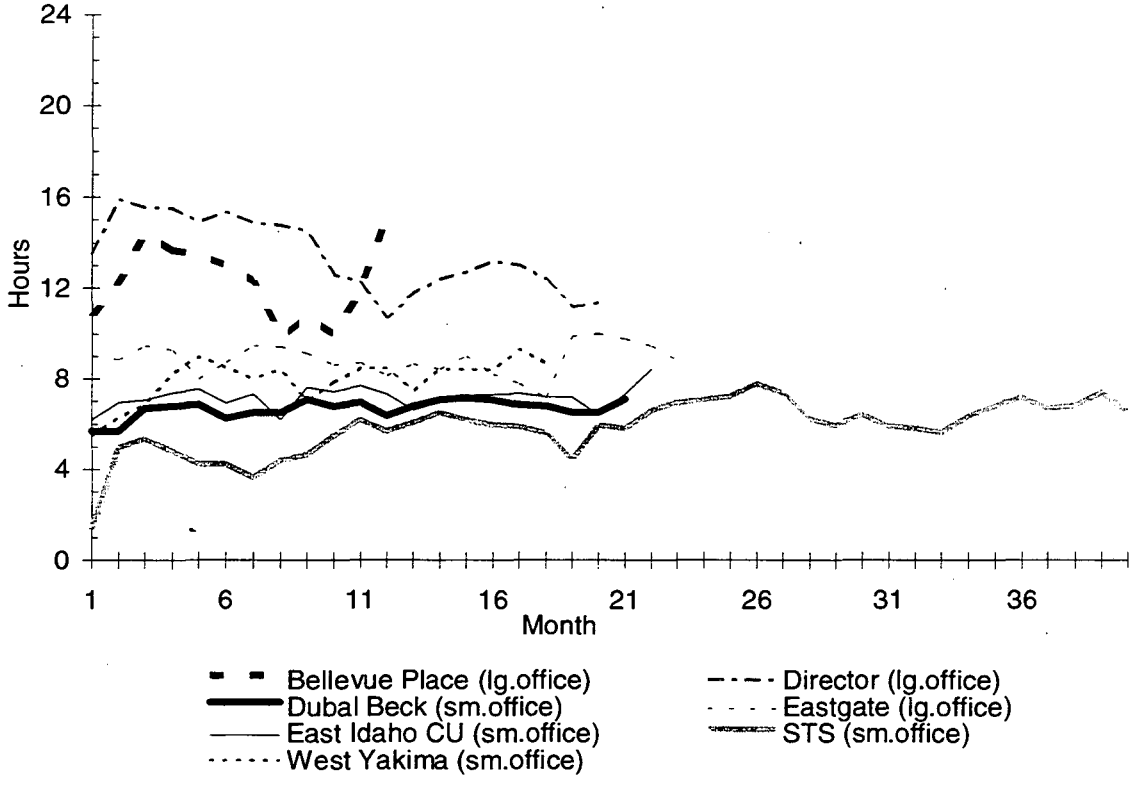
Building	Location	Size (kft^2)	Comments
Siskiyou	Ashland, OR	3.0	meters were disconnected by building manager for six months
STS	Ellensburg, WA	4.3	
East Idaho Credit Union	Idaho Falls, ID	5.3	
Dubal Beck	Portland, OR	8.5	
West Yakima	Yakima, WA	16.2	
Eastgate	Bellevue, WA	25.1	multiple tenants
Director	Portland, OR	79.7	multiple tenants
Bellevue Place	Bellevue, WA	389.0	multiple tenants

There are several methods available for computing hours of operation from hourly kW data: we could estimate the hour each day at which most of the lights turn 'on', and the hours at which most of the lights turn 'off'. Instead we have opted to estimate full load hours, which normalize lighting kW by the maximum lighting load. Thus, the full load hours calculation incorporates both lighting load and lighting duration information. The equation for daily full load hours is:

$$daily\ full\ load\ hours = \frac{\sum_{t=1}^{24}(kW_t)}{kW_{max}}$$

Where kW_t is the lighting kilowatt load in hour t, and kW_{max} is the maximum kW load for the building over the entire metered period. Weekly full load hours are the sum of 7 days' full load hours. Figure 4-3 illustrates the average full load hours over time for the eight Energy Edge buildings.

Figure 4-3. Weekday full load hours over time for Energy Edge Office Buildings



The lighting full load hours vary from site to site, as well as over time. The largest offices, Bellevue Place, Director, and Eastgate, show considerable variation over time, this variation could reflect changes in building occupancy. The buildings were all new, or recently commissioned, at the time the metering began. The majority of buildings, large and small, begin the metered period (which began on different dates for each site) with a one to three month ramp-up of average full load hours. This may be due to tenants gradually moving in over a period of several months after commissioning. Because the meters connected to the lighting circuits in the Siskiyou office building were turned off for six months, we omit this building from the analysis.

The data in Figure 4-3 suggest that hours of operation vary over time in commercial buildings. The metering activities for these buildings begun just after the buildings were commissioned, so an initial ramping-up of hours of operation can be seen for some of the buildings. Part of this variation is also seasonal, as described in section 4.1.4.

The method used to estimate the precision associated with metering the kW lighting load for a building involves subsampling from the data used to construct Figure 4-4. The subsampling strategy subsampled the full load data in continuous segments, varying in duration from two weeks to six months. By overlapping the samples, the maximum number of subsamples are taken for each duration.

To estimate the change in subsample precision, the average full load hours estimate for each subsample is compared to the long-term (utilizing all available data) full load average. For all subsamples of a given duration, the standard deviation of the difference between the subsample estimates of full load hours and the long term estimate of full load hours provides us with an estimate of the error associated with limited term metering. The standard deviation of this difference for each subsample was used to calculate the 90% upper and lower confidence levels for the range of metering durations. In order to utilize all available full load hours data, subsampling was performed on all seven of the Energy Edge office buildings, and the results were combined using a simple average.

Figure 4-4 illustrates the improvement in the precision of the short-term estimate as metering duration increases.

Because an equal number of subsamples under and over-estimate average long-term full load hours, the error across all subsamples averages about to about zero. Clearly, there is a substantial increase in precision as metering duration increases from two to four weeks. Metering longer than four weeks, however, reaps only linear increases in precision.

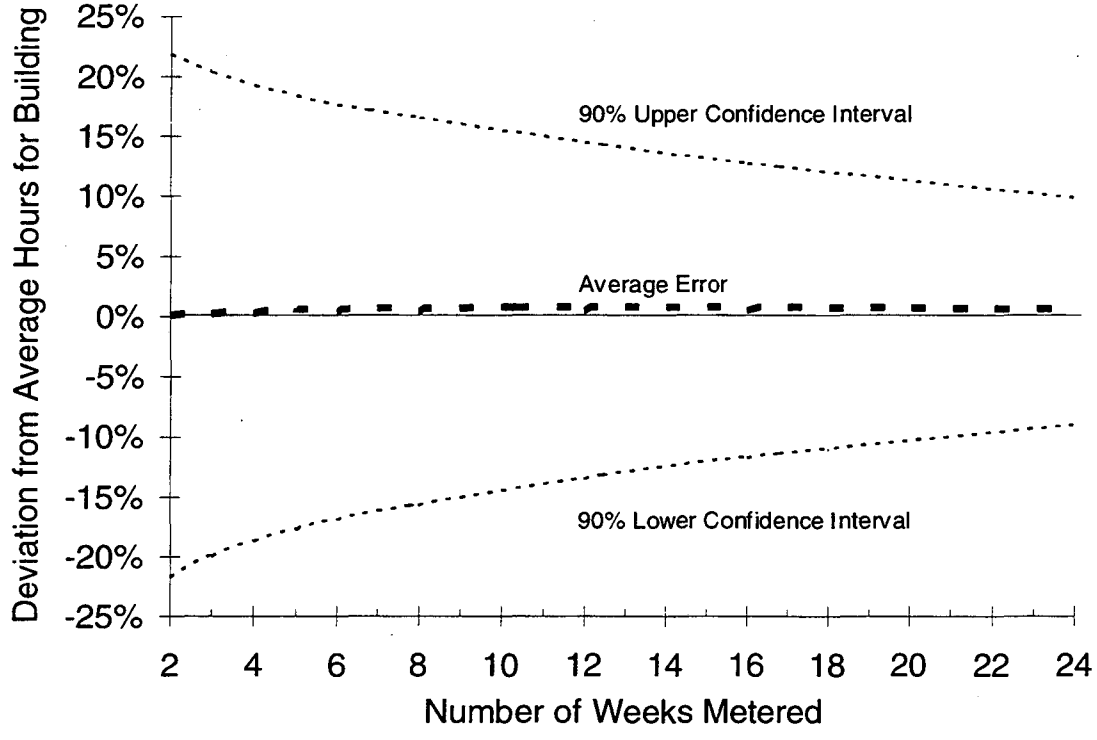
If this Energy Edge data are even partially representative of hours of operation changes over time for most buildings, then an important component of the uncertainty in most metering studies has not been given appropriate weight in past studies. Many DSM metering studies assume the hours of operation are constant, i.e., that the hours of operation measurements taken during sampling have zero variance over time. The only imprecision in metering quantified in a traditional analysis is the imprecision of extrapolating from a metered sample to the entire population of participants.

The inherent imprecision of short term metering, as expressed in Figure 4-4, directly affects the precision of any annual savings estimate based on short term metering. If the hours of operation data exhibited in Figure 4-3 are typical of commercial office buildings, then any short term metering study that omits a correction factor for metering duration will overestimate the precision of the annual savings estimate.

A correction factor can be read off the plot in Figure 4-4, based on the duration of the metering activity. This factor can be combined with the estimated precision of the metered estimate of savings. By acknowledging that the hours of operation vary over time, the actual precision of end-use metering estimates of savings are reduced. However, by estimating the increase in overall precision as the duration of subsample metering is

increased, one could explicitly tradeoff the expense and increased precision of longer metering.

Figure 4-4. Precision of estimate improves with increase in duration of metering

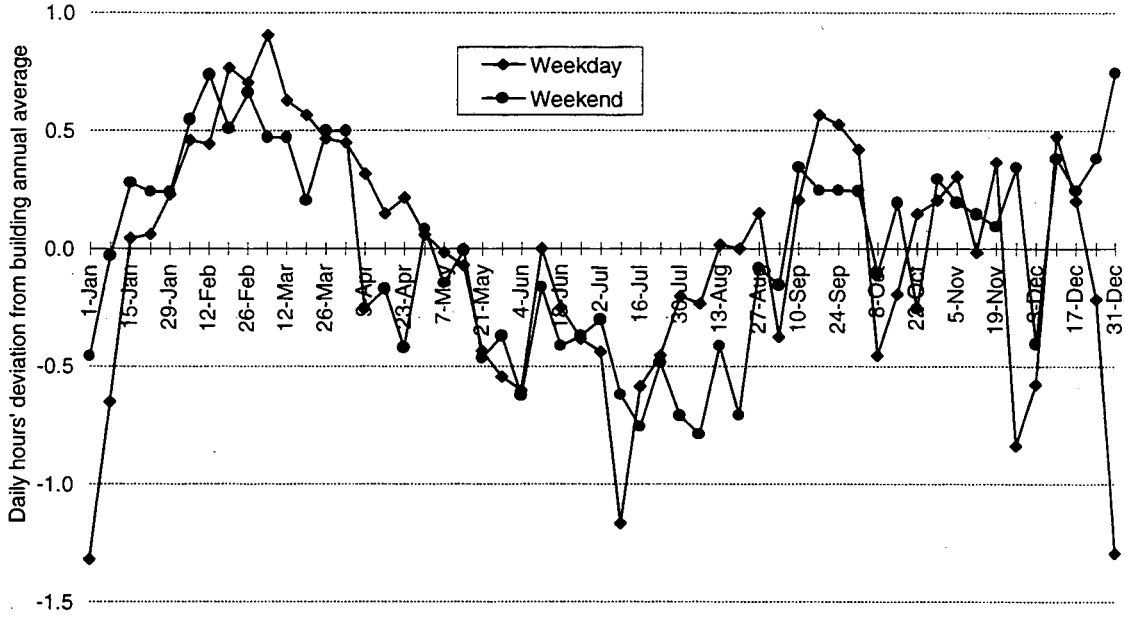


4.1.4 Bias in Metered Results: Seasonality of Hours of Operation

Another important implication of the results in this section involves an evaluator's ability to 'game' the evaluation results: If periods of greater or lesser hours of operation can be anticipated, metering can be performed during those periods when the resulting estimates of savings can be higher or lower than actual average savings. Thus, shorter duration metering can be used to generate estimates of annual savings that increase utility shareholder incentive payments or justify a program that, in reality, is not cost-effective.

We performed some simple time-series analysis of the Energy Edge hours of operation data in order to estimate the average change in mean daily hours of operation from season to season. A time-series plot of the average annual cycle of hours of operation for all the Energy Edge office buildings is given in Figure 4-5. Hours of operation are expressed as a change from the annual average.

Figure 4-5. Seasonal Variability in Hours of Operation



From Figure 4-5 it is apparent that a seasonal variation in average hours exists, but it is difficult to conclusively quantify because of the effects of holidays. Independence Day, Thanksgiving, Christmas, and other business holidays influence the average daily hours of operation of numerous work weeks each year. Filtering out these holidays results in a seasonal pattern of hours of operation where winter hours are 30 minutes longer and summer hours are 30 minutes shorter than the annual average.

In order to provide an accurate annual average hours of operation for a building or group of buildings, this seasonal bias, as well as the bias apparent in Figure 4-5 due to holidays and other disruptions to the work week, an evaluator must identify the extent of the bias and adjust annual hours of operation estimates accordingly.

4.1.5 Bias in Metered Results: Interaction of Lighting with Heating and Cooling Loads

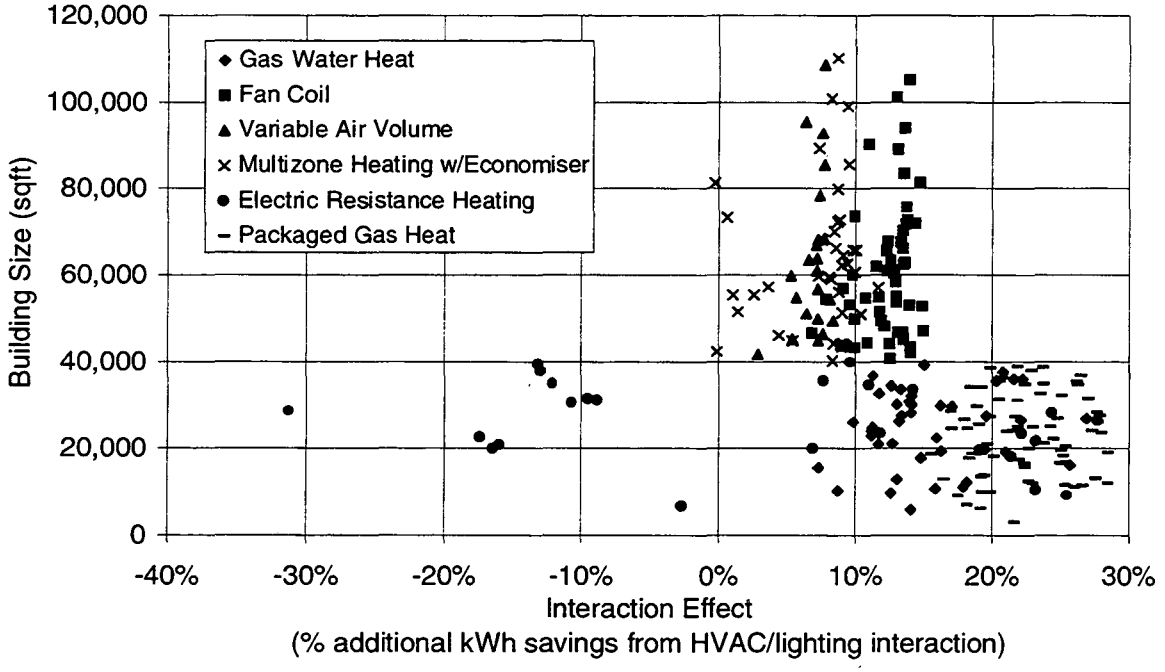
In addition to reducing lighting energy consumption, programs which install efficient lighting equipment can also affect a building's heating and cooling requirements. More efficient lighting systems generate less waste heat than standard lighting systems. In an office building, this reduction in waste heat can reduce cooling and increase heating loads. In most cases, the reduced cooling loads save more electricity than consumed by the increased heating loads (most commercial buildings do not use electricity to generate space heat). End-use metering studies for lighting programs do not typically measure changes in heating and cooling loads to supplement primary lighting savings. This omission biases metering results so that they generally underestimate program savings. In some cases, an adjustment to program savings is made across all sites to account for interaction effects. This adjustment is usually linear, and usually adds an additional 5% to 15% savings to each buildings' annual estimate. Unless based on some engineering or

metering data, such an adjustment may under or overstate actual savings due to heating/cooling-lighting interactions.

Past studies have estimated this additional energy savings, called an HVAC/Lighting interaction effect, and noted that the magnitude of the effect is dependent on building characteristics, weather, and heating and cooling equipment.[5] Thus, it is difficult to assume *a priori* a 5 or 10% increase in electricity savings, and add this to metered estimates of savings, to include interaction effects. The savings may be larger or smaller than this, depending on the location and characteristics of the participating buildings. The simulation exercise described in the next chapter uses DOE2-1E to estimate the effects of a lighting rebate program on 250 commercial buildings. We present one of those results here because it illuminates the issue of the size of the interaction effect, as determined by building size and choice of heating and cooling systems. The results are summarized in Figure 4-6.

For each heating and cooling system modeled, a definite range exists for the magnitude of the interaction effect. The widest range exists for the electrically heated buildings, while the other buildings experience increases in gas or oil consumption due to increased heating

Figure 4-6. HVAC System Determines Magnitude of Interaction Effect



loads which are not reflected in this graph. A few of the buildings incorporating electric heating experienced net increases in electricity use. Because each heating and cooling system demonstrates a different range of interaction effects, increasing total savings by

[5] Sezgen, O.A., Huang, Y.J., "Lighting/HVAC interactions and their effects on annual and peak HVAC Requirements in Commercial Buildings", *Proceedings from the 1994 ACEEE Summer Study*, 3:229-239

say, 10%, may over or underestimate the interaction effect, depending on the distribution of heating and cooling systems among participants. In the simulation, a cutoff of 40,000 square feet was used to distinguish small and large buildings, which determined available HVAC systems, and building materials and characteristics.

The limited simulation results presented here suggest that the HVAC systems used in larger buildings experience interaction effects on the order of 10%, and HVAC systems in smaller buildings experience interaction effects on the order of 20%. A comprehensive characterization of these effects would include data on the more prominent heating fuels of oil and natural gas, and compare the costs of increased heating to the savings from reduced cooling. The analysis here is meant only to illustrate that interaction effects can introduce a bias into metering results, and that the potential exists to reduce this bias using information about participating buildings and building energy consumption simulation results. If the simulation results described in Figure 4-6 are similar to the actual interaction effects, it seems that omission of all interaction effects can result in underestimating annual savings by as much as 20% and inclusion of a flat estimate of interaction effects of 16% (the average in our simulated sample of buildings) can under or overestimate actual savings by around 10%. Furthermore, this over or under estimation seems dependent on the characteristics of participating buildings. If HVAC system information were collected for each building, it may be possible to reduce the uncertainty around the size of the interaction effect.

4.1.6 Assessing the Bias and Precision of End-Use Metered Estimates of Savings

The previous sections identified three factors that affect the precision and bias of annual savings estimates obtained using end-use metering: sample representativeness, metering duration, and HVAC/lighting interaction effects. In this section we integrate these uncertainties to estimate the overall precision and bias for a number of past metering studies.

We summarize the uncertainties from each of these factors in Table 4-2. The three factors involve a combination of systematic and stochastic uncertainties, which are difficult to combine. Combining these factors is subject to some qualification because of the relatively limited sources of information on which we can base our characterization. Our limited sample prohibits a thorough assessment of each uncertainty. We also assume that the factors shaping the uncertainties are independent. If the uncertainties were correlated, the actual precision could be significantly less or more than estimated here.

We first present the results that integrate the stochastic biases into estimate of end-use metering results. Then we present the range of possible results using the information on systematic biases.

Table 4-2. Uncertainties in End-Use Metering

Uncertainty	Systematic or Stochastic Error	Potential Magnitude
Lack of sample representativeness	Systematic; metered equipment may save more or less than equipment in population	10%-300%
Limited metering duration	Stochastic; due to nonseasonal factors,	10%-20%
	Systematic; due to seasonal factors	-5%-+5%
HVAC/lighting interactions	Systematic error with surrounding uncertainty, dependent on building characteristics	116% of metered savings estimate ± 9%

The basic results from the metering studies are given in Table 4-3. We are able only to present the precision and estimated mean value for each study; an estimate of the bias is incalculable without further information on the *actual* savings for program participants.

Table 4-3. Summary of Metering Results

Study	Sample Size	Population Size	Duration of Metering	Mean estimate of savings (alt. est.)	Precision est. by eval. report (alt. est.)	Std. Dev. of Metered Sample	Estimate description (alternative estimate)
NU Energy Saver Lighting Rebate (1991)	30	6,100	2 weeks pre, 2 weeks post	0.79 of track. db.	± 54%	± 1.37	Single Ratio estimate using tracking database 90% CI around std. dev. of mean
NEES Small C&I (1991)	21	2,483	2-3 weeks, pre, 2-3 weeks post	0.96 of track. db.	± 16.7%	± 0.44	Single ratio estimate using tracking database 90% CI based on std. dev. of mean
NEES Energy Initiative (1991)	23	4114	2-3 weeks pre, 2-3 weeks post	0.677 of track. db.	± 14.5%	± 0.28	Single ratio estimate using tracking database 90% CI based on std. dev. of mean
PG&E Express (1992)	16	4,454	0-15 weeks pre, 1-15 weeks post (usu. 2-4 weeks)	1.31 of track. db. (1.07 of track. db.)	± 65% (± 39%)	± 2.0	Single ratio estimate using tracking database 90% CI based on std. dev. of mean (Dbl ratio estimate using engineering models, tracking database)
PG&E Customized (1992)	36	1,509	0-15 weeks pre, 1-15 weeks post (usu. 2-4 weeks)	0.66 of track. db. (0.75 of track. db.)	± 11% (± 12%)	± 0.25	Single ratio estimate using tracking database 90% CI based on std. dev. of mean (Dbl ratio estimate using engineering models, tracking database)

All five programs install efficient lighting equipment in commercial buildings, and were evaluated using both tracking database information and end-use metering with spot-watt meters and data loggers to collect run-time data. The studies described in Table 4-3 express program savings as a ratio of metered results to tracking database estimates for the same sites. The precision of each estimate is based on a 90% confidence interval around the standard deviation of the mean ratio values. Mean estimates of savings show that metered estimates of savings range from 66% to 131% of tracking database estimates. The precision around these mean estimates varies widely, from 11% to 65%. While the precision is affected slightly by sample size differences across the programs, the wide variations in precision are due mainly to the differences in standard deviations for the samples. The PG&E programs supplemented the evaluations with engineering models of a superset of the metered buildings, and then used this information to adjust the final estimates of savings.

Using the results of our analysis on metering duration and precision, we can adjust the estimated precision of the metered estimates to account for variability in hours of operation over time. As demonstrated in Figure 4-4, if we know the duration of the run-time metering, then we can estimate the imprecision inherent in the resulting estimate of hours of operation. We make use of addition in quadrature to propagate an estimate of the error from limited duration metering into the existing, metered estimates of savings in Table 4-3. The sum of all pre and post-installation metering in each program is used to determine the size of the adjustment in precision. The equation which we will use to combine the precisions (known as addition in quadrature) is:

$$\frac{\delta q}{|q|} = \sqrt{\left(\frac{\delta x}{x}\right)^2 + \cdots + \left(\frac{\delta z}{z}\right)^2}$$

Where q is the product of $x,\ldots,z$, δq is the uncertainty in q (expressed here as a standard deviation), and δx is the uncertainty in x. Combining the standard deviation of the metered estimate with the average standard deviation for a hours of operation estimate of a given duration, we obtain the results in Table 4-4.

We can also incorporate our estimate of potential bias due to HVAC interaction effects into the metered estimates. Data from our simulation of commercial office buildings indicate that interaction effects increase electricity savings so that total savings are 116% ± 9% of metered savings. To adjust metered estimates, then, we need only to multiply the metered estimate of savings by this 116% ± 9% adjustment factor. We again use quadrature to estimate the propagation of error through products. The results for the estimates of savings and precision adjusted for limited duration hours of operation and interaction effects are given in the final two columns of Table 4-4.

Table 4-4. Adjustments and results of adjustments to metered estimates

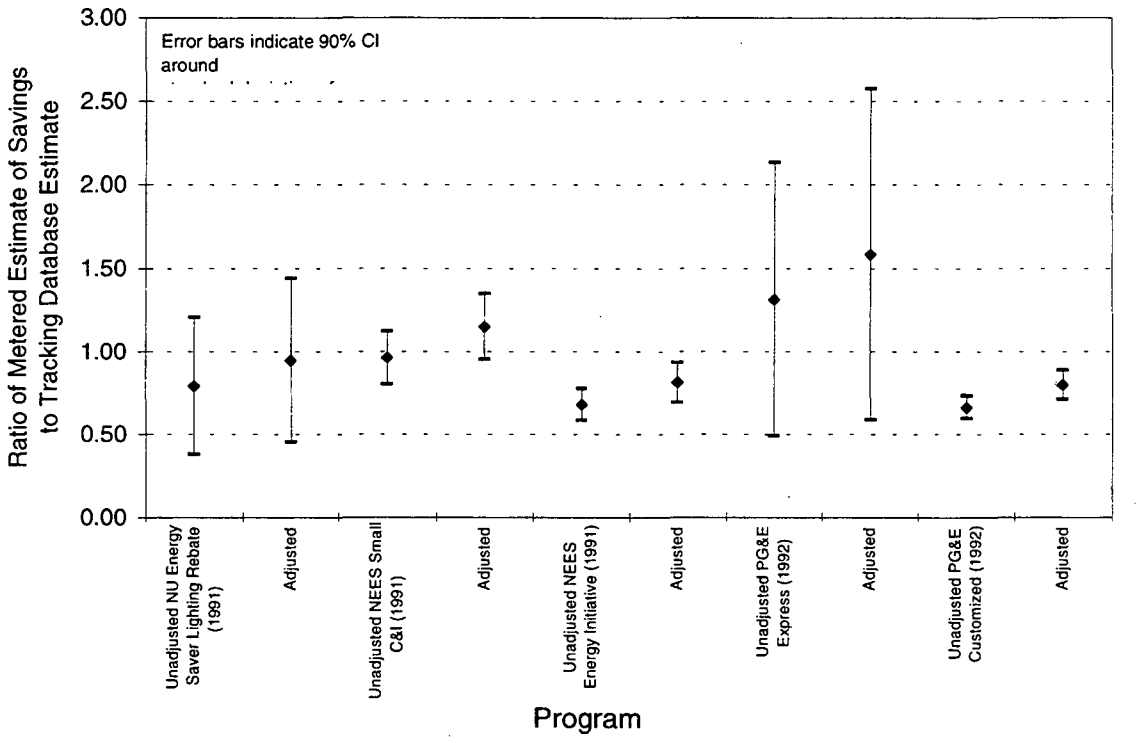
Study	Mean estimate of metered sample	Std. dev. of metered sample	Mean estimate of hours bias	Std. dev of hours bias	Std. dev of sample with hours adjustment	Mean of inter-action effect	Std. dev. of inter-action effect	Mean of adjusted estimate	Std. dev of adjusted estimate	Increase in standard deviation
NU Energy Saver Lighting Rebate (1991)	0.79	± 1.37	1.03	± 0.12	1.41	1.16	± 0.09	**0.94**	± **1.64**	20%
NEES Small C&I (1991)	0.96	± 0.44	1.03	± 0.11	0.47	1.16	± 0.09	**1.15**	± **0.55**	24%
NEES Energy Initiative (1991)	0.68	± 0.28	1.03	± 0.11	0.30	1.16	± 0.09	**0.81**	± **0.35**	25%
PG&E Express (1992)	1.31	± 2.00	1.04	± 0.10	2.08	1.16	± 0.09	**1.58**	± **2.42**	21%
PG&E Customized (1992)	0.66	± 0.25	1.04	± 0.10	0.27	1.16	± 0.09	**0.80**	± **0.32**	27%

The combined effect of the imprecision of limited duration metering and imprecision of the interaction effect increase the standard deviation of metering estimates by 23%, on average. This increase in standard deviation corresponds to an increase of the 90% confidence interval around the mean estimate as well. A comparison of the original estimates of metering savings and precision to the adjusted estimates is given in Figure 4-7. Note that we cannot present an absolute measure of bias because we do not know the actual savings for each program. The mean savings in Figure 4-7 for each program is again presented as a ratio of the metered estimate to the tracking database estimate. The HVAC interaction effect increases the value of this ratio for the each program.

The preceding discussion has centered mainly on the precision of the metered estimates. Now we turn to the issue of metering susceptibility to sampling bias. As described in earlier in this chapter, selection of an unrepresentative sample, both in terms of site selection as well as equipment selected for monitoring at each site, can result in a biased estimate of savings. The size of the bias introduced depends on the characteristics of installations for all customers, and the particular sites, building zones, and equipment metered in the evaluation.

Most metering studies use stratified sampling techniques to develop a sample of buildings representative of the population of participants by both size and type. However, the majority of metering studies only sample a few lighting circuits within each building. Thus the typical metering study could introduce bias by not metering a representative sample of zones within each building. It has been suggested by some critics of evaluation that "convenience sampling", where meters are installed on equipment in the most accessible locations in a building, often occurs. We can estimate the probable effect of convenience sampling with a brief thought experiment.

Figure 4-7. Comparison of adjusted and unadjusted metering results



Let us assume that a program is composed of identical lighting systems, installed in a variety of building zones. Due to convenience sampling, a nonrepresentative sample of these measures is metered to obtain estimates of hours of operation. The nonrepresentative sampling results in a bias in the estimate of hours of operation. The equipment and sampling distributions, and the resulting bias in hours of operation are displayed in Table 4-5.

Table 4-5. Example of potential bias in hours of operation from nonrepresentative sample

Building Zone	Mean Hours of Operation	% of Measures Installed in Zone	% of Metered Sample in Zone
Hall	6,522	15%	20%
Lobby	4.645	10%	15%
Sales Areas	5,388	15%	20%
Open Office	4,067	20%	30%
Other	3,706	10%	5%
Private Office	2,551	20%	10%
Storage Areas	2,282	5%	0%
Conference	1,946	5%	0%
Weighted Mean Hours	3,308	3,693	4,043

The average annual hours of operation across all zones is 3,300 hours. The effective average hours of operation for installed equipment is slightly higher, 3,693 hours per year. The hypothesized effect of convenience sampling is to inflate the estimate of program equipment hours of operation by about 10%, to 4,043 hours. This causes an concomitant 10% increase in the annual savings estimate.

This example does not suggest that all metering studies are subject to a bias of 10%. As stated earlier, the bias can vary considerably. If the metered sample is not representative of the program equipment in every respect, a bias may exist. Additional data on the sampling schemes used and on the resulting metered data would be required to estimate the true extent of this bias in metering studies.

4.2 Comparing Accuracy to the Costs of Data Collection

In this section, we integrate the previous analyses of the chapter to compare estimates of savings from different metering methods with their data collection costs. Our estimates of the performance of each method's results are subject to several qualifications:

We only examine a handful of programs in this analysis. Thus, we describe each method's bias and precision in the context of these programs; we cannot definitively determine the bias and precision of each evaluation method under all conditions. With a larger sample of programs one could produce a more definitive estimate of each method's abilities.

Our estimates of each method's precision are based on variabilities in the program data and ways in which the evaluation methods are used to calculate estimate precision. Our estimates of the bias of metering results are dependent on the factors just mentioned, and most importantly, on the representativeness of the metered sample. We have estimated the range of the potential bias stemming from a nonrepresentative sample, but extensive data on all participants and metered sites and equipment would be required to precisely estimate the bias induced by a specific nonrepresentative sample.

In order to obtain estimates of evaluation data collection and data analysis costs, we reviewed the DSM literature, and we sent a short questionnaire to five DSM evaluation practitioners. Table 4-6 lists the resulting estimates for the cost of metering-based bottom-up evaluation methods.

Table 4-6. Estimates of data collection and analysis costs for bottom-up evaluations

Estimate Type	Data Collection Costs/Site	Data Analysis Costs/Site	Economies of Scale
Lighting Loggers / Hours of Use Logging	$1,300-$1,500	$1,300-$1,500	No
Pre-Post Run-Time Logger (load meter)	$1,300-$4,000	$1,300-$4,000	No

The costs given in Table 4-6 are only rough approximations, based on the judgment of several consultants who regularly conduct these evaluations. The actual costs for a specific evaluation are dependent on the types and sizes of customer buildings, the variety of measures installed by the program, and the specific monitoring equipment used. There are some economies of scale for projects which include large numbers of site visits. Other methods do not provide significant cost reductions with larger sample sizes. For the following cost/precision comparisons, we use the middle value of each range of costs in Table 4-6.

4.2.1 Costs, Bias, and Precision of Metering estimates of Savings

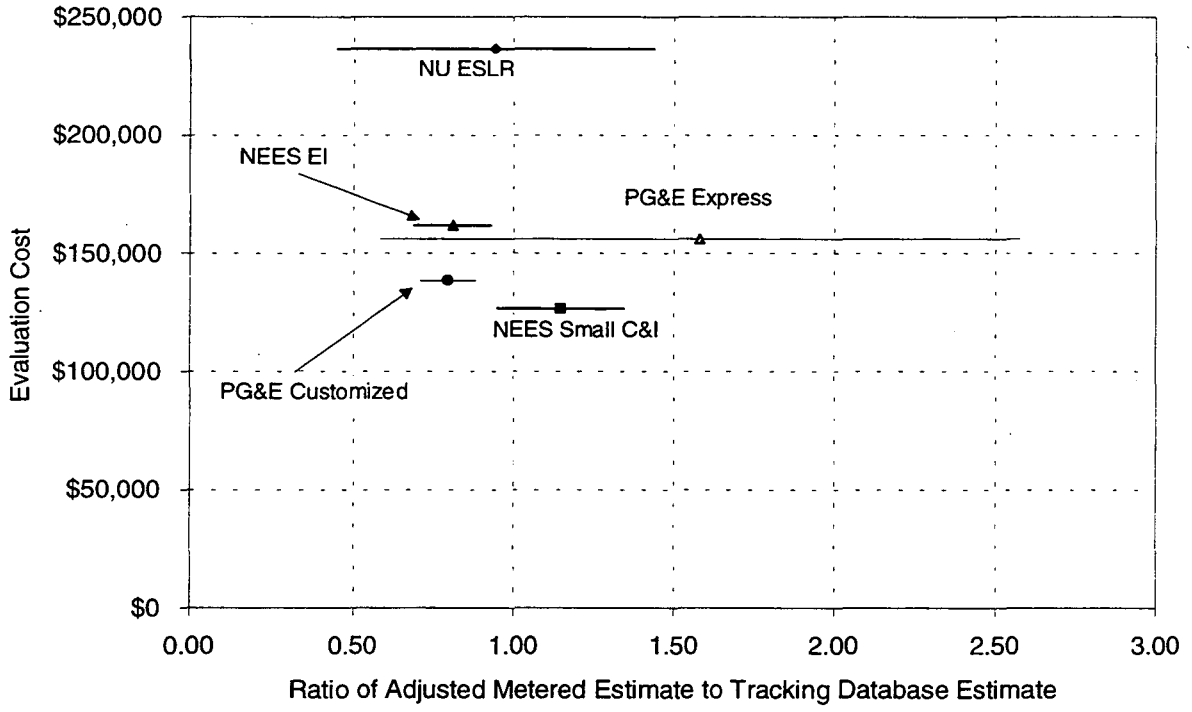
Figure 4-8 presents a comparison of precision and bias of metered savings estimates with evaluation cost. Here again, cost estimates for each evaluation vary due to differences in program size; much of each evaluation's cost is the cost of the tracking database. The absolute bias of these evaluations is unknown, since we have no better estimate of savings for these programs than the metered estimates, adjusted for interaction effects and limited-duration hours of operation logging. Assuming that there are no significant problems with metered sample representativeness, these estimates are unbiased. The position on the x-axis only indicates the ratio of the adjusted, metered estimates to original tracking database estimates. The precision of each estimate has also been adjusted to account for limited duration metering and HVAC/lighting interaction effects, and is expressed in Figure 4-8 as the 90% confidence interval around the mean estimate of savings for each program.

There is a wide range of precisions represented in Figure 4-8. The differences between programs are due to differences in participant characteristics, the variety of program equipment installed, and tracking database accuracy.[6] For example, if all participants installed exactly the same type and quantity of measures, and used them in an identical fashion, the precision on the ratio of metered estimates of savings to tracking database estimates of savings would be much tighter.

In Appendix C, we compare the precision of the evaluations incorporating metering in Figure 4-8 with evaluations utilizing customer billing data and econometric methods. The next chapter investigates the bias and precision of methods which use whole-premises customer billing data.

[6] For the PG&E Express program, the tracking database estimates of savings were imprecise because they were based on standard equipment usage and baseline equipment assumptions rather than participant-specific data. The imprecision in the tracking database estimates of savings propagated to the population-level savings estimate described in Figure 4-8.

Figure 4-8. Cost, Precision and Bias of Metered Estimates of Savings



4.3 Conclusions

Contemporary end-use metering studies omit potentially important uncertainties from their estimates of savings: the imprecision of limited duration metering and the effects of HVAC/lighting interactions reduce the actual precision of metered results, regardless of whether these uncertainties are explicitly acknowledged. We find that these uncertainties reduce the precision of end-use metering estimates by approximately 20%. This reduction could be tempered by 1) longer-duration metering, or 2) a better understanding of interaction effects coupled with detailed information about each customer's HVAC system.

Additional issues arise when bias in metering studies is examined. Bias can result from limited duration metering, ignoring HVAC/lighting interaction effects, and from nonrepresentative metered sample selection.

Our sample of office building hours of operation data suggest that hours vary seasonally. On average, hours of operation are half an hour longer in the winter, and half an hour shorter in the summer than the shoulder months. Neglecting to account for the season metering is performed could bias the estimate of hours of operation, and the resulting estimate of annual savings. Hours of operation are similarly biased by metering occurring during holidays or other days during which the normal work schedule is disrupted.

HVAC/lighting interaction effects increase program savings for most office buildings. Omitting this effect from consideration can result in a 5-15% downward bias in annual savings estimates.

Small sample metering studies depend heavily on the representativeness of the metered sample and proper stratification. Most evaluators already stratify the population to select a representative sample of participants, and then select representative equipment within each facility. Detailed metering results from evaluations would allow an assessment of differences in equipment operation across participants, facility types, and facility zones. Until such detailed reporting is commonplace, enabling analyses of equipment operating differences to be performed, the representativeness of current metered samples will remain in question.

Using Simulation Techniques to Assess Performance of Top-Down Evaluation Methods

A fundamental problem in assessing the bias and precision of evaluation methods which use customer billing data is that the actual savings are never known, so there is no point of comparison from which to contrast different methods. In this chapter, we employ a technique which uses Monte Carlo and building simulation methods to generate a set of buildings and building electricity consumption data for which energy savings are known in advance. Knowing the savings from our sample in advance provides a baseline for each evaluation method's performance. Thus, both bottom-up and top-down evaluation methods can be used to estimate program-related energy savings for these (imaginary) buildings. The results of the evaluation methods can then be compared to the 'true' savings for each building or group of buildings. Varying characteristics of our simulated buildings, resulting in changes in electricity consumption over time or across the sample of buildings, can reveal strengths and limitations of particular evaluation methods. Parametrically or probabilistically varying characteristics to represent a range of real-world buildings will allow us to estimate the bias and precision of each method.[1]

A schematic of the analysis described in this chapter is given in Figure 5-1.

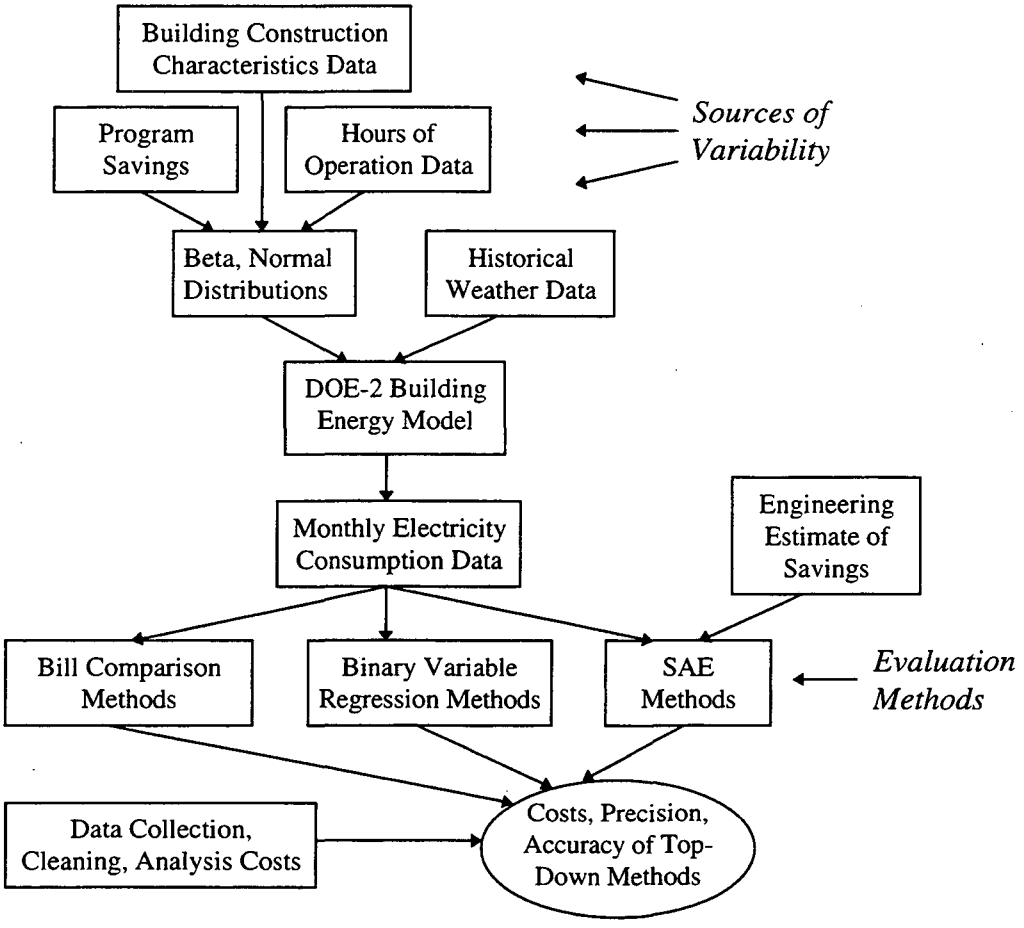
5.1 Creation of Buildings and Building Consumption Data

Examining the top-down evaluation methods requires a set of billing data for which the participants' electricity savings are known. We use Monte Carlo techniques with normal, uniform, and beta (similar to a log-normal, a beta distribution is asymmetric with a long tail) distributions, based on a combination of industry data and expert judgment, to create 500 simulated commercial buildings (250 buildings participate in the program, and 250 do not). Using DOE2,[2] a building energy consumption simulation based on heat-transfer principles, we generate two years worth of monthly electricity consumption records for these 500 buildings.

[1] It is important to note that our simulation techniques deal only with gross savings issues; we do not assess bias in the results of billing analyses that estimate annual savings net of free riders. Our analysis also sets aside the problems of self-selection bias in regression models. Our focus allows us to concentrate on the basic regression forms used and their susceptibility to errors in variables issues.

[2] B. Birdsall, W.F. Buhl, K.L. Ellington, Erdem, A.E., Winkelman, F.C., "Overview of the DOE-2 Building Energy Analysis Program", Simulation Research Group, Lawrence Berkeley Laboratory, Berkeley, CA, LBL-19735, Feb. 1990.

Figure 5-1. Overview of Top-Down Analysis



To simulate the intervention of a lighting retrofit program, participants' lighting intensity (in W/sqft) is reduced by some amount after the final month of the first year. The savings induced by the model varies for each building and is sampled from a beta distribution.

Building materials and construction characteristics were modeled after information from several national and utility building surveys. Appendix B contains a full explanation of the distributions used to represent building characteristics and their sources.

5.2 Varying Building and Consumption Characteristics

Evaluation of energy conservation programs is complicated by the dynamic nature of energy consumption: consumption changes based on business decisions, personal decisions, weather, and other difficult to predict factors. These factors affect savings and the ability of evaluation methods to estimate savings. In order to test the capabilities of different evaluation methods, we vary three factors which affect program savings and overall consumption patterns: the size of the DSM program's kWh savings (by varying the number of measures retrofit and the kW

savings per measure), weather conditions over time, and building hours of operation over time.

Varying the size of kWh savings allows us to determine if the ability of certain evaluation methods to estimate savings is dependent on the size of savings relative to other factors that vary in the simulation. For example, it may be true that regression of monthly billing data may be unable to detect savings in programs that save less than a certain percentage of total consumption.

Weather is arguably the most significant factor affecting energy consumption over time. Heating and cooling demands can double during the winter and summer, respectively. Heating and cooling loads vary not only with each season but also stochastically within seasons. Lighting hours of operation is also affected by changes in ambient (solar) light and by seasonal changes in the length of days.

There are many business-related factors that affect a building's energy consumption: changes in occupancy, changes in business or economic climate, and building renovations can all affect building energy consumption. Rather than attempting to gather data on the frequency and possible effects of all these factors, we have chosen to use changes in hours of operation as a proxy for a wide variety of possible factors which affect energy use. We base our estimates of changes in hours of operation on data from the Energy Edge project described in the previous chapter.

5.2.1 The Size of Program kWh Savings

The distributions being sampled to determine each building's energy savings can be adjusted so that the average effect size of the DSM program can be specified. For example, a bulb rebate program may result in a 5% reduction in total electricity consumption, but a more comprehensive direct install program may result in a 20% reduction in total electricity consumption. For this analysis, three effect sizes have been selected based on a survey of estimated effect sizes from utility commercial lighting programs. These average effect sizes, based on the distributions of measures installed and lighting intensity reductions in Table 5-1 of Appendix B, are given in Table 5-1.

Table 5-1. Average Program Effect Sizes for Participating Buildings

Effect Size	Lighting Electricity Saved	Total Electricity Saved
Small	7%	4%
Medium	16	9
Large	25	14

The actual reduction for each participant is determined by the product of two random variables sampled from a beta distribution: % of measures retrofit, and % of watts saved per measure. The post-installation lighting energy intensity in watts per square foot after the retrofit are calculated as:

$$New\ Watts/Sqft = \frac{Watts\ Saved/Measure}{Original\ Watts/Measure} \times \frac{\#\ Measures\ Changed}{Total\ Measures\ in Building} \times Original\ Watts/Sqft.$$

Representing a lighting retrofit by specifying a reduction in lighting energy intensity has the added benefit of allowing DOE2 to compute interactions between lighting, heating and cooling loads. For a double peaking utility (i.e., having both a substantial summer and winter peak demand period), these interactions can affect total building energy savings significantly, increasing individual building electricity savings.[3]

5.2.2 Changes in Weather over Time

We have run two simulation cases: (1) Identical weather for pre- and post-program years, and (2) dramatic changes in weather for pre- and post-program years. The 'dramatic changes' case uses weather (solar radiation and temperature) data from Chicago for the pre-program year, and data from Washington DC for the post-program year. The difference in average weather conditions between these two sites is more varied than the difference in weather across adjacent years for any single city.

As an improvement to the simulation, more realistic (and less severe) changes in weather could be simulated using the following method: use different historical years of weather data from a single city to simulate more probable weather changes over time. Then identify the warmest and coolest years on record for the city, as well as the years with an 'average' change in weather conditions. The warmest and coolest years could be used to simulate a severe change in weather conditions from pre to post-program years, while the more similar weather years could be used to represent more average conditions. Using both the average and severe weather data would enable us to assess each evaluation method's ability to control for changes in energy use due to weather.

5.2.3 Changes in Building Hours of Operation Over Time

Changes in hours of operation acts as a proxy for a wide variety of human-technology interactions which can affect consumption and thus the size of energy savings, as well as an evaluation method's ability to detect savings. Changes in productivity or output, changes in staffing, and changes in office schedules can all be approximated by changes in operating hours. A description of the development of an set of simulated hours of operation data is in Appendix B.

There are myriad other parameters which reasonably could vary over time or across the customers. Hours of operation and weather are two parameters for

[3] Sezgen, O.A., Huang, Y.J., "Lighting/HVAC interactions and their effects on annual and peak HVAC Requirements in Commercial Buildings", *Proceedings from the 1994 ACEEE Summer Study*, 3:229-239

which substantial data exists, and are known to have a significant effect on customer consumption patterns. Future studies could, using the framework outlined and implemented here, proceed to investigate the effects of other sources of variability on evaluation accuracy, including changes in building occupancy, changes in building operation, and changes in energy costs.

5.2.4 Engineering Estimate of Savings in SAE Regression

As discussed in Chapter Two, an SAE regression incorporates an estimate of annual savings as an explanatory variable in the regression model. If a tracking database estimate of savings is used, the variable would equal the product of annual hours of operation, watts saved per measure installed, and the number of efficient measures installed. The building simulation uses these values as inputs, so we have an 'exact' representation of the tracking database. The end-use metering studies from Northeast Utilities and New England Electric System (described in Chapter Three), however, suggests that tracking databases do not consistently provide the actual values for these parameters. Using the data from NU and NEES, the discrepancies between tracking database estimates, site survey estimates, and 'actual savings' (metered estimates) shown in Table 5-2 were compiled. Using the distributions characterized in Table 5-2 to adjust the tracking database information in the building simulation dataset, we construct savings estimates which mimic the bias and precision of actual tracking database and site inspection estimates of savings. In this way, we can simulate tracking database and site survey estimates of savings, and use these as independent variables in the SAE models. The 'exact' tracking database estimate, taken directly from the building simulation dataset, will also be used in one version of the SAE model to assess the accuracy of the SAE model with perfect savings information.

Table 5-2. Errors associated with tracking database and site survey estimates of savings.

Ratio of: to:	End-Use Metering Tracking Estimate	End-Use Metering Site Inspection
Average	80%	88%
Standard Dev.	40%	22%

5.3 Adjustments to the Regression Models

All regression models include explanatory variables representing building size (sqft) and annual hours of operation. The values of these values are taken directly from the building simulation. This amounts to an assumption that auditors and building managers provide perfect information regarding building size and hours to program auditors and evaluators. Specifying exact values for these variables allows us to more clearly observe the effects of varying weather, hours of operation, program effect size, and tracking database estimates of savings in the models and consumption datasets. While other variables are used in practice,

including variables representing facility type, business and economic conditions, and recent renovations, we only included explanatory variables relevant to the building simulation output.

Some of the regression models required adjustments in order to derive robust estimates of savings. In this section we describe the adjustments which were examined, including corrections for multicolinearity, autocorrelation, and heteroskedasticity.

Both monthly and annual billing data were available for use with each regression model. The time-series, time-series cross-section, and lagged dependent variable models were more accurate using monthly data together with an explanatory variable representing cooling degree days, rather than annual data. The lack of variation in annual cooling degree days prevented use of weather-related explanatory variables in the models with annual consumption data. Identification of the relationship between each explanatory variable and the dependent variable, energy consumption, as separate from noise or other anomalies in the data, requires sufficient variation in the dataset for each explanatory variable.

Inclusion of an explanatory variable representing heating degree days resulted in multicolinearity due to the strong negative correlation (0.85) between heating and cooling degree days. Used separately, the heating degree day coefficient was found to be less significant than the cooling degree day coefficient. The minor significance of the heating degree day coefficient is not surprising since a minority of the buildings were heated with electricity. As a result, the heating degree day coefficient was dropped from the regression models.

5.3.1 Autocorrelation

The cross-section time-series model had a Durbin-Watson statistic near zero, suggesting significant autocorrelation. The pre and post equations were differences in an attempt to remove the autocorrelated error structure. The differencing approach substantially improved the standard errors of the explanatory variables for the cross-section, time-series model, and the results of the differencing equation were used for the cross-section time series regression model reported in Table 5-4.

Durbin-Watson test statistics on the time-series models suggest significant (at the alpha = 5% level for models using monthly data and at the alpha=1% level for models using annual data) autocorrelation errors. The first order correlation estimate given by SAS[4] is consistently positive (ranging from 0.2 for models incorporating annual data to 0.9 for models incorporating monthly data). However, the results of the simple time-series regression model were not significantly different from the corresponding difference model, which corrected for autocorrelation.

[4] The SAS statistics package, by the SAS Institute, Inc. of Cary, NC was used to calibrate all regression models.

The Durbin-Watson test statistic is not suitable for use with time-series models that include a lagged dependent variable. For these models Durbin's h-test was used to investigate autocorrelation. The h-test gives a value based on a standard normal variable, when the result is significantly different from one, autocorrelation may be present in the model. The h-test result was not significantly different from zero for any of the lagged dependent variable models, including the SAE models.

5.3.2 Heteroskedasticity

The effect of heteroskedasticity is usually to bias the standard errors upwards. Stratification of the sample was performed based on building size, with a new stratum defined for each 20,000 sq. ft. interval. In no case did stratification of the sample measurably improve the standard errors of the estimates. In the case of the pooled time-series cross-section model, several strata had statistically insignificant coefficients on the participation indicator variables. The reduced explanatory power of the stratified model could be a result of the reduced sample size and variability in model data. Two stage, weighted least squares methods were also employed as a correction for possible heteroskedasticity. The weighted least-squares models only improved precision by 3%, on average.

5.4 Results

The results of implementing the evaluation methods on the simulated datasets are summarized in Table 5-4. When possible, simulations were run in with nine different datasets: a baseline condition where only the program reduction in lighting intensity affects electricity consumption, a semi-baseline condition where hours of operation also vary on a monthly basis, and a full-variation condition where both weather and hours of operation vary. No static or constant-weather simulations have been run for the comparison group of nonparticipants, so cross sectional models were not implemented for those datasets. Table 5-3 summarizes the average annual kWh savings per building for each program effect size under the full-variation condition.

Table 5-3. Average annual kWh savings for each program effect size

Effect Size	Lighting Savings	Interaction Savings	Total
Small	30,814	3,866	34,680
Medium	67,799	8,395	76,194
Large	105,686	12,900	118,586

Baseline savings are calculated by examining results from the baseline simulation which only varies lighting intensity for the participating buildings. All other variables, including weather and hours of operation, remain constant. The average of the difference in consumption for participating buildings in these two years is defined as true savings. Note that this quantity is not a measure of actual savings in any given year: weather, hours of operation, etc., legitimately affect actual savings amassed in any particular year. However, by defining savings as a

reduction in consumption occurring under static conditions, we obtain a concrete estimate against which the other model results can be compared.

Each cell in Table 5-4 presents the savings estimate as a fraction of the true savings and the precision of the estimate based on a 90% confidence interval. The closer the ratio is to 1.0, the more accurate the evaluation method. The smaller the precision, the more precise the evaluation method. For example, for a program with a small effect size (saving 4% of each building's electricity use), with both weather and hours of use variations incorporated into the dataset, the time-series cross-section regression model overestimated energy savings by 5% (i.e., a ratio of 1.05) with a precision of +/- 38% at a 90% confidence interval.

Table 5-4. Application of Top-Down Evaluation Methods on Simulated Building Data

Variability in data:	Ceteris parebis (lighting intensity change only)			Hours vary by month			Hours and weather vary		
Effect Size	Small	Med	Large	Small	Med	Large	Small	Med	Large
Time-Series (TS) Comparison	1.00 By Definition			1.01	1.00	1.00	0.41	0.75	0.85
Cross-Sectional (CS) Comparison							0.53	0.82	0.90
TS, CS Comparison							0.99	1.02	1.02
TS Regression	1.00 ± 39%	1.00 ± 17%	1.00 ± 11%	1.02 ± 38%	1.01 ± 17%	1.01 ± 11%	1.05 ± 38%	1.04 ± 17%	1.03 ± 11%
TS, CS Regression							0.99 ± 13%	1.02 ± 10%	1.02 ± 9%
CS Regression with lagged dep. variable							0.97 ± 9%	1.00 ± 5%	1.01 ± 3%
Tracking SAE Regression	0.02 ± 45%	0.02 ± 51%	0.01 ± 45%	0.01 ± 89%	0.01 ± 64%	0.01 ± 66%	0.01 ± 100%	0.01 ± 66%	0.01 ± 66%
CS Tracking SAE Regression							0.04 ± 33%	0.05 ± 25%	0.04 ± 25%
Site-Insp. SAE Regression	0.73 ± 6%	0.67 ± 6%	0.60 ± 7%	0.62 ± 14%	0.62 ± 9%	0.57 ± 8%	0.61 ± 17%	0.62 ± 10%	0.58 ± 9%
CS Site-Insp. SAE Regression							0.84 ± 8%	0.85 ± 4%	0.84 ± 4%
Perfect SAE Regression	0.99 ± 1%	0.99 ± 1%	0.99 ± 1%	0.89 ± 10%	0.94 ± 5%	0.95 ± 4%	0.83 ± 14%	0.91 ± 7%	0.94 ± 5%
CS Perfect SAE Regression							0.95 ± 7%	0.99 ± 3%	1.01 ± 2%

Because there are no other factors besides the lighting retrofit program and weather which affect average savings (hours of operation vary monthly, but are equal, on average, for the pre- and post-program years), the evaluation methods which incorporate both participant and non-participant data, or which incorporate participant data and weather data, perform well. In most cases these methods verify savings almost exactly, with good precision.

When hours of operation and weather conditions vary over time, larger program effect sizes allow most methods to more accurately and precisely estimate savings. This can be explained intuitively by recognizing that the larger the effect size, the smaller the other sources of variability by comparison. Increasing the effect size had the most beneficial effect on simple time-series and cross-sectional comparisons. For these methods, the larger effect size improved the estimate of savings from 41% to 85% of actual savings for the time-series comparison and from 53% to 90% of actual savings for the cross-section comparison.

Of the regression-based methods, the least successful are those which include a tracking database or site-inspection estimate of savings as an explanatory variable in the regression equation. Unless a perfect engineering estimate of savings is used as the explanatory variable, these models perform worse than the regression methods which use binary variables to indicate program participation and/or post-program year consumption. This finding has important implications for the current practice in DSM evaluation, where because of the straightforward interpretation of an SAE coefficient,[5] SAE models are a popular alternative to regression models using a binary indicator variable.

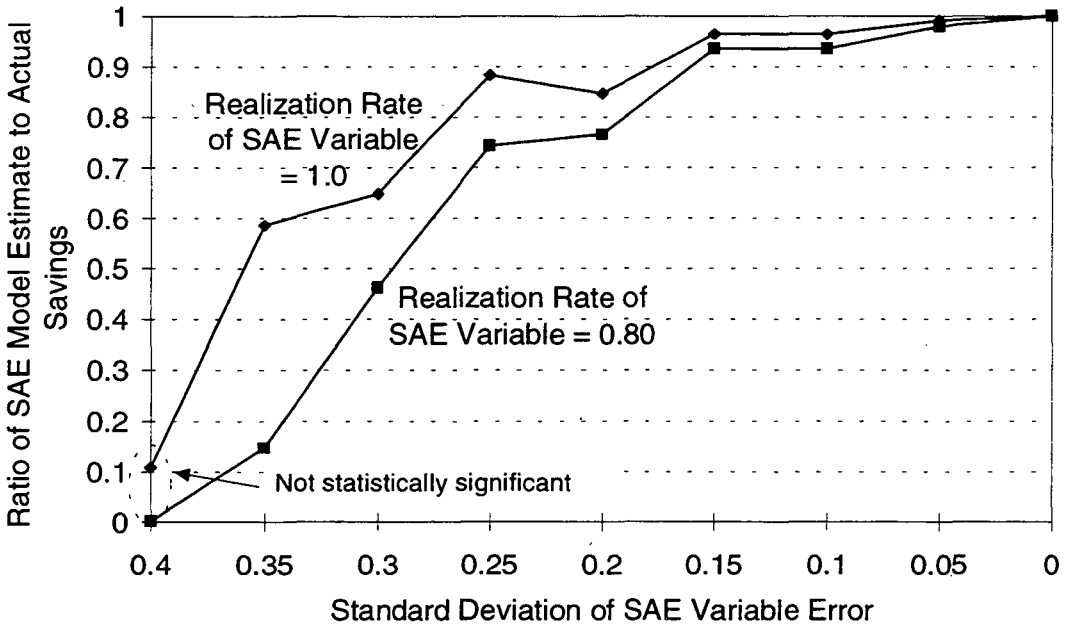
While the SAE models which include the least accurate *ex ante* estimate of savings (with a 40% standard deviation) are clearly unable to provide a reasonable estimate of annual savings, the models which include a more accurate, but not exact, *ex ante* estimate of savings (with a 20% standard deviation) consistently *under*estimate actual savings. The results of the SAE models which incorporate site inspection estimates are also deceiving because of the relatively narrow precisions estimated around the ratios of model-estimated savings to ex-ante savings. Actual program savings are not within a 90% confidence interval around the models' estimates of savings. The SAE model incorporating a 'perfect' *ex ante* estimate of savings provide an unbiased estimate savings, when interaction effects (which are not included in the *ex ante* estimate) are taken into account.

The SAE model results strongly suggest a general conclusion about SAE models: the accuracy of the model is limited by the accuracy of the savings estimate used as an explanatory variable (i.e., an extension of the univariate regression, classical errors-in-variables problem). Given the accuracy of the tracking database and site inspection estimates of savings from the NEES and NU programs, SAE models are not as robust as regression models that use binary indicator variables to estimate program savings. More accurate bottom-up estimates of savings are needed before inclusion in the regression models can improve savings estimates. In order to explore the dependence of SAE model accuracy on engineering estimate bias and precision, we calibrated the SAE regression model with engineering estimates (used as the SAE variable in the regression model) possessing a range of precisions and two accuracies (no bias and 20% bias). The results of this investigation are

[5] As explained in Chapter Two, the SAE coefficient describes the fraction of the engineering estimate of site-level savings which is verified by the econometric model. A ratio of ex post and ex ante savings estimates, such as the one provided by an SAE model, is referred to as a realization rate by DSM evaluators.

displayed graphically in Figure 5-2, which plots SAE model result bias as a function of engineering estimate precision. Recall that in Chapter 3 we estimated the error in tracking database and site inspection estimates of annual savings to have a standard deviation of ±40% and ±20%, respectively.

Figure 5-2. SAE Model Bias Reduced with Precision of Engineering Estimate



As the engineering estimate of savings becomes more precise (indicated by moving right along the x-axis) the bias of the SAE model is reduced, until the model, using our simulated data, is perfectly accurate when using a perfectly accurate engineering estimate of savings for each site. A biased engineering estimate of savings results in a significantly less accurate model result when compared to a unbiased, but similarly imprecise, engineering estimate. When the engineering estimate is fairly precise (i.e., when the engineering estimate's error has a standard deviation of less than ±10%), the biased and unbiased engineering estimates provide similarly accurate estimates of savings when used in the SAE regression model.

Comparisons of SAE analyses from several Northeastern utilities (discussed in Chapter 2) with their corresponding metering studies suggest that SAE models with imprecise engineering estimates may underestimate actual savings as suggested by our regression results. While the metering studies estimate realization rates between 80-100%, the SAE models estimate realization rates of only 50-70%.

The source of the errors in variables problem in SAE models is that the SAE variable is measured with error. Because the true value of the SAE variable (representing participant savings) is not precisely known, the variable used in the

regression is an imperfect measure, known as a "proxy" variable. [6] In the classical error of variables model, the direction of the bias in the variable measured with error (here, the SAE variable) is downward (towards zero). The magnitude of the bias is dependent on the ratio of the variance of the error in the SAE variable to the variance of the SAE variable itself. A basic result of the errors in variable model is that the bias is equal to:[7]

$$Bias\ in\ SAE\ Coefficient = -\beta \frac{\sigma_u^2}{Var(SAE)}$$

Where β is the estimated SAE coefficient, σ_u, is the standard deviation of the error in the measured SAE variable, and var(SAE) is the variance in the SAE variable itself (not the SAE variable error).

In our simulations, we implemented three effect sizes, small, medium, and large savings per participant, based on three different normal distributions (described in Appendix B). Each effect size has a different value for the variance of the SAE variable. To illustrate these variances, the histograms for the three distributions of SAE variables, one for each effect size, are shown in Figure 5-3.

Despite the significant differences in SAE variable variance across effect sizes, SAE models perform equally regardless of effect size.

5.5 Costs of Improving SAE and Time-Series Models

Given the simulation and evaluation results described in the preceding section, it is straightforward to construct a curve which approximates the cost of reducing the bias of an SAE model. For this exercise, we assume that the accuracy of the SAE model is dependent primarily on the precision of the SAE estimate, and that the cost of increasing the precision of the SAE estimate can be approximated roughly by estimates of data collection costs. Table 5-5 provides the per site cost estimates (in $1994) for data collection activities in a commercial lighting retrofit program from a San Francisco Bay Area consulting firm.

---

[6] For a formal explanation of errors in variables resulting in biased coefficients, see Maddala, G.S., *Introduction to Econometrics*, Macmillan, New York, NY, 1988, pp.383-391. The authors are indebted to Roger L. Wright, Mimi Goldberg, and Jeff Schlegel for categorizing the simulation results as an errors in variables problem. We also wish to thank Dan Violette and Greg Rodd, who searched our simulated dataset for anomalous entries that might invalidate our conclusions.

[7] This assumes that the other explanatory variables are measured with little error (relative to each variable's variance) or, if other variables are measured with significant error, there is little or no correlation between the value of the other variables and the SAE variable. Maddala, G.S., *Introduction to Econometrics*, Macmillan, New York, NY, 1988, p.383

Figure 5-3. Distribution of SAE Variable for the Three Effect Sizes

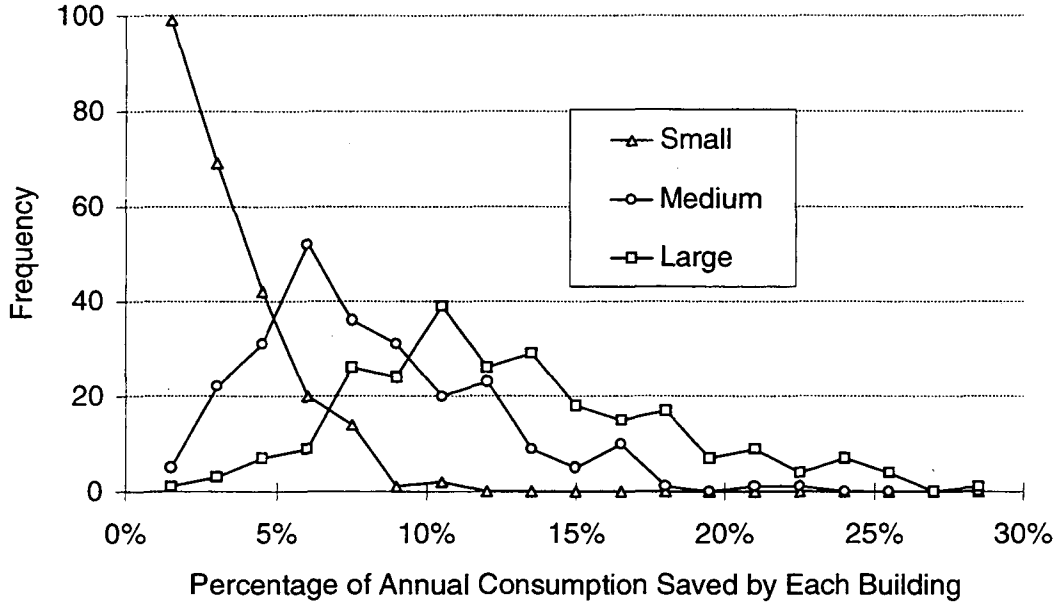
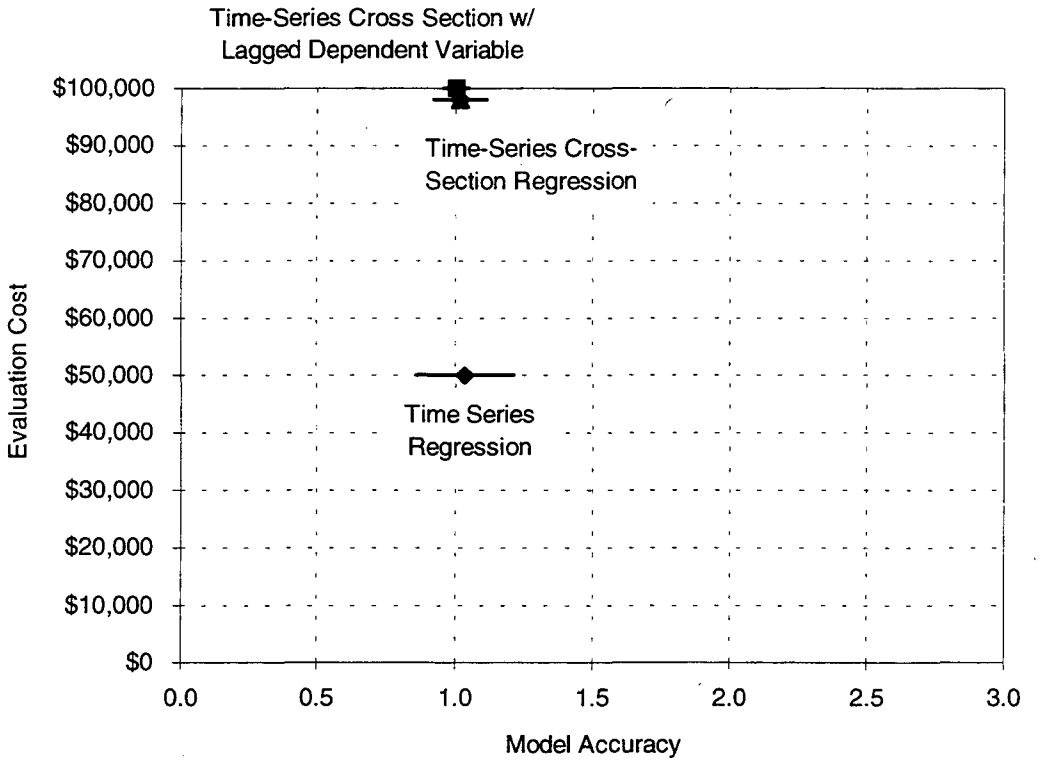


Table 5-5 also shows the cost for 250 sites for each method. While there may in fact be economies of scale associated with a larger number of sites, the information we have collected on DSM consulting firms' pricing policies do not reflect them. Using the data in Table 5-5 we can compare the cost and accuracy of regression models with time-series data, and models with time-series and cross sectional data. The cost/accuracy curves for these models are displayed in Figure 5-4.

Table 5-5. Estimates of Data Collection Costs

Collection Activity	Cost per Site	Cost for 500 Sites
Follow up surveys	$25	$12,500
Participation surveys	$50	$25,000
On site surveys	$1,500	$750,000
On site survey with spot metering	$3,000	$1,500,000
End-Use Metering	$5,500	$2,750,000

Figure 5-4. Regression Model Accuracy vs. Data Collection Costs



While total evaluation costs remain relatively small, incorporating nonparticipant data in the regression does double the data collection costs.[8] The additional data collected improves the precision of the savings estimate considerably, narrowing the 90% confidence interval by more than two-thirds, in the case of the lagged dependent variable model. The lagged dependent variable model, in this instance, provides a more precise and more accurate result than the time-series cross-section model utilizing the same data. Both of the models which incorporate nonparticipant data perform better than the participant-only model. If the increased accuracy of the result is of sufficient value to the evaluator (the value of increased accuracy being determined by the desired use of the savings estimate information), it can be worthwhile to incorporate nonparticipant data in the regression model.

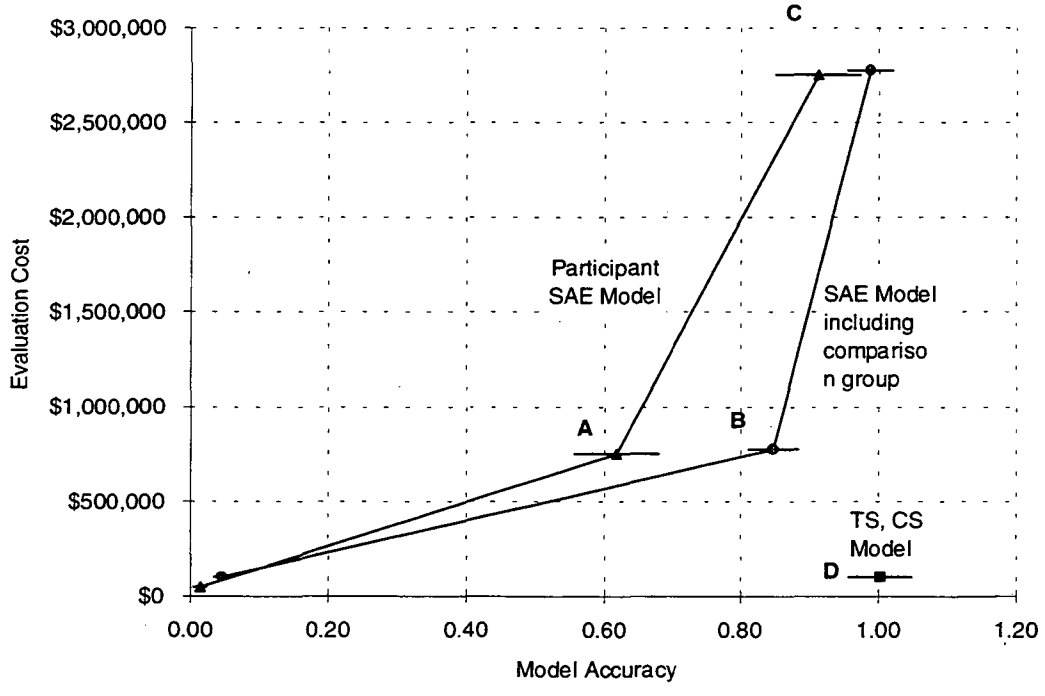
As mentioned in the previous section, the regression models that utilize binary indicator variables perform better than most of the models which use site-specific estimates of customer energy savings in place of the binary variables. Using the data in Table 5-5 we can plot the costs of data collection against the accuracy of

[8] For this analysis we assume that a data collection activity must be undertaken to collect nonparticipant data. Some utilities may have the requisite customer data, including building size and hours of operation estimates, as well as monthly billing and weather data, without the need for additional collection activities.

the corresponding SAE models. Such a plot is shown in Figure 5-5 for a program with a medium effect size. Two variations on the SAE model are shown, one which includes data from non-participants (a cross-sectional model) and one which does not. The cross-sectional model requires nonparticipant site hours of operation and square footage data in addition to billing records. Data collection costs for nonparticipant sites are assumed to be similar to the cost of a participation survey. Model accuracy is expressed as the ratio of the model result to the actual savings. The brackets represent 90% confidence intervals around each model's savings estimate.

Comparing the standard and cross-sectional models results in an insightful conclusion: For a small increase in cost, including nonparticipant data significantly increased the accuracy of the resulting savings estimate (from point A to point B). This is not the case for increases in spending on the data collection for the SAE variable (from point A to point C): If site inspection data has already been compiled, three fold increases in evaluation spending are required to increase model accuracy another 20-30%. Of course, no evaluator would use end-use metering on every building just to use this information in a regression of customer billing data; end-use metering is too costly and metering every site would in itself produce an extremely accurate estimate of savings. Rather, we use this data point to illustrate the stringent requirements on SAE variable precision to enable the SAE model to produce more, yet still not completely unbiased results.

Figure 5-5. SAE Model Accuracy vs. Model Data Collection Costs



This suggests that before an evaluator considers metering or other resource-intensive activities to improve SAE explanatory variable accuracy, it is more cost effective to add comparison data from a group of nonparticipants to the model.

Furthermore, the results in the previous section suggest that a more appropriate technique would be to utilize a simpler econometric model using a binary variable to indicate program participation. Because such a model does not incorporate site-level or tracking database estimates of savings with their inherent imprecision and bias, the resulting model would provide a more accurate estimate of mean savings (across all participants) than the SAE model. For comparison, a time-series, cross-section regression with a binary indicator variable and lagged dependent variable is plotted as point D.

Any conclusions regarding evaluation method superiority and cost-efficiency made using the simulated datasets and evaluation method exercises must be couched in an understanding of the limitations of this analysis. The building consumption datasets' realism is dependent upon the ability of DOE2 to realistically model energy consumption, which has been the subject of extensive research at LBL, as well as at other institutions, and upon the authenticity of the static and varying input parameters input into the DOE2 simulation. To the extent that either DOE2 or the input conditions we have created deviate from real-world conditions, our evaluation model results cannot be generalized to real-world DSM programs.

5.6 Conclusions

In this chapter, we examined several time-series and cross-sectional comparison and regression models, including SAE models. Despite its popularity in the evaluation community, we found that the Statistically Adjusted Engineering Method has dubious value unless the tracking database estimate used in the regression is very precise and reasonably unbiased. Yet, knowing this *a priori* is difficult, if not impossible. While the SAE models' result may provide a convenient metric for expressing savings, the inaccuracies common in tracking databases threaten to hinder, rather than help, time-series regression of customer billing data.

Inclusion of comparison groups in time-series regression can greatly improve the precision of annual savings estimates, at moderate costs. When the DSM program reduces customer consumption by a small amount (4% in our simulation), incorporating nonparticipant data improves the precision of savings estimates by a factor of three. For programs which save a larger proportion of customer electricity consumption, the improvement is smaller, but still significant.

Our simulation, by design, has not allowed us to investigate other, important problems associated with regression models which examine energy savings: namely self-selection and savings net of free riders. By setting these issues aside, we have been able to look at more fundamental issues in regression model specification and savings estimate bias and imprecision.

Integrating Annual Savings Results with Measure Lifetime Estimates

Top-down and bottom-up methods estimate a single year of savings, while the success (defined here as cost-effectiveness) of most DSM programs is dependent on the program-installed equipment operating efficiently at participant sites for 5-20 years. One measure of program costs and savings can be obtained by levelizing the total cost of a DSM program over the number of years the equipment operates, and dividing the result by the annual program savings. The resulting figure has units of *¢/kWh electricity saved* and can be compared to other DSM programs when planning future DSM programs, to specific supply-side options when comparing DSM activities as an alternative to power plant construction, and to the average cost of electricity[1] for a utility when assessing the overall cost-effectiveness of a program.

In this chapter we integrate annual savings estimates into a model for estimating savings over the lifetime of the efficient equipment. These results will be used to describe the uncertainty surrounding estimates of a program's cost of conserved energy.

An importance analysis is undertaken to determine the contributions of annual (bottom-up and top-down) savings estimates, measure lifetime estimates, and free ridership estimates to the overall uncertainty in the lifetime savings estimate. The results of the importance analysis, in tandem with method cost information, allows us to identify cost-minimizing methods to improve the precision of savings estimates.

6.1 Characterizing Measure Lifetimes

Manufacturers provide estimates of the usable life of their equipment, but these estimates do not account for variations in real-world use, or for premature retirement due to building remodeling and renovation. Lacking more accurate data on which to base estimates of lifetime, most evaluations use manufacturers' estimates of equipment life. Only a handful of studies shed light on the actual lifetimes of efficient equipment in the field. In this section, we describe the results of two of the most significant studies and use them to develop estimates of measure lifetimes that can be integrated into a model to calculate lifetime energy savings.

[1] Decision analytic calculations can trade off both kWh (energy savings) and kW (load reductions) against avoided costs of both generation and peak capacity. In this paper, we concern ourselves with generation only. The framework can be extended to include avoided kW and capacity charges.

A Long Island Lighting Company (LILCO) persistence study in 1993 examined program measures at 600 participant sites two, three, four, five, and six years after measure installation.[2] This study provides us with good estimates of short-term persistence for a variety of lighting measures used in commercial buildings. For the LILCO study, Table 6-1 lists the percentage of measures that, as of 1993, were still functioning or had functioned for their estimated lifetime (if the utility estimate of average lifetime had already been exceeded) and then had been replaced with efficient or inefficient equipment. The averages given in the table are weighted by the expected lifetime kWh savings. Thus, the average value represents the persistence of savings, rather than the persistence of measures.

Table 6-1. Results of LILCO persistence study

Technology	Expected Lifetime[3]	lifetime kWh savings / unit	Year Installed: (percentages are measures persisting as of 1993) 1987	1988	1989	1990	1991
Efficient Ballasts	15 years	56	99%	97%	95%	99%	81%
Fl. Current Limiters	12.5	N/A	95%	99%	100%	93%	100%
Fl. Fixture	10	1253	95%	100%	99%	100%	100%
High Int. Dis. < 200W	2	866	85%	92%	98%	100%	100%
High Int. Dis. > 200W	4	810	100%	100%	98%	98%	100%
Optical Reflector	12.5	223	100%	100%	100%	100%	100%
CFL	2	223	100%	99%	94%	98%	99%
34W Fl. Tubes	4	21	100%	96%	95%	93%	86%
60W Fl. Tubes	2.4	55	100%	98%	92%	78%	93%
Weighted Average			94.5%	97.9%	98.1%	99.0%	99.4%
Standard Deviation			5.0%	2.6%	2.9%	7.1%	7.2%

Since the LILCO data represent four separate program years, we cannot make explicit time series comparisons (e.g., logically, persistence after five years cannot be higher than persistence after two years). However the data suggest that for this sample of five years worth of participants, overall savings persistence in the first six years is probably around 95%. It is important to note that high persistence for participants in any given year does not seem to guarantee commensurate persistence for participants in previous or subsequent years. For example, 1991 participants who installed efficient ballasts experienced a lower persistence than participants in any previous year. This variability suggests that, until the reasons for such variability are understood, it may be useful to monitor persistence for all

[2] Applied Energy Group. 1993. *Persistence Study of Energy Conservation Measures Implemented in LILCO's Commercial Audit and Dollars & Sense Programs*, Long Island Lighting Company, Woodbury, NY.

[3] Expected lifetime and lifetime kWh savings/unit values are from *Directory of Commercial Lamps*, The New York State Energy Office, December 1992.

program years, rather than simply extrapolating from those years where persistence studies have been performed.[4]

If we assume, based on the LILCO study, that short-term persistence does not generally appear to be a problem for DSM programs in commercial buildings, we move to the issue of long-term persistence: Do the measures remain installed and operational for their entire assumed lifetimes? A study performed in the Pacific Northwest by Bonneville Power Administration examined the distribution of equipment ages within 600 buildings.[5] The Bonneville study used two different methods to analyze the on-site data and estimate measure lifetimes: A method based on the distribution of ages of existing equipment, and a method based on the reported rates of change of equipment (which was not as successful, due to data limitations).

The method based on the age distribution of existing equipment assumes that the mean of the current age distribution approximates the average measure lifetime. This method assumes that the equipment has been in the marketplace long enough for the age distribution to be in a steady state, and that when equipment is retired, it is replaced with an identical piece of equipment.

Unfortunately, calculating lifetime estimates of some efficient equipment was hindered by the first assumption above; some efficient equipment has not been on the market long enough to possess a steady state distribution of ages. Among lighting-related equipment, the distribution of ages for electronic ballasts was very low, suggesting that this technology has not been in use long enough to consider it in steady state. Because of this shortcoming for some efficient technologies, we use the aggregated results from the Bonneville study, which combine efficient and less efficient versions of different lighting technologies. The estimates of measure lifetime for lighting equipment in commercial buildings, along with 90% confidence intervals for each estimate, are given in Table 6-2.

The confidence intervals, based on the sample size and standard deviation of observed estimates, range from ±5% to ±36%. While the BPA study was the largest of its kind, the results utilize data from only two DSM programs, both of which were implemented in the Pacific Northwest. Ideally, the BPA data could be used with data from other evaluations in other parts of the country to more completely characterize equipment lifetimes. Each evaluation's final estimate of an equipment's lifetime could be considered a single data point, and the standard deviation for this group of data points could be used as a rough characterization of

[4] It should also be noted that these site inspections do not investigate the possibility of degradation of savings over time.

[5] Skumatz, L.S., Hickman, C.H., "Effect Energy Conservation Measures and Equipment Lifetimes in Commercial Buildings: Calculation and Analysis", Proceedings from the 1994 ACEEE Summer Study, Asilomar, CA, 1994, v.8, p.193-204.

Table 6-2. Measure Lifetime Estimates from BPA Study

Equipment[6]	*Mean of Utility Estimates*	*Mean Observed Lifetime (years)*	*90% Confidence Interval*	*Standard Deviation Implied by 90% CI*
Ballast	13.3	10.0	± 17%	± 1.0
Bulb	4.9	4.2	± 10%	± 0.25
Control	—	22.9	± 5%	± 0.70
Fixture	20.0	21.0	± 5%	± 0.58
Reflector	—	6.2	± 36%	± 0.4

the uncertainty in actual lifetimes for each piece of equipment. Because no other studies as thorough as BPA's exist, we rely upon the BPA study for a rough estimate of lifetime variability.

Measure life estimates based on expert judgment from five utilities were averaged to obtain the values in the first column of Table 6-2. While the mean utility estimates are similar to the observed results, the range of utility estimates for many of the technologies was wide, ranging from 50-150% of the observed lifetimes. This suggests that some of the utilities may be significantly under- or overestimating measure lifetimes, and thus lifetime savings. Because we only have observed lifetime data for a small group of measures in a single region, however, it is dubious to assume that the utility estimates which deviate from the observed data are wrong. Thus, this analysis focuses on the precision of measure lifetime estimates rather than the bias created by the use of under- or overestimates of measure lifetimes.

Because the BPA study investigated equipment longevity in only one part of the country, for equipment installed in two, fairly similar, DSM rebate programs, we believe that the BPA equipment lifetime estimates probably underestimate the true variability in measure lifetimes which occurs in commercial lighting DSM programs generally. But since so little attention has been given to careful measurement and estimation of measure lifetimes thus far in DSM program evaluation, we present the analysis in this chapter as a conservative estimate of the importance of robust estimation of equipment lifetimes in future evaluation efforts.

In the next section we use these lifetime estimates to calculate the uncertainty in estimates of program costs per kWh of electricity saved over the lifetime of program equipment.

6.2 Cost to Society: Calculating the Cost of Conserved Energy

Unlike the equation used to calculate annual savings, the cost of conserved energy is not a simple product of its components. The cost of conserved energy is

[6] Ballasts include standard and magnetic efficient ballasts, and electronic ballasts. Bulbs include incandescent, compact fluorescent, and fluorescent. Controls include mechanical on/off, multiswitch, and timer switching. Fixtures include floods, can lighting, spot lighting, and strip lighting.

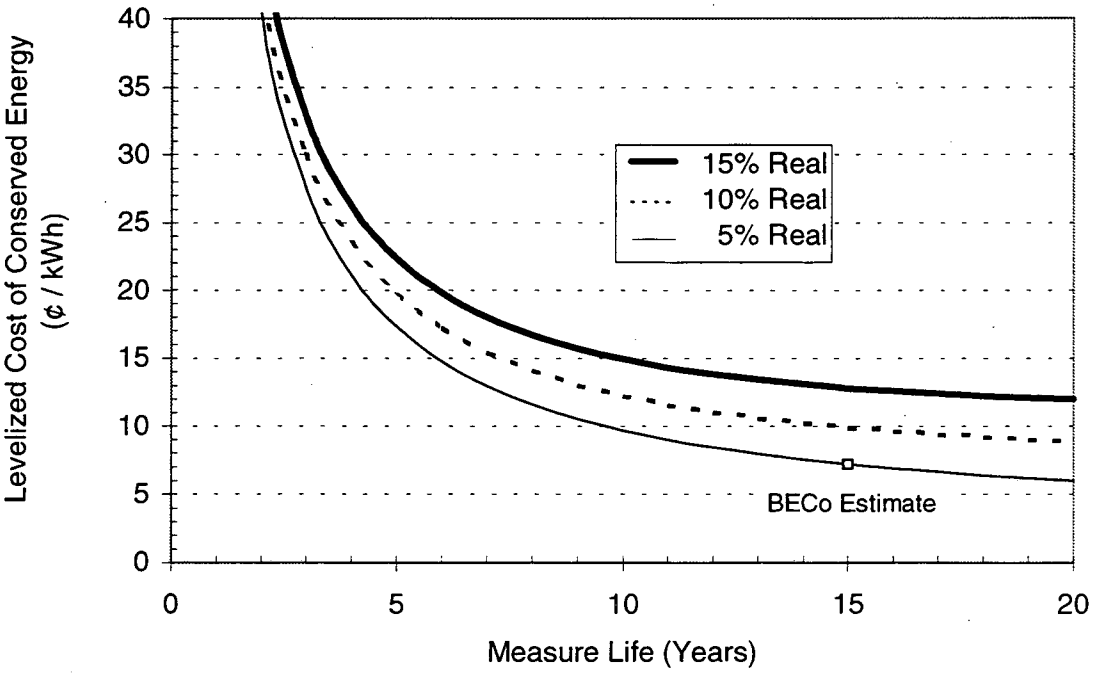
obtained by levelizing a program's cost across the years the program equipment saves electricity, and then dividing the levelized cost by the annual electricity savings of the program. The equation is shown below.

$$Cost\ /\ kWh\ saved = Cost \times \frac{i(1+i)^n}{(1+i)^n - 1} \div Annual\ Savings$$

Where i is the discount rate and n is the lifetime of the program equipment. The term used to levelize program costs over the life of the equipment is known as the capital recovery factor. *Cost* in the equation is the sum of all costs of installation, maintenance, and evaluation, borne by either the utility or by program participants. Annual savings include electricity savings of all participants, regardless of whether they would have installed identical equipment in the absence of the program. In this way, the cost of conserved energy can be thought of as the cost to society of saving energy through DSM. In this framework, it is unimportant who pays for the program and who saves electricity; so long as the costs and benefits can be tallied comprehensively and consistently.

The relation between the equipment life n and the cost of conserved energy is exponential, as demonstrated in Figure 6-1 for a sample DSM program (based on actual program data from a eastern utility). The cost of conserved energy is shown for three discount rates: 5, 10, and 15% real.

Figure 6-1. The Cost of Conserved Energy for a range of measure lifetimes



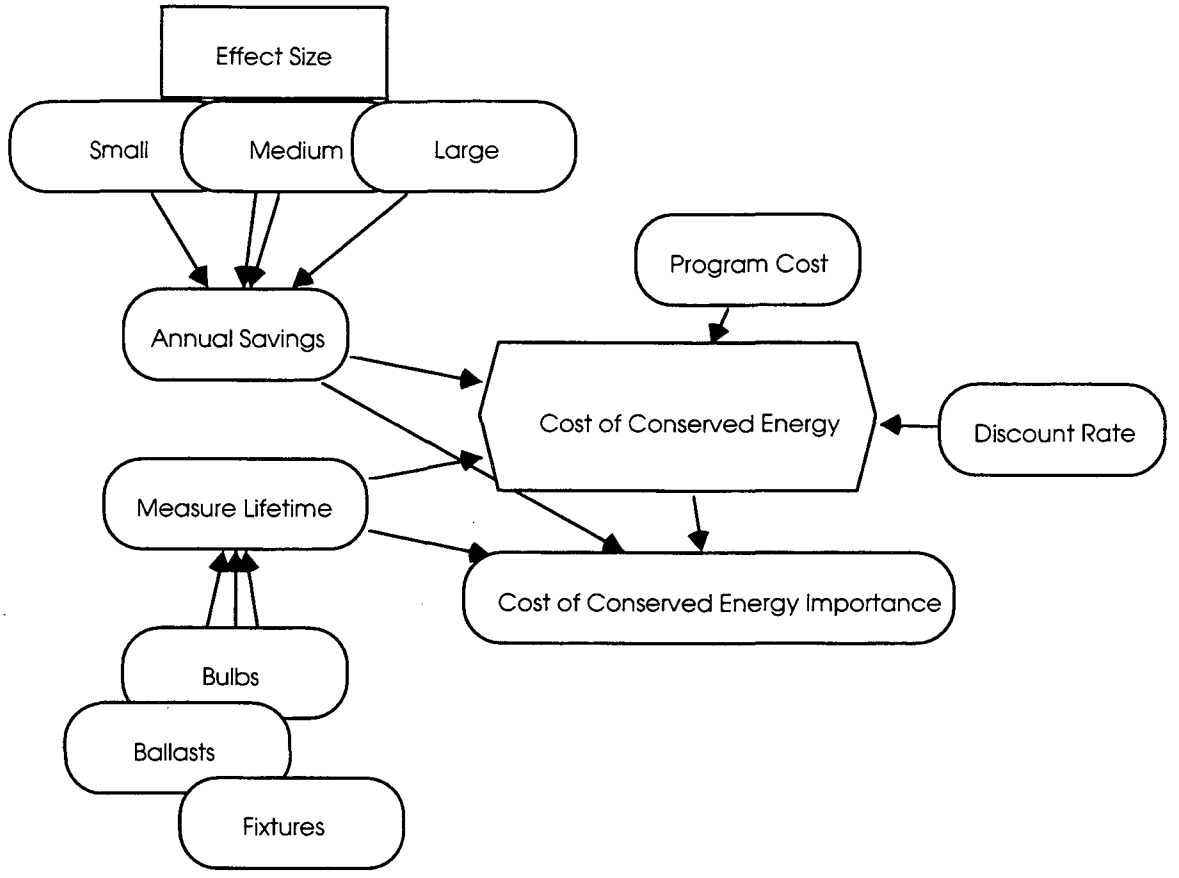
As one would expect, the longer the equipment is used, the lower the overall cost/kWh of electricity savings. The cost of conserved energy decreases dramatically as the equipment lifetime increases from 2 to 10 years. After 10 years, however, longer equipment lifetimes have a less dramatic effect on the cost of conserved energy because electricity savings garnered 10-20 years in the future have little value today, and because the percentage increase in savings from an additional year decreases as the equipment lifetime increases. For equipment which lasts longer than 10 years, each additional year of program savings reduces the cost per kWh of electricity by about half a cent. The non-linear relation between the equipment lifetime and the cost of conserved energy has important implications for the uncertainty of the cost of conserved energy. These implications are explored using a Monte Carlo model in the next section.

6.2.1 A Monte Carlo Model of the Cost of Conserved Energy

A Monte Carlo model allows us to characterize the uncertainty of the cost of conserved energy for commercial lighting programs. We can also estimate the importance of annual savings and measure lifetime estimates to the total uncertainty. The model we use possesses two main inputs, the measure lifetime and annual program savings. The measure lifetime estimates were described in the previous section in this chapter. The program savings estimates are based on econometric regression models and end-use metering analyses undertaken in previous chapters. We can also vary the discount rate, an important input to the cost of conserved energy, and the program effect size, which affects the precision of regression-based estimates of annual savings. A schematic of the model is given in Figure 6-2.

In the previous chapter we compared annual savings estimates generated using top-down and bottom-up methods. While top-down methods appear to perform better than bottom-up methods, and at a lower cost, our estimation of top-down methods' precision and bias relied upon a synthetic data set which only partially mimics the complexity of real-world, top-down analyses. In addition, other factors besides bias and precision guide the decision to use top-down or bottom-up methods in an evaluation. Thus, we do not exclude the results of bottom-up methods in this chapter's uncertainty analysis, and instead present our results here using a variety of top-down and bottom up methods. We incorporate savings and cost information from actual utility DSM programs into the model. We will then compare the results for different methods and summarize the quantitative differences in precision of the cost of conserved energy when different methods are used to estimate the annual savings.

Figure 6-2. The Monte Carlo model for the cost of conserved energy



6.2.2 Cost of Conserved Energy Estimates using Top-Down Methods

In this section we review the results of the Monte Carlo models using top-down methods' estimates of annual savings. We use results from the time-series model, the time-series cross-sectional model, and the time-series cross-sectional model with a lagged dependent variable. We discuss two results obtained with the Monte Carlo model: estimation of the uncertainty in the cost of conserved energy and the importance of annual savings and measure lifetime estimates' contributions to that uncertainty.

Table 6-3 presents the cost of conserved energy results of the Monte Carlo analysis using uncertain estimates of measure lifetime (based on BPA data) and annual savings (based on the building simulation in Chapter 5). Because the top-down methods' performance is dependent on the program effect size, we describe the cost of conserved energy results for each effect size in Table 6-3. The nonlinear nature of the cost of conserved energy function results in asymmetric distributions, thus, we present both the mean and median cost of conserved energy estimates.

Table 6-3. Cost of Conserved Energy for a Hypothetical Commercial Lighting Program

Top-Down Method	Effect Size (savings per participant)	Mean (¢/kWh)	Median (¢/kWh)	Standard Deviation	90% Prediction Interval
Time-Series	*Small*	4.3	4.0	± 1.4	± 57%
	Medium	4.1	4.0	± 0.64	± 26%
	Large	4.0	4.0	± 0.57	± 23%
Time-Series Cross	*Small*	4.1	4.0	± 0.62	± 25%
Section	*Medium*	4.0	4.0	± 0.55	± 23%
	Large	4.0	4.0	± 0.54	± 22%
Time-Series Cross	*Small*	4.0	4.0	± 0.58	± 24%
Section w/Lagged	*Medium*	4.0	3.9	± 0.52	± 21%
Dependent Variable	*Large*	4.0	3.9	± 0.52	± 21%

If point estimates of annual savings and measure lifetime were used to generate an estimate of the cost of conserved energy, we would obtain a result of 4 ¢/kWh. Using the Monte Carlo model with probabilistic estimates of annual savings and measure lifetime results in a comparable mean estimate, but augments the point estimate with additional information about the precision.

When the uncertainty surrounding measure lifetime estimates is combined with annual savings uncertainties, the resulting cost of conserved energy estimate, even in the best case, does not meet the 90/10 criteria[7]: a 90% confidence interval with 10% precision. Only the time-series model is affected by changes in effect size: smaller effect sizes reduce the ability of the model to estimate savings precisely.

Table 6-4 presents the rank correlations between the uncertainty in model inputs and the uncertainty in the resulting cost of conserved energy for each top-down method. The rank correlation takes a value between zero and one; the closer to one, the higher the correlation between input uncertainty and result uncertainty and the larger the contribution of that input to overall uncertainty in the result. For each model, the input with the larger rank correlation with the result is responsible for more of the uncertainty in the result.

[7] The 90/10 criteria is used as a measure of appropriate evaluation by some regulatory bodies, such as the California PUC.

Table 6-4. Importance of Uncertainty in Cost of Conserved Energy Inputs

Top-Down Method	Effect Size (savings per participant)	Rank Correlations	
		Annual Savings Estimate	Measure Lifetime
Time-Series	*Small*	0.88	0.49
	Medium	0.61	0.74
	Large	0.46	0.87
Time-Series Cross Section	*Small*	0.50	0.82
	Medium	0.39	0.90
	Large	0.37	0.92
Time-Series Cross Section w/Lagged Dependent Variable	*Small*	0.44	0.90
	Medium	0.23	0.98
	Large	0.22	0.99

When the simple time-series model is used to calculate annual savings and the program has a small effect size (defined as a program which saves only 4% of a customer's total electricity consumption) the uncertainty in the annual savings estimate overwhelms the result. However, when the time-series model is used with programs with medium or large effect sizes, the uncertainties are comparable, and when the time-series model is used with a program with a large effect size, the measure lifetime uncertainty overwhelms the result.

With the exception of the simple time-series model, the uncertainty associated with annual savings estimates is consistently less important than the uncertainty associated with the measure lifetime estimate. In most cases the correlation between measure lifetime uncertainty and the cost of conserved energy uncertainty is more than twice as large as the correlation between annual savings uncertainty and the cost of conserved energy uncertainty. The implication for evaluation activities is that more resources should be devoted to reducing the uncertainty in measure lifetime estimates.

6.2.3 Cost of Conserved Energy Estimates using Bottom-Up Methods

The results of the analysis of bottom-up methods in Chapter 4 demonstrated that the precision of end-use metering can vary dramatically. Because the variation is based on the quality of the tracking database and the variability of measure types and customer consumption, no single precision could represent the outcome of all end-use metering studies. Moreover, we have inadequate data to characterize a distribution for end-use metering study precision.

Rather than attempt to represent end-use metering method precision with a single value, we use three estimates of precision to represent good, average, and poor precision which can result from end-use metering. The three estimates are based

on the precision of the end-use metering studies examined in Chapter 4, and are presented in Table 6-5.

Table 6-5. Parametric estimates of end-use metering precision

Poor Precision	± 50%
Average Precision	± 25%
Good Precision	±10%

Use of these precisions in the Monte Carlo model of the cost of conserved energy results in the cost of conserved energy estimates presented in Table 6-6. Only the model using the 'good' end-use metering precision results in a cost of conserved energy estimate with precision comparable to the estimates obtained using top-down methods.

Table 6-6. Cost of Conserved Energy for a Hypothetical Commercial Lighting Program

Bottom-Up Precision	Mean (¢/kWh)	Median (¢/kWh)	Standard Deviation	90% Prediction Interval
Poor	2.5	4.0	± 86.	—
Average	4.4	4.0	± 1.7	± 70%
Good	4.1	4.0	± 0.65	± 27%

Unless an evaluator can be sure that a metering evaluation can provide results in the range of our 'good' precision estimate, the resulting estimates of savings will be much less precise than the results of less expensive top-down studies.

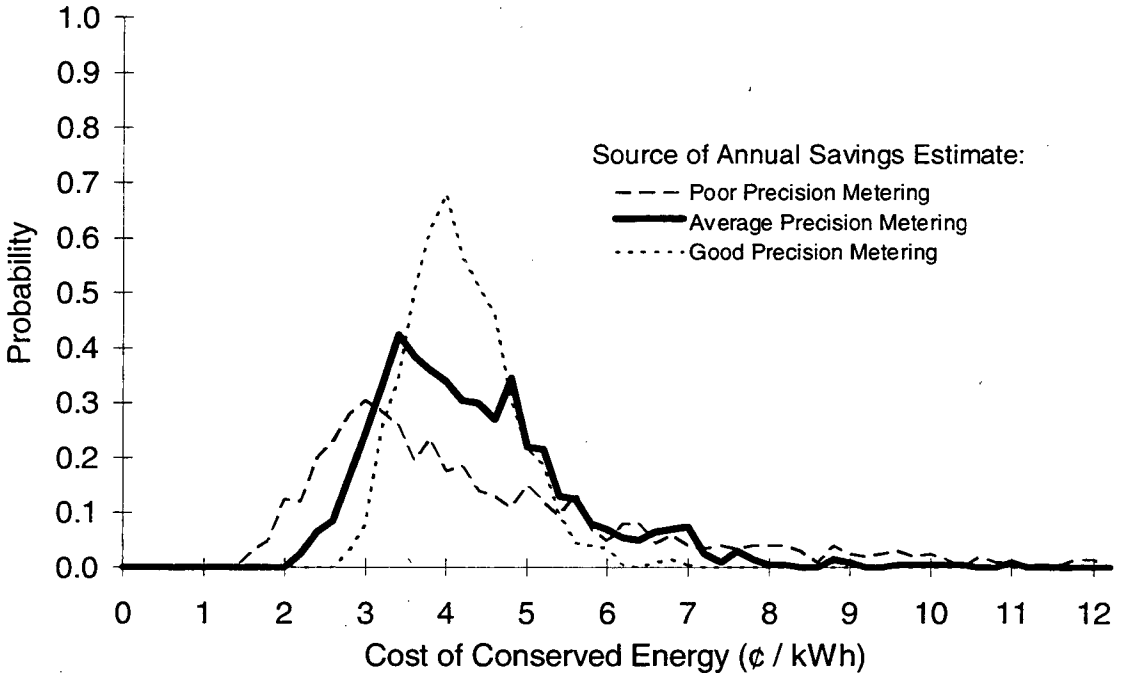
Table 6-7. Importance of Uncertainty in Cost of Conserved Energy Inputs

	Rank Correlations	
Bottom-Up Precision	Annual Savings Estimate	Measure Lifetime
Poor	0.83	0.20
Average	0.90	0.46
Good	0.59	0.76

As one would expect, the uncertainty in annual savings estimates from metering overwhelm the uncertainty in measure lifetime estimates for poor and average estimates of metering data precision. Only when precision is in the 'good' range, does measure lifetime uncertainty dominate the resulting lifetime savings. Thus, the majority of the uncertainty in the cost of conserved energy originates in the metered estimate of annual savings.

Figure 6-3 illustrates the probability distributions for the cost of conserved energy estimates calculated using metered estimates of annual savings. As the precision of the meter estimate of annual savings decreases, the distribution widens and becomes asymmetrical, with an elongated tail on the right-hand side.

Figure 6-3. Distribution of CCE Estimates calculated with metered estimates of annual savings



6.3 Conclusions

Integrating the annual savings estimates with measure lifetime data has allowed us to estimate the overall uncertainty in the cost of conserved energy. This study is unique because of its treatment of both annual savings and measure lifetime as uncertain quantities.

Acknowledging that measure lifetime estimates are uncertain quantities increases the overall uncertainty of cost of conserved energy estimates. When even the most precise estimates of annual savings are used, the cost of conserved energy does not meet a 90/10 criterion of precision. By using point estimates of measure lifetimes in their calculations of lifetime savings and the cost of conserved energy, utilities overstate the precision of their findings.

Despite our use of a conservative estimate of measure lifetime uncertainty, we observe that in many cases the measure lifetime estimate is responsible for the lion's share of uncertainty in the cost of conserved energy estimate. An analysis of the cost of reducing measure lifetime uncertainty should be undertaken to

determine if it is cost effective to concentrate more resources on measure lifetime uncertainty and less on annual savings estimates.[8] Only when annual savings estimates based on end-use metering are incorporated into the calculation of the cost of conserved energy is the measure lifetime uncertainty responsible for a minority of the uncertainty in the result.

The nonlinear nature of the cost of conserved energy equation affects the shape of the resulting probability density function. The cost of conserved energy is asymmetric with an elongated right-hand tail. This asymmetric distribution underscores the importance of precise estimation of the components of the cost of conserved energy: Higher cost of conserved energy estimates cannot be ruled out by the analyst unless inputs to the equation are of sufficiently high precision.

[8] A future iteration of this analysis could incorporate these.

Calculating the Uncertainty in Program Cost-Effectiveness Estimates

Past discussions of DSM program evaluation have suggested that the appropriate level of evaluation is dependent on the cost and performance of each evaluation technique, and the value of the resulting information to the evaluator, regulator, or program planner.[1] We agree, and in this chapter present a case study exploring the appropriate level of evaluation for a particular objective. In this chapter, we relate the precision and bias of evaluation methods to estimates of a program's cost-effectiveness, and consequently to a program screening exercise which uses evaluation results to determine the programs that will receive funding to operate for another year.

We use Monte Carlo techniques to estimate the effects of imprecision and bias in DSM program savings estimates and in the resulting program cost-effectiveness estimates. We review the potential biases of the evaluation methods described in previous chapters. The results of these calculations enable us to discuss the appropriate levels of DSM program evaluation with the objective of ensuring cost-effectiveness. Stated differently, we examine the level of permissible bias and imprecision in evaluation when evaluation results are used to verify program cost-effectiveness.

7.1. Introduction

Estimates of the cost-effectiveness of DSM are based on evaluations of program impacts. The evaluation methods used are, to an extent not well understood, subject to errors of imprecision and bias. Evaluation imprecision can reduce evaluator confidence in estimates of program cost-effectiveness, and evaluation bias can result in non-cost-effective programs being mislabeled as cost-effective. We assess the uncertainty in estimates of DSM program cost-effectiveness for evaluation methods of varying precision and accuracy. By first examining the effects of imprecision and bias, we can then assess the impact of evaluation method choice on our confidence in the cost-effectiveness of a program. The results of these calculations enable us to discuss the appropriate levels of DSM program evaluation with the objective of confidently assessing cost-effectiveness.

In this chapter, we begin with a review of the implications of bias and imprecision in evaluation results. We then discuss the range of cost-effectiveness estimates observed in recent commercial lighting rebate programs. Based on this range of

[1] Hummel, Phillip E. (1993), "Resource Allocation and DSM Program Evaluation Planning", *Proceedings of the 1993 Energy Program Evaluation Conference*, Chicago, IL, pp. 637-642, August and Wirtshafter, Robert, Les Baxter (1991), "Establishing Priorities for Future Evaluation Efforts", *Proceedings of the 1991 Energy Program Evaluation Conference*, Chicago, IL, pp. 137-142. August.

cost-effectiveness estimates we assess the effects of imprecision and bias on evaluator confidence in program cost-effectiveness, and discuss the implications of these findings for future evaluations.

7.2. The Implications of Biased and Imprecise Evaluation Results

Biased, i.e., under- or over-estimates of savings, have important implications on several levels: For the utility, biased estimates of savings misinform about program cost-effectiveness. Biased over-estimates of savings may cause utilities to retain DSM programs which are not, in reality, cost-effective. At the state regulatory level, overestimates of savings will result in utility overcompensation for lost revenues (for lost revenues which, in fact, were never lost) and payment of excessive shared savings incentives. Thus, the utility is allowed to collect additional, unjustified revenue from ratepayers. At the national level, plans to reduce national dependence on fossil fuels or reduce power plant emissions using DSM activities may fall short of desired goals if plans are based on studies which exaggerate actual savings.

An *imprecise* estimate of savings has some slightly different implications: Imprecision in annual savings or measure lifetimes can affect the mean cost of conserved energy estimate, and reduce confidence that a marginally cost-effective program is really cost-effective. Most of the regulatory concern regarding precision suggests a fundamental desire for a precise estimate, but this desire is not necessarily based in the requirements of any particular use of the evaluation results. In many cases the 90/10 criteria is applied to estimates of annual savings, without a similarly rigorous criteria being required for lifetime savings or for the resulting estimates of the cost of conserved energy. Vine and Kushler discuss the history of regulatory mandates for evaluation precision.[2]

We suggest that a precision criteria of 90/10 is usually unnecessary for confidently verifying cost-effectiveness. Bias in evaluation results, depending on the evaluation methods used, appears to be a greater threat to accurate cost-effectiveness calculations. The importance of bias is compounded by the necessity of considering imprecision around the true (unbiased) mean, not the biased mean, which is usually all that is contemplated in evaluation practice today. In the following sections we consider the implications of precision and bias separately for clarity, although the two must be addressed concurrently in evaluation practice.

7.3. Assessing Cost-Effectiveness

The cost-effectiveness of utility DSM programs is gauged by comparing a program's cost of conserved energy, the levelized cost of the program over the installed equipment's anticipated lifetime, to the sponsoring utilities' avoided

[2] Vine, Edward L., Martin Kushler (1995), "The Reliability of DSM Impact Estimates", *Proceedings of the 1995 Energy Program Evaluation Conference,* Chicago, IL.

costs.[3] A program that provides kWh savings at a levelized cost equal to or less than the levelized avoided costs is considered cost-effective, and has a total resource cost (TRC) test ratio greater than one.[4]

Even if an estimate of savings results in a TRC test ratio greater than one, the evaluator cannot rule out the possibility of the program not being cost-effective without some estimate of the savings, cost, and avoided cost estimates' precision. Due to the nature of the cost of conserved energy calculation, a more imprecise savings estimate increases the probability that a program's cost of conserved energy is larger than anticipated, which can shift the mean TRC test ratio to less than one. Under certain circumstances, an imprecise estimate of savings can dramatically reduce confidence in program cost-effectiveness.

While an imprecise estimate of savings can reduce confidence in a program's cost-effectiveness, a biased estimate of savings can misrepresent a non-cost-effective program as cost-effective. Because assessment of bias requires an independent estimate of the 'true' savings for comparison, our characterization of bias is understandably less complete, but not necessarily less important, than our characterization of savings estimate imprecision.

7.4. Cost-Effectiveness Estimates for Commercial Lighting DSM

The recent DEEP Commercial Lighting Report estimated the cost of conserved energy and reported utility-estimated avoided costs for 20 commercial lighting programs.[5] Examining the ratios between the estimates of avoided costs and total resource costs for these 20 programs provides some insight regarding the distribution of typical (but probably biased) cost-effectiveness estimates. The distribution of reported cost-effectiveness estimates allows us to impute the bias and imprecision conditions under which the cost-effectiveness of these programs would be erroneously reported. Table 7-1 lists the utility estimated avoided costs, the cost of conserved energy, and the TRC test ratio for the 20 commercial-sector lighting programs examined in the DEEP report.

When point estimates of the cost of conserved energy were compared to each utilities' estimate of their avoided costs, all of the programs examined in the DEEP report were cost-effective, i.e., had TRC test ratios greater than or equal to one. A few (15%, but only 1% by energy savings) were only marginally cost-effective, with ratios less than 1.5. The majority (55%, 50% by energy savings) had cost-

---

[3] Avoided costs are also levelized over the life of the efficiency measures using a discount rate equivalent to the utilities' cost of capital.

[4] California Public Utilities Commission and California Energy Commission (1987), "Economic Analysis of Demand-Side Management Programs." *Standard Practice Manual*, P400-87-006, December.

[5] Eto, Joseph, Ed Vine, Leslie Shown, Richard Sonnenblick, Chris Payne (1994), "The Cost and Performance of Utility Commercial Lighting Programs", Lawrence Berkeley Laboratory, LBL-34697, May.

Table 7-1. Total Resource Costs and Avoided Costs from 20 Commercial Lighting Programs

Sponsoring Utility	Annual Savings (GWh)	Cost of Conserved Energy (¢/kWh)	Avoided Costs (¢/kWh)	Total Resource Cost Test Ratio
BPA	2.4	4.5¢	4.7¢	1.0
BHEC	2.1	4.7¢	5.0¢	1.1
IE	1.1	4.4¢	4.8¢	1.1
NMPC	101.4	6.0¢	9.0¢	1.5
BECo	8.3	7.2¢	11.2¢	1.6
GMP - Small C/I	3.0	7.6¢	12.1¢	1.6
PG&E	115.7	5.0¢	8.5¢	1.7
SDG&E	2.0	4.1¢	7.2¢	1.7
SMUD	43.7	6.5¢	11.2¢	1.7
CHG&E	16.1	3.7¢	6.8¢	1.9
GMP - Large C/I	16.3	6.3¢	12.1¢	1.9
SCL (Pilot)	1.1	2.5¢	4.7¢	1.9
Con Edison	91.9	6.8¢	14.0¢	2.1
NEES - Small C/I	23.5	5.2¢	10.8¢	2.1
CMP	15.7	1.8¢	4.6¢	2.5
NEES - EI	104.3	3.7¢	10.0¢	2.7
NU - ESLR	149.8	2.5¢	8.1¢	3.2
NYSEG	53.9	2.3¢	10.0¢	4.3
SCE	72.8	1.2¢	7.2¢	5.8
PEPCO	40.5	1.2¢	7.5¢	6.4

effectiveness ratios ranging from 1.5 to 2.1. A final group (30%, 50% by energy savings) had cost-effectiveness ratios ranging from 2.5 to 6.4. These three groups form the basis for our parameterization of cost-effectiveness estimates. We can simulate three programs with mean cost-effectiveness equal to the mean from each of the three groups.[6] We can then estimate the effects of an imprecise estimate of savings on the cost-effectiveness estimate for each program. Table 7-2 summarizes our parameterization of cost-effectiveness estimates.

[6] Avoided cost calculation is a complicated matter. A complete accounting involves estimation of a utilities' fixed and variable costs per kWh and per kW supplied. These costs will vary over the life of program measures, and a thorough understanding of the utilities' resource acquisition plans is required to estimate future changes in avoided kWh and kW costs. Finally, DSM program characteristics also affect the calculation of pertinent avoided costs: A program that saves energy on-peak will have a larger avoided kW cost component than a program that only saves energy during off-peak hours.

With this in mind, it is clear that avoided cost estimation is itself subject to considerable uncertainty. While we recognize the importance of correct avoided costs calculation, an in-depth discussion of the uncertainties associated with avoided cost estimation, or of the utility and customer-borne cost elements in the cost of conserved energy, is beyond the scope of this paper.

Table 7-2. Parameterization of TRC test ratios

Range	Mean Total Resource Cost Test Ratio	Range of TRC Test Ratios	% of the 20 DEEP Sample Programs in Range	% of Annual Savings in Range
Low	1.1	1.0 - 1.1	15%	1%
Medium	1.8	1.5 - 2.1	55%	50%
High	4.2	2.5 - 6.4	30%	49%

7.5. Precision of Bottom-Up and Top-Down Evaluation Methods

This section summarizes estimates of evaluation precision from our analyses of both top-down (econometric methods based on whole-premise billing data) and bottom-up (metering methods utilizing information on specific equipment installed) estimates. Myriad factors can affect the precision of both methods, and the estimates of precision given here are based on limited program data and a subset of all available evaluation methods. Thus, these estimates of precision do not universally apply to every econometric or metering study one could conduct, but rather provide a rough estimate of the range of precisions one could expect using a variety of methods.

It is also important to note that estimates of precision obtained with different evaluation methods are not strictly comparable. A value's precision is entirely dependent on the implicit assumptions that govern which aspects of a quantity are thought to be imprecise. The precision of an end-use metering- derived savings estimate is typically based on information on the sample size and sample homogeneity when compared to the participant population. The precision of an econometrically derived savings estimate is based on the capacity of the econometric model to systematically explain variability in the participant billing data. The statistical assumptions inherent in multivariate regression (e.g., normality and independence) also implicitly affect the calculation of estimate precision.

In order to create these rough estimates of the precision (and rough estimates of bias, which we discuss later in the chapter) associated with different evaluation methods for commercial lighting programs, we have performed a number of detailed analyses based on both actual program and simulated program data, as described in previous chapters. To estimate the bias and precision of end-use metering methods, we compared results from a handful of short and long-term metering studies, investigating hours of operation, sample size and selection, and interaction effects between heating cooling, and lighting equipment. To investigate the bias and precision of econometric methods we used the building energy modeling program DOE2 to simulate a set of participant and nonparticipant buildings' monthly energy consumption, and estimated econometric models using the results.

Table 7-3 presents the range of relative precisions (at the 90% confidence level) we obtained in the aforementioned analyses. To represent a diversity of evaluation methods, we parameterize the precision of these evaluation methods with low,

Table 7-3. Parameterizations of Annual Savings Estimate Precision

Range	Precision from Econometric Analysis with Simulated Data	Precision from Analysis of End-Use Metering Data
Low	15%	50%
Medium	10%	25%
High	5%	10%

medium, and high precision estimates. One should not deduce from Table 7-3 that econometric methods are inherently superior to metering methods. Our method of obtaining estimates of econometric precision used simulated consumption data which probably understated the variability in an actual set of monthly billing data. As mentioned earlier, estimates of precision from different methods are based on different statistical assumptions, and are therefore not strictly comparable. Finally, end-use metering provides a wealth of additional evaluation information above and beyond simple estimates of annual program savings.

7.6. The Effect of Imprecision on Cost-Effectiveness Estimates

In this section, we use the previous sections' information on the imprecision of evaluation method results and the parameterization of cost-effectiveness to estimate the effects of imprecision on confidence in program cost-effectiveness estimates. We utilize a Monte Carlo model to propagate uncertainties because the method and results are easily grasped without a detailed understanding of calculus or other analytic propagation of error techniques, and because Monte Carlo techniques allow more freedom in specification of uncertain quantities and functional relationships.

Additional uncertainty is incorporated into the cost-effectiveness calculation with the incorporation of an uncertain measure lifetime estimate, based on inventories of efficient equipment installed in lighting programs in the Pacific Northwest.[7] Most regulators focus on the precision of annual savings. By incorporating an uncertain estimate of measure lifetime, we can estimate the precision of lifetime savings and program cost-effectiveness.

7.6.1. Monte Carlo Model Results

Three examples of the resulting distributions of the TRC test ratio from the Monte Carlo model are given in Figure 7-2.[8] The distributions displayed reflect annual savings estimates of average precision obtained through end-use metering. Each

[7] Skumatz, Lisa S., Curtis Hickman (1994), "Effect of Energy Conservation Measures and Equipment Lifetimes in Commercial Buildings: Calculation and Analysis", Proceedings from the 1994 ACEEE Summer Study, Asilomar, CA, v.8, pp. 193-204.

[8] The Monte Carlo model sampled 1000 points from each distribution, obtained using median hypercube sampling.

of the three distributions represents a different mean estimate of the TRC test ratio, representing the three parameterizations described in Table 7-2.

The distributions with some portion of their area to the left of 1.0 represent programs which, given the precision of the evaluation methods used, could be non-cost-effective even though the mean estimate, which might be submitted alone as an estimate of cost-effectiveness in a regulatory hearing, is greater than 1.0.

Table 7-4 lists the fraction of each distribution that lies below 1.0, indicating the likelihood of non-cost-effectiveness. Only the distributions for programs with low mean total resource costs have a significant portion of their area below 1.0. Thus, the risk of mistakenly labeling a program cost-effective when it actually is not is highest for programs whose mean estimates of the TRC test ratio are close to 1.0. This result is intuitive: imprecise measurement which results in a ratio close to one has a greater chance of actually being below one than a similarly imprecise measurement which results in a ratio much larger than one.

Figure 7-1. Distributions of the Total Resource Cost Test Ratio for Medium Precision Metering

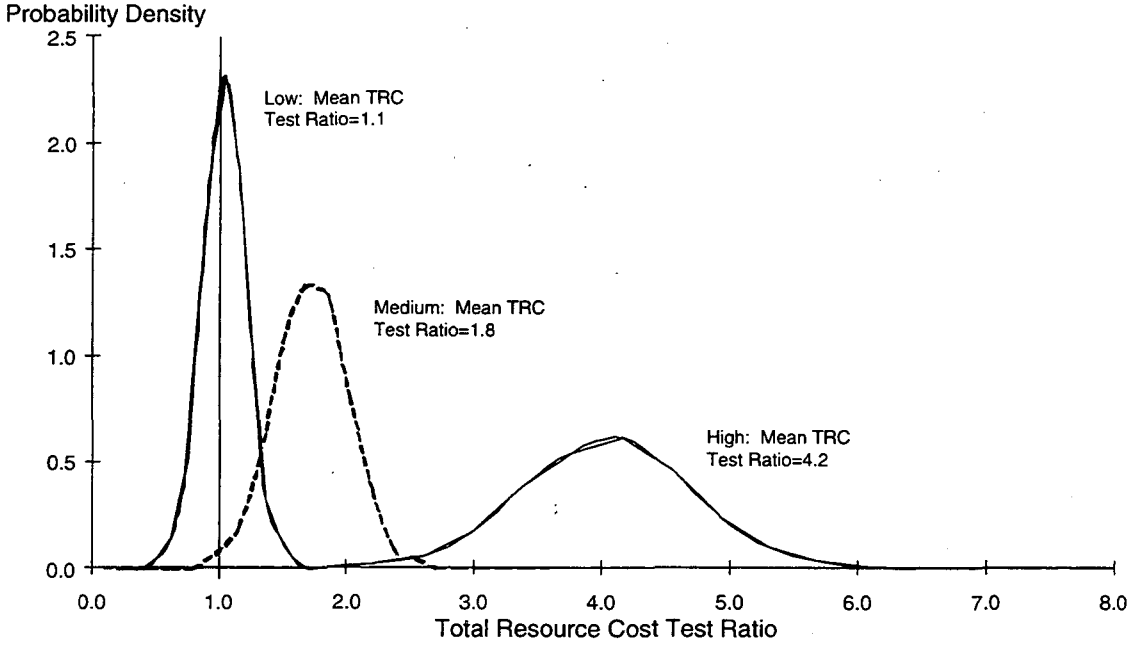


Table 7-4. Fraction of Distributions Representing Non-Cost-Effective Programs

Mean TRC Test Ratio	Savings Estimation Method	Precision	Percent of Distribution Less Than 1.0
Low (1.1)	End-Use Metering	Low (±50%)	40%
		Medium (25%)	29%
		High (10%)	11%
	Econometric	Low (15%)	19%
		Medium (10%)	11%
		High (5%)	3%
Medium (1.8)	End-Use Metering	Low	7%
		Medium	—
		High	—
	Econometric	Low	—
		Medium	—
		High	—
High (4.2)	End-Use Metering	Low	1%
		Medium	—
		High	—
	Econometric	Low	—
		Medium	—
		High	—

7.6.2. Implications of Estimate Imprecision

The Monte Carlo results summarized in Table 7-4 have important implications for the level of precision required to confidently assess DSM program cost-effectiveness. The answer to the question, "Is a 90/10 criterion necessary to confidently assess the cost-effectiveness of a DSM program?" is "No". Only for programs with mean TRC test ratios near 1.0 is a level of precision approaching 90/10 necessary to confidently determine whether the program is truly cost-effective. Even for the lowest precision evaluation, a program with a 'medium' mean TRC test ratio is cost-effective at the 90% confidence level.

Should these results change the way in which evaluations are conducted? We see two ways to proceed from this analysis: In the distribution of TRC test ratios from the DEEP sample of 20 commercial lighting programs, we observe that the majority of programs fall into the 'medium' category. Thus, in the majority of cases, a 90/10 criterion would be excessive for the determination of cost-effectiveness. It follows that a less stringent precision requirement should be adopted.

Alternatively, program planners and evaluators may have some previous estimate of the mean TRC test ratio associated with the program, perhaps based on a previous year's evaluation, or on program planning estimates. A determination of evaluation requirements could be made based on this estimate of cost-effectiveness: programs with preliminary cost-effectiveness ratios near 1.0 would be allocated additional evaluation resources to ensure a confident assessment of *ex ante* cost-effectiveness.

If cost-effectiveness verification were the primary goal of an evaluation, we would advocate a combined approach, whereby programs without preliminary or planning estimates of cost-effectiveness are allocated enough evaluation resources to assess cost-effectiveness for a program with a TRC test ratio in the 'medium' range, while programs with some cost-effectiveness information would be evaluated as dictated by these ratios.

A fundamental hurdle in this type of evaluation planning is our inability to estimate the cost of attaining a given level of evaluation precision. The programs for which we have been able to collect detailed evaluation data in our research represents too limited a sample to conclusively characterize the program attributes, participant characteristics, and evaluation method uncertainties required to understand the precision-evaluation tradeoff. Thus, the most practical and immediately applicable result of our analysis here is that a 90/10 criterion for relative precision of annual savings estimates is almost always excessive for determining cost-effectiveness.

7.7. The Effect of Bias on Cost-Effectiveness Estimates

Thus far, we have focused on the importance of precision in assessing cost-effectiveness. However, it is crucial, and potentially more important, to consider the role of bias as well. Despite the importance of estimate accuracy, our understanding of evaluation bias is less developed than our characterization of precision due to the difficulty of characterizing bias, which requires an independently estimated, unbiased estimate for comparison (i.e., the true, actual program savings). Our limited sample of program evaluations also hindered a more thorough characterization of evaluation method bias.

Just as imprecise savings estimates pose the greatest threat to programs with mean cost-effectiveness near one, those same programs may actually be non-cost-effective programs with biased estimates of savings. Table 7-5 reviews the biases identified in our research. The biases in Table 7-5 are given as percentage deviations from the unbiased value. A negative bias means that the cost of conserved energy is underestimated and a positive bias means that the cost of conserved energy is overestimated.

The effect of these biases is cumulative; a cost of conserved energy estimate based on limited duration metering and manufacturer estimates of measure lifetimes would be subject to multiple biases which could double or halve the cost of conserved energy. Some utilities implicitly acknowledge the bias inherent in their cost-effectiveness estimates by only implementing and continuing programs with a TRC test ratio significantly above 1.0, using, for example, a threshold of 2.0 or higher to screen programs.

Table 7-5. Sources of Bias in Cost of Conserved Energy Estimates

Parameter	Source of Bias	Magnitude of Bias in the Cost of Conserved Energy
Bottom-Up Savings Estimates	Seasonality of Hours of Operation	-5% to +5%
	HVAC/Lighting Interactions	+5% to +15%
	Nonrepresentative Metered Sample	unknown
Top-Down Savings Estimates	Engineering Estimate Uncertainty in SAE Models	+5% to +50%
Measure Lifetime	Use of Mfr. Estimates	-40% to -5%
Free Riders*	Free Riders Over Time	at least -11%
Free Drivers	Omission of Free Driver Savings	positive, but unknown

*Free riders are relevant only for utility cost test ratios, not TRC test ratios.

For TRC test ratios close to one, even a bias of a few percent could result in a non-cost-effective program being erroneously labeled cost-effective (or vice-versa, labeling a cost-effective program as non-cost-effective). However, by considering bias and precision together, the effect of bias is even more pervasive. A negative bias in the cost of conserved energy means that the distribution of the true TRC test ratio in Figure 7-2 is further to the left and closer to 1.0 than the supposed distribution. For programs with mean TRC test ratios close to one, a larger fraction of the distribution would move below the cost-effectiveness threshold of 1.0, revealing an increased probability that the program is not cost-effective.

If these biases are large enough or several negative biases are applicable, even a program with a mean TRC test ratio in the 'medium' range could, in actuality, be non-cost-effective. For example, a bottom-up metering study could meter equipment in the winter, overestimating annual hours of operation by approximately 5%, and savings could be coupled with biased manufacturer estimates of equipment lifetimes, overestimating lifetimes by as much as 40%. In the worst case, the combined bias could overestimate lifetime program savings by 45%, which would cause a marginally non-cost-effective program to appear to have a (biased) TRC test ratio approximately equal to 1.6.

Most of the evaluations for the 20 programs reviewed in Table 7-1 are subject to at least one of the biases listed in Table 7-5: (1) Metering studies that did not adjust for seasonality or interaction effects; (2) SAE models that used imprecise tracking database estimates of savings; (3) Measure lifetime estimates based on manufacturers' estimates of equipment operation, and; (4) Free ridership estimates which only discuss free riders in the first program year (relevant only for utility cost test ratio estimation).

Given the potential importance and pervasiveness of these biases in the current practice of evaluation, it seems prudent that some evaluation resources should be allocated to reduce bias, and not only imprecision, in the cost of conserved energy. In the next section, we discuss the potential costs of reducing these biases.

7.7.1. Implications of Estimate Bias

The preceding discussion demonstrates the significance of evaluation bias when assessing program cost-effectiveness. How can these biases be handled in the program evaluation? Ideally, the costs and potential impacts of each bias would be compared, and resources would be spent to identify and reduce the largest biases at the least cost. Because of the variability associated with the impacts of the biases, it is difficult to definitively prioritize the biases in order of their importance so that they can be addressed effectively given available resources. A larger sample of program evaluation data than is presently available is required to better characterize each evaluation method's biases. To begin to prioritize the treatment of the biases in evaluation, we present some qualitative estimates of the evaluation costs associated with reducing the biases.

Table 7-6. Estimates of the Cost of Addressing Biases in Commercial Lighting Evaluation

Source of Bias	Method Used to Reduce Bias	Approximate Marginal Cost
Seasonality of Hours of Operation	Seasonality Adjustment	Low
	Longer Term Metering	Med/High*
HVAC/Lighting Interactions	Metering of HVAC Equipment	High
	Modeling of HVAC/Lighting in Prototypical Buildings	Low/Med
Nonrepresentative Metered Sample	Proper Participant Stratification and Selection of Equipment to Meter	Med
		Low
Eng. Est. Uncertainty in SAE Models	Switch to non-SAE model	Low
Use of Mfr. Estimates of Lifetimes	Verify Continued Operation with Site Surveys	Med*
Free Riders Over Time	Analyze Equipment Sales to Nonparticipants During Life of Program Equipment	Med*
Omission of Free Driver Savings	Customer and Vendor Surveys	Low
	Analyze Equipment Sales in Diffusion Framework	Med/High*

*These methods require considerable additional time for the compilation of sufficient data.

Even with only rough guidelines regarding evaluation costs, we can draw some conclusions. Many of these biases can be at least partially addressed with minimal additional evaluation resources: Metered samples can be adjusted to control for seasonal effects and carefully stratified based on equipment, facility, and building zone characteristics; SAE models can be used only when tracking database estimates are of sufficient precision; and customer and vendor surveys can be used to obtain first-order estimates of free driver and spillover effects. Incorporating these changes into evaluation practice would improve the accuracy of estimates of cost-effectiveness of lighting programs at minimal additional cost. For those

evaluation improvements which require substantial commitments of time and money, a decision analytic framework, delineated in the next section, could be used to determine if the preliminary estimate of the TRC test ratio warranted additional efforts to reduce estimate bias.

The generalizability of information regarding these biases may also represent a justification for additional evaluation: If information regarding a bias from a particular evaluation can be used to estimate the magnitude of the same bias for other programs and evaluations, the cost of the additional evaluation is effectively spread among multiple programs.

As with estimate imprecision, we find that evaluation biases threaten the cost-effectiveness of programs with TRC test ratios closer to 1.0. Unlike imprecision, however, biases could threaten claims of cost-effectiveness for programs with TRC test ratios in the medium range (~1.8) as well. When imprecision is considered in addition to bias, reduced statistical confidence in even higher TRC test ratios may result. If the TRC ratios of our sample of 20 commercial lighting programs are representative of DSM in general, and all program evaluations are subject to biases and imprecision on the order of those described here, then as many as 55% of recently implemented programs, representing 50% of energy savings, could be erroneously labeled cost-effective as a result of ignorance of evaluation biases and incognizance of estimate imprecision.

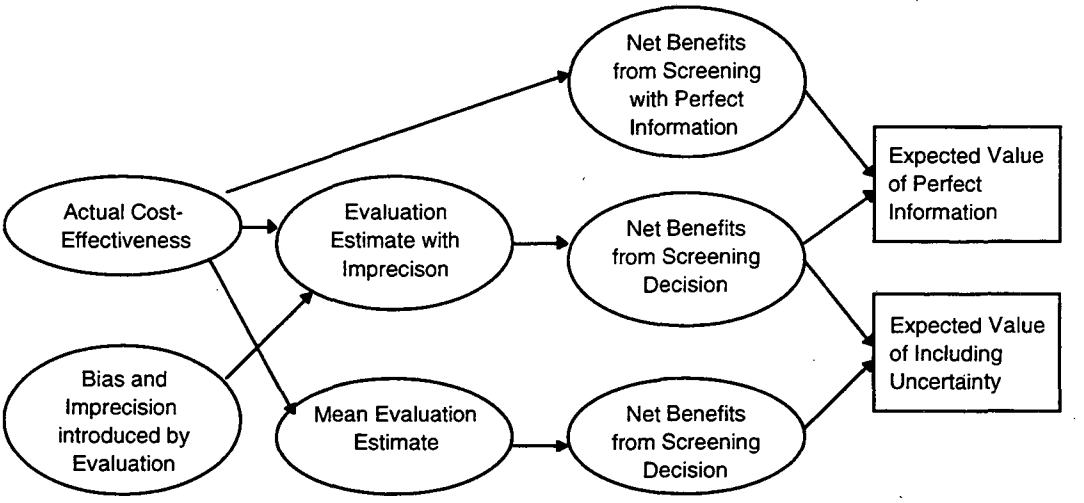
7.8. The Value of Correctly Assessing Cost-Effectiveness

In this chapter, we've discussed the role of precision and bias in assessing cost-effectiveness. The next logical step is to devise a method for the optimal allocation of evaluation resources. Ultimately, the cost of improving the precision and accuracy of evaluation method results should be traded-off against the value of obtaining increasingly accurate and precise estimates of program cost-effectiveness. A common use of cost-effectiveness information is program screening: ongoing programs are screened to determine if they should be funded for the next program year. In the following paragraphs, we briefly illustrate a procedure for estimating evaluation value and comparing it to evaluation cost. We present this decision analytic approach as an intriguing topic for future research, to demonstrate the applicability of these methods even when information on evaluation costs and evaluation results are not known with certainty.

A decision analytic approach to determining the appropriate level of additional evaluation to reduce imprecision and bias requires: (1) a subjective estimate of the chances that the program is actually non-cost-effective given any initial evaluation results, and (2) an estimate of the resources that would be (potentially) misallocated to the program in the following year (i.e., next year's program budget), when the decision to fund a program in the coming year is based on a) a mean evaluation estimate of cost-effectiveness, or b) an estimate of cost-

effectiveness that includes imprecision.[9] The difference of the benefits accrued between 2a) and 2b) is the value of including uncertainty of evaluation estimates in the program screening decision. The product of 1) and 2) is the expected value of future misallocated resources (also known as the expected value of perfect information), and would represent the maximum marginal evaluation expenditure justified if the resulting evaluation provided an estimate of cost-effectiveness with 100% certainty. A flow diagram for this analysis is presented in Figure 7-2.

Figure 7-2. Flow Diagram for Evaluation of Cost-Effectiveness for Program Screening



We have implemented a Monte Carlo model to estimate the expected value of including uncertainty and the value of perfect information while assuming a range of biases and imprecisions for the evaluation estimate. We have also parameterized our analysis into three ranges of evaluation-derived TRCs, shown in Table 7-2, representing the reported TRC ratios given in Table 7-1. We assume the program in question requires a $2,000,000 annual expenditure. If the evaluation estimate of cost-effectiveness is greater than one, the program will be funded in the next year. When imprecision is considered, if the program has a greater than 90% chance of being cost-effective, the program will funded in the next year.

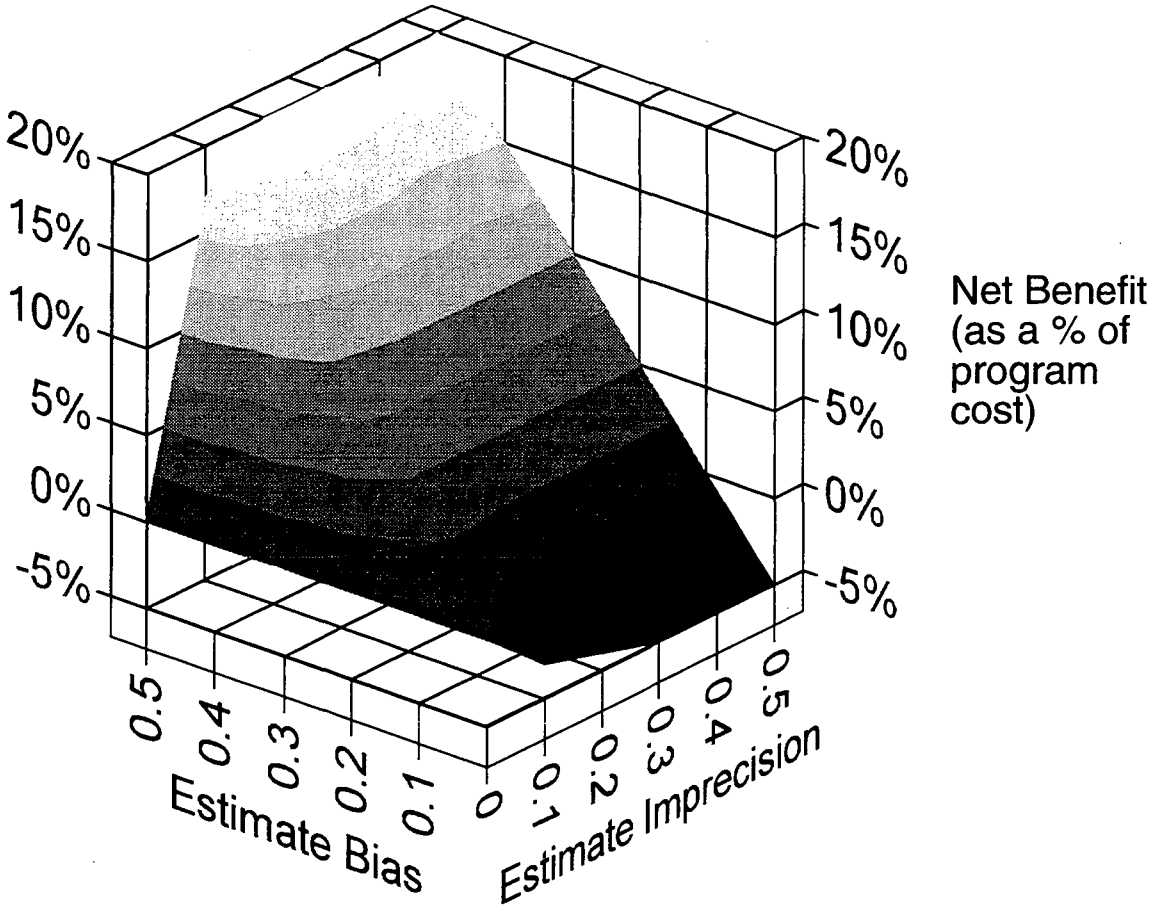
7.8.1. The Expected Value of Including Uncertainty

In this section, we present the value of including uncertainty (i.e., the imprecision around the evaluation estimate of cost-effectiveness) as a factor in the decision to

[9] A similar method is used to investigate utility planning uncertainties by Hobbs, Benjamin, & Prakesh Maheshwari (1990), "A Decision Analysis of the Effect of Uncertainty upon Electric Utility Planning", *Energy*, 15:9, pp. 785-801

rerun the program for another year. Results are given for a range of evaluation imprecisions and biases. Figure 7-3 illustrates the value of including uncertainty, as a function of total program cost, for the case of a 'Low' mean cost-effectiveness estimate.

Figure 7-3. Expected value of including uncertainty: TRC estimates in the low (mean=1.1) range



This value can be interpreted to be the additional evaluation expenditure justified, given the initial evaluation result is of the specified imprecision and bias, in order to obtain an estimate of the imprecision of the cost-effectiveness estimate. The value of including the evaluation result imprecision in the decision to continue program funding increases with increasing imprecision and increasing bias. When the imprecision is 0, as expected, the two decisions are equivalent and the value of including uncertainty is zero.

Because the evaluation cost-effectiveness estimates are in the low range, with a mean of 1.1., there is considerable value in estimating the imprecision of the evaluation estimate: an imprecise estimate can suggest the program has a high probability of not being cost-effective, even though the evaluation's point estimate is greater than 1. Also, imprecision without bias results in a negative benefit; this is due to the opportunity cost of cost-effective programs being canceled.

the value of including uncertainty for 'Medium' and 'High' mean cost-effectiveness estimates is negligible (with a maximum of 6% for 'Medium' and 0% for 'High'). As the reported TRC mean estimates increase in cost-effectiveness, the imprecision and bias must be greater to require the incorporation of imprecision into the program screening decision. In the high range of estimated cost-effectiveness ratios, adding imprecision information to the screening decision, regardless of the initial estimate's bias and imprecision, does nothing to improve the decision to rerun the program; no bad decisions are made even when estimate imprecision is ignored.

Calculating the expected value of including uncertainty allows the evaluator and program planner to place a rough estimate of value on one use of the evaluation imprecision information. Stated differently; we assert that evaluators who ignore the imprecision of their evaluation results pay a price that is dependent on the magnitude of estimate imprecision and bias, and can be quantified. Evaluators who underestimate imprecision, incur a smaller, but potentially large net cost.

7.8.2. The Expected Value of Perfect Information

The following three figures illustrate the expected difference in net benefits from a program when the screening decision uses an evaluation-derived estimate (which can be biased and imprecise) vs. when the true cost-effectiveness is used. If subsequent evaluation could inform the screening decision with perfect information, the evaluation would be worth the expected value of perfect information.

Figure 7-4. Expected value of perfect information: TRC estimates in the low (mean=1.1) range

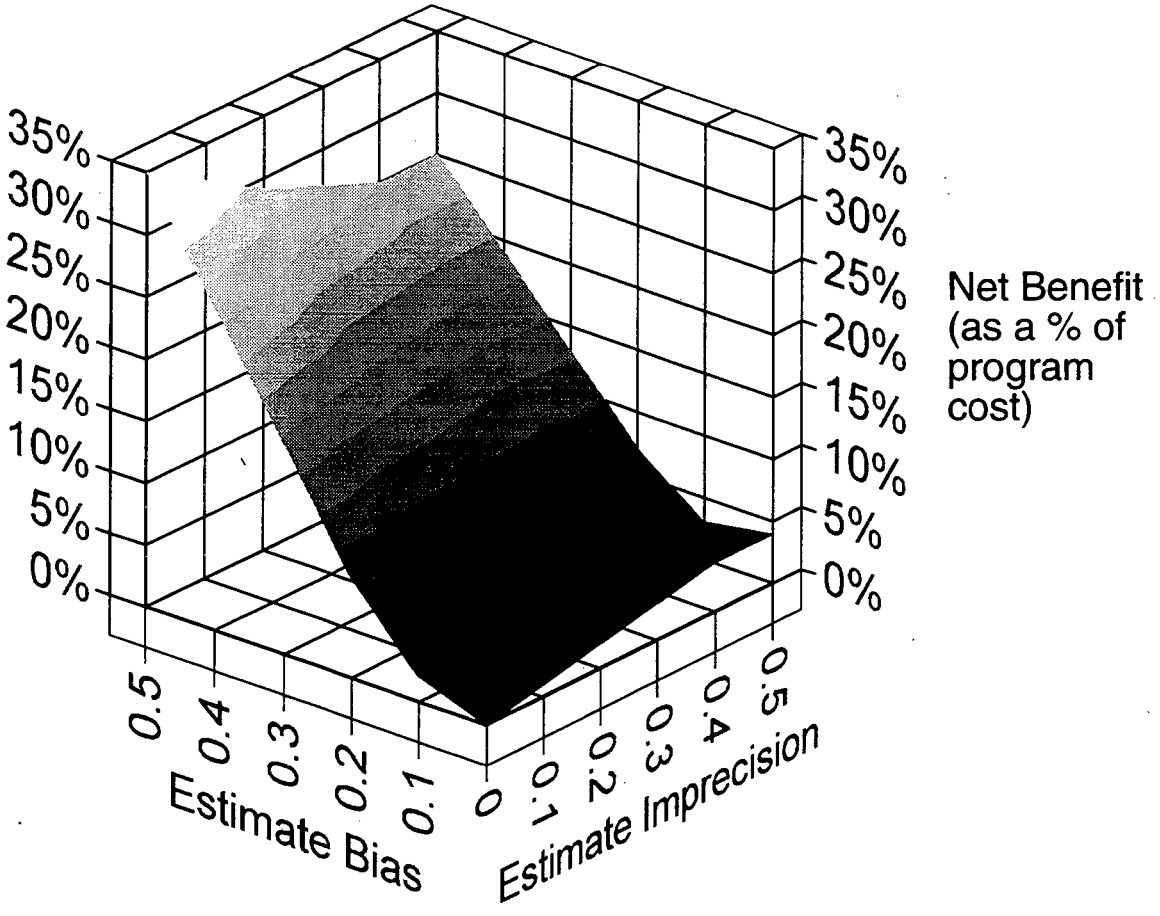
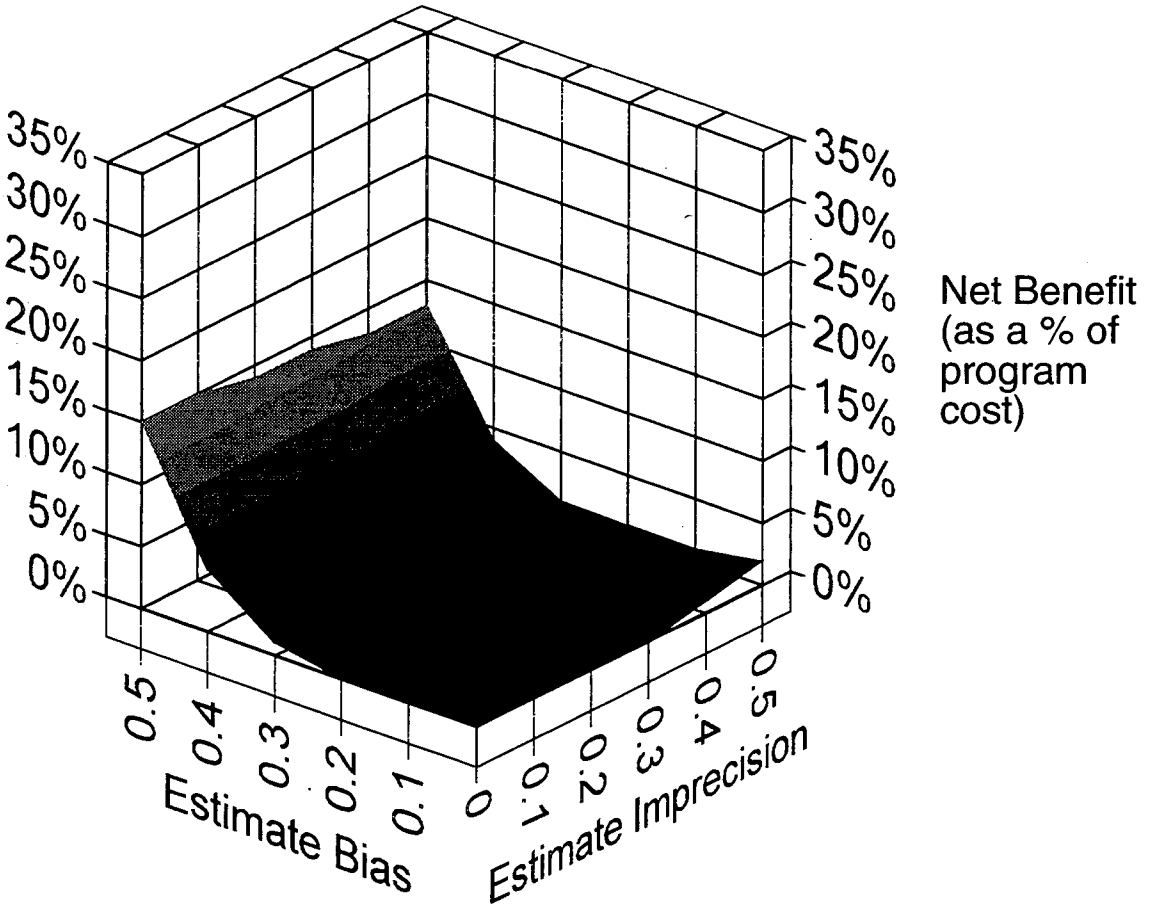


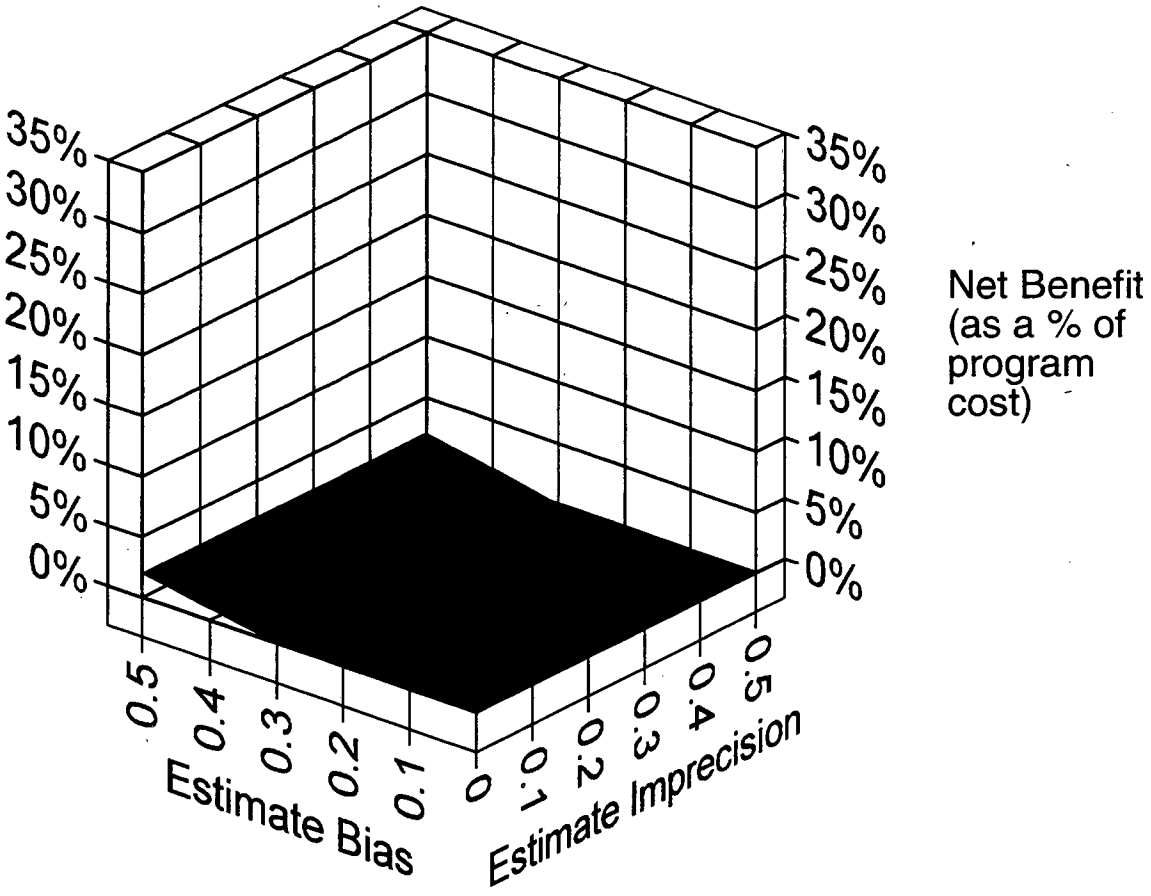
Figure 7-5. Expected value of perfect information: TRC estimates in the medium (mean=1.8) range



Figures 7-4, 7-5, and 7-6 illustrate the value of a *perfect* evaluation; however, it is unlikely that any evaluation techniques could provide a result utterly free of bias and imprecision. Thus, the net benefits in these figures represent an upper bound that could justifiably be spent on evaluation. An evaluation which reduced the bias by 50% could, roughly speaking, be allocated about half of the resources indicated in the above figures.

These figures make it clear just how critical the size of the initial, mean cost-effectiveness estimate is in determining the value of additional evaluation. While a screening decision based on a low cost-effectiveness estimate could be improved even if its bias and imprecision are minor, at higher cost-effectiveness levels the benefit of additional evaluation, even when bias and imprecision are significant, is slight.

Figure 7-6. Expected value of perfect information: TRC estimates in the high (mean=4.2) range



Using cost-effectiveness estimates for program screening and budgeting is just one example of how savings estimates are used and evaluation resources should be apportioned. Additional applications of evaluation information such as shared savings calculation, lost revenue recovery hearings, and load forecasting may justify more accurate and precise evaluation results, and may require a different selection framework. For example, when considering shared savings incentives earned by the utility, evaluation expenditures may be justifiably apportioned to programs with high TRC test ratios, because these programs can potentially provide the utility with the largest monetary rewards[10], as opposed to cost-effectiveness screening, where programs with lower TRC test ratios would justify increased evaluation resources. The appropriate level of evaluation expenditures could be set and justified by considering the value of evaluation using one, several, or all, of these applications of evaluation results.

[10]High TRC test ratios result in larger incentive awards with all other things (e.g., program size) being equal. A secondary effect, where larger programs usually have higher TRC test ratios and therefore larger shared savings incentives, also exists.

7.9. Chapter Summary

In this chapter we describe and implement a framework to assess the effects of bias and imprecision on estimates of program cost-effectiveness. The framework allows program evaluators and program planners to explicitly handle the uncertainties inherent in the complex evaluation of a DSM program. By estimating the effects of these uncertainties on estimates of program cost-effectiveness, program planners can ascribe confidence to their results and adopt levels of evaluation expenditures which are justified by the uses of the evaluation results. This is superior to the current practice in the industry, which overemphasizes the importance of method precision, ignores method bias, and does not base evaluation needs on information value.

Our implementation of this framework suggests that imprecision in the cost of conserved energy is significant for programs with mean TRC test ratios close to one, while higher ratios guarantee cost-effectiveness even with considerable estimate imprecision. A 90/10 criteria for precision seems excessive for most programs when screening for cost-effectiveness, in light of these findings.

However, bias in savings estimates can threaten the confidence of cost-effectiveness estimates for programs with ratios approaching 2.0, especially when estimate imprecision is also considered. Much of the contemporary concern with precision should be redirected to examine bias in evaluation estimates, given the results we present here.

Savings estimate biases and imprecision stem from a multiplicity of factors, some of which require expensive additions to evaluation procedures, and some of which require only slight changes in evaluation methods. While we recommend that all evaluations should include the least-cost methods to reduce estimate bias, additional expenditures should be traded off against the value of accurately assessing cost-effectiveness. The value of other evaluation information applications, such as demand forecasting and program improvement, require additional, explicit tradeoffs between information value and evaluation costs.

Implications for Future DSM Program Evaluations

As for me, all I know is I know nothing. Socrates, Phaedrus, sect 235

We began our study by suggesting that there is no single best evaluation method. Instead, we argued that what is best depends on a host of situation-specific circumstances that are independent of any particular method: What is the objective of the evaluation? How much money rides on the outcome of an evaluation? What is our level of confidence in the information we already have, prior to conducting an evaluation? We believe that formally addressing these questions is essential for choosing a best method and efficiently allocating evaluation resources.

Our thinking is based on two basic observations: First, evaluation methods vary greatly in cost, primarily due to the cost of various data collection strategies. Second, all evaluation methods are susceptible to error. Given that budgets are finite and methods are not equal, a utility must ration evaluation resources. We have developed and demonstrated a framework, drawn from decision analysis, for making these decisions systematically, based on the uses of evaluation information.

Implementing this framework requires explicit recognition of the errors associated with every evaluation method. Generally, this error is characterized as an imprecision around an evaluation result and, by assuming the error is normally distributed (i.e., with a bell-shaped curve), reported as a symmetric confidence interval around the point estimate. Unfortunately, we observe that reporting errors associated with evaluation methods is not common practice. Without this basic level of reporting, we believe it is difficult, if not impossible, to judge the value of the information actually obtained by an evaluation.

We strongly recommend future evaluations explicitly report and discuss the imprecision of their findings.

Imprecision should describe the uncertainty of the result based on the practical and theoretical limitations of the evaluation technique(s) used. For example, techniques that sample only a segment of the participant population are subject to some uncertainty based on the size and variability of the sample relative to the entire population. Calculation of imprecision can also involve subjective judgments, as in the case of persistence of savings throughout a measure's assumed measure lifetime: A subjective estimate of imprecision, based on program designer and evaluator expert judgment regarding persistence of savings over time, could be used to bound the annual savings estimate. What is important is that an effort be made to quantitatively estimate and communicate the limitations of the evaluation methods used. Assuming an estimate is thought to be accurate to +/-5% is very different from thinking the same estimate is accurate to +/-50%.

While consistent reporting of precision is an important first step in providing the information necessary to better judge evaluation findings, it is equally if not more important to recognize the limitations inherent in current methods for calculating precision. In current practice, estimates of precision are obtained via a series of straightforward calculations based on the (theoretical) statistical imprecision of the method and the variability in the data being evaluated. Most calculations correctly begin with the sampling (for end-use metering methods) or explanatory variable coefficient (in the case of regression models for billing analyses) imprecision, which is based on the variability and completeness of the data at hand. However, this procedure is insufficient to calculate the actual imprecision of an evaluation's result. We, thus, maintain that the imprecision calculated using such methods is better thought of as representing a lower bound on the actual imprecision of the savings estimate.

The estimate of precision represents a lower bound because of the possibility that the methodological assumptions (representative sampling of the population, explanatory variables in regression models measured without bias or excessive imprecision, etc.) on which the precision calculations are based may be flawed. When the estimate of precision associated with a savings estimate is taken directly from the sampling protocol, or regression model, this amounts to an implicit statement that the assumptions on which the method's viability is dependent are true with 100% certainty.[1] When estimates of measure persistence and free ridership are incorporated without concomitant (or at least subjective) estimates of precision, the stated imprecision of lifetime savings estimates should be seen as even more optimistic.

At this point, one may begin to question the value of pursuing the systematic assessment of evaluation trade-offs we advocate. That is, if imprecision is rarely reported and what is reported is known to be an underestimate, what advantages can our more formal approach offer over current more or less ad hoc methods? We believe there is one primary advantage: Increased defensibility, both to internal utility and external regulatory audiences.

Formally acknowledging and systematically incorporating what we do and do not know in committing evaluation resources ensures that the decision has taken full advantage of all available information. There is at least as much or, some would argue, far more value in acknowledging what we do not know as there is in presenting what we do claim to know. Remaining silent on what we do not know misrepresents the robustness of evaluation findings and consequently their defensibility.

[1]Deviation of program conditions from each method's assumptions can also result in biased estimates, which can be represented by an asymmetric estimate of precision.

Given the prospect of a more competitive utility business, we expect that there will be greater scrutiny of all future spending decisions; evaluation budgets will be no exception. Formal representation of what is and is not known in developing these budgets will allow for greater explicitness assessing the value of proposed evaluation activities. In principle, this explicitness should lead to better decisions.

Taxonomy of Evaluation Objectives

When discussing the general objectives of DSM program evaluations, program savings estimates are considered the objective of impact evaluations, and assessments of program administration, delivery, and customer satisfaction are considered the objective of process evaluations. Rather than distinguishing between process and impact evaluations, it is more instructive to examine the evaluation requirements for more specific objectives. Table A-1 summarizes the information requirements for different evaluation objectives.

Table A-1 includes evaluation requirements that are compiled in both process and impact evaluations. Both process and impact evaluations provide important information to program planners, evaluators, and regulators. While this research project has focused mainly on impact-oriented evaluation results, process evaluation information is equally essential to efficient, properly targeted DSM programs.

Process evaluations can provide insights on how to increase consumer satisfaction and market penetration of a DSM program. Process evaluations can also identify cost-cutting and efficiency measures associated with program implementation and delivery. Short-term, interim process evaluations can provide a sanity check for program implementers by pinpointing which aspects of a program are working as expected, and which are falling short of expectations. Evaluation results can also give insight into the equity effects of DSM programs. A careful analysis of who is participating in DSM programs and who is not participating is required to uncover cross-subsidization of DSM programs through the utility rate structure.

Impact evaluation results have several different uses: (i) They are used to plan the future of DSM programs; (ii) They are used to inform long range forecasts of demand and capacity requirements; and (iii) They are used in utility hearings to illustrate the magnitude of lost revenues, enabling utilities to adjust their rate structures or obtain PUC-promised incentives . The value of increasing evaluation expenditures is dependent upon the impact of the resulting information on planning, which in turn is dependent on the size of the DSM program. The larger the program and the higher the program's market penetration, the greater the impact of a miscalculation of savings on planning.

A.1. PUC Incentive and Cost Recovery

DSM has been impeded in many states due to regulatory structures that penalize utilities for implementing conservation programs by not providing cost recovery mechanisms. Conservation verification protocols are being established in many states to provide a means for utilities to verify program-induced conservation, and then recover lost revenues in addition to shareholder incentive payments. For these circumstances, it is not enough to know how much energy was conserved during a post-program period. The appropriate question is how much conserved

Table A-1. Taxonomy of evaluation requirements for different objectives

Requirement ↳ / Objective ⇩	Estimate of energy savings (impact measure)	Adjust estimates for technology failure/ misuse	Control for exogenous factors (changes in weather, price, facility use)	Adjust estimates for takeback effects	Adjust estimates for customer intent to install outside of program	Identification/ quantification of technology failure/misuse	Identification/ quantification of takeback effects	Analysis of customer satisfaction and adoption process	Assessment of program administration and delivery
Inform PUC rate, incentive, cost recovery hearings	✔	✔	✔	✔	✔	(if contested by regulators)	(if contested by regulators)	(if contested by regulators)	(if contested by regulators)
Prioritize program within DSM plan	✔	✔	✔	✔	✔			✔	✔
Inform demand forecasts	✔	✔	✔	✔	✔	✔	✔		
Identify methods to reduce program costs								✔	✔
Identify methods to increase number of participants								✔	✔
Identify methods to improve savings per participant						✔	✔	✔	✔

energy is specifically attributable to the DSM program in question. Thus, techniques which control for all exogenous factors should be used. Free riders must be estimated in order to adjust savings estimates to only include savings induced by the program.

A.2. Program Prioritization

Prioritization of programs requires an assessment of program costs and benefits. Once again, it is important to measure only benefits directly attributable to the program. Measurement of free riders is important; free riders reduce a program's net benefits, but they are also an indicator that the program may no longer be unnecessary or marketed to the wrong population. Analyzing customer satisfaction and adoption provides clues to future participation rates and can detect socioeconomic gaps in market penetration.

A.3. Demand Forecasts

The conservation estimates incorporated into most demand forecasts today are based on engineering models of energy consumption. Sometimes these models are adjusted to account for technological and behavioral idiosyncrasies uncovered during onsite surveys and end-use metering. Statistical models that are based on measured energy and demand savings, as opposed to conceptualizations of those savings, offer planners comprehensive estimates of electricity use reductions (together with a characterization of estimate uncertainty) resulting from previous conservation programs. Given the magnitude of most current DSM programs, however, there is little value gained in incorporating (relatively small) estimates of electricity savings in (relatively large) estimates of future demand.

A.4. Reducing Program Costs

Audits of internal program administration and delivery are useful in pinpointing program inefficiencies. An understanding of customer needs and attitudes can uncover cost-effective ways to interact with customers.

A.5. Increasing Program Participation

Understanding the customer and the effectiveness of different information delivery methods is crucial to obtaining program participation. Surveys of program participants and non participants can identify customers' economic, social, and physical barriers to program participation

A.6. Improving Per Participant Savings

Improving per participant savings requires a detailed understanding of customer-technology interactions following program intervention. Technology assessment in a simulated environment or in a sample of customers' dwellings, can provide important information regarding the suitability of particular DSM measures to a target market.

Description of Building Characteristics and Hours of Operation Variability

In this appendix, we detail the distributions used to construct our simulated dataset of 500 office buildings.

B.1 Building Construction Characteristics

In order to create a set of buildings which approximate the diversity of efficient equipment and construction materials and proportions in commercial buildings in the Northeast United States, I sample from a large number of distributions representing different building characteristics. EIA data was used to estimate a beta distribution of square footage. A 40,000 sq. ft. cutoff is used to distinguish between small and large offices, which have different equipment and construction characteristics. A floor is added to the building for each 20,000 sq. ft. The distributions describing the types and proportions of HVAC systems are from a soon-to-be-released LBL report from Ellen Franconi (See tables B.3 and B.4 in that report).

Distributions of building material R-Values, and lighting and equipment intensities have been taken from Akbari, H., J. Eto, S. Konopacki, A. Afzal, K. Heinemeier, and L. Rainer, *Integrated Estimation of Commercial Sector End-Use Load Shapes and Energy Use Intensities in the PG&E Service Area*, LBL-34263, December 1993. In this document building characteristics of a sample of ~75 large and ~75 small office buildings were tabulated. Each characteristic surveyed included min, max, mean, median, and the standard deviation, which provided enough information to estimate an equivalent beta distribution. Separate distributions are used for small (<40,000 sq. ft.) and large buildings, to distinguish between common characteristics of small buildings and those of larger buildings.

The distributions for all parameters used to generate buildings and building energy savings for the DOE2 simulation are given in Table B-1.

Table B-1. Distributions for DOE2 input parameters

Variable	Office Size	Distribution	Min	Max	Parameters
Building size (sqft)		Beta	1,000	200,000	2,8
sqft_person	Small	Beta	3.6	1260	1.5,4
sqft_person	Large	Beta	20.7	1145	1.5,3
ach		Beta	0.7	1.0	2,3
wall r value	Small	Beta	1.6	16.4	1.2,3
wall r value	Large	Beta	1.3	23.2	3,8
window r value	Small	Beta	1.1	2.0	4,6
window r value	Large	Beta	1.1	2.8	4,9
window shading coefficient	Small	Beta	0.6	0.85	2,8
window shading coefficient	Large	Beta	0.6	0.85	7,6
roof r value	Small	Beta	1.8	51.1	2,9
roof r value	Large	Beta	1.7	27.7	2,9
equip intensity	Small	Beta	0	6.8	1.3,3
equip intensity	Large	Beta	0.1	10.5	1.3,8
glass/wall ratio	Small	Beta	0	0.65	3,6
glass/wall ratio	Large	Beta	0	0.95	1.6,1.9
HVAC type	Small	Poisson	30% gas H_20 50% packaged gas heat 20% electric resistance heating		
HVAC type	Large	Poisson	30% Multizone heating w/economiser 30% Variable air volume system 40% Fan Coil		
Weekday Hours	Both	Truncated Beta	6	18	1.6,1.9
Weekend Hours	Both	Truncated Beta	1	8	1.5,3
lighting intensity	Small	Beta	1	7.5	2,8
lighting intensity	Large	Beta	0.5	4.6	1.5,2
% Measures changed	Small Effect Size	Beta	10	50	1.6,3
%W/measure saved		Beta	20	50	1.6,3
% Measures changed	Medium Effect Size	Beta	30	80	1.6,3
%W/measure saved		Beta	20	60	1.6,3
% Measures changed	Large Effect Size	Beta	60	100	1.6,3
%W/measure saved		Beta	20	60	1.6,3

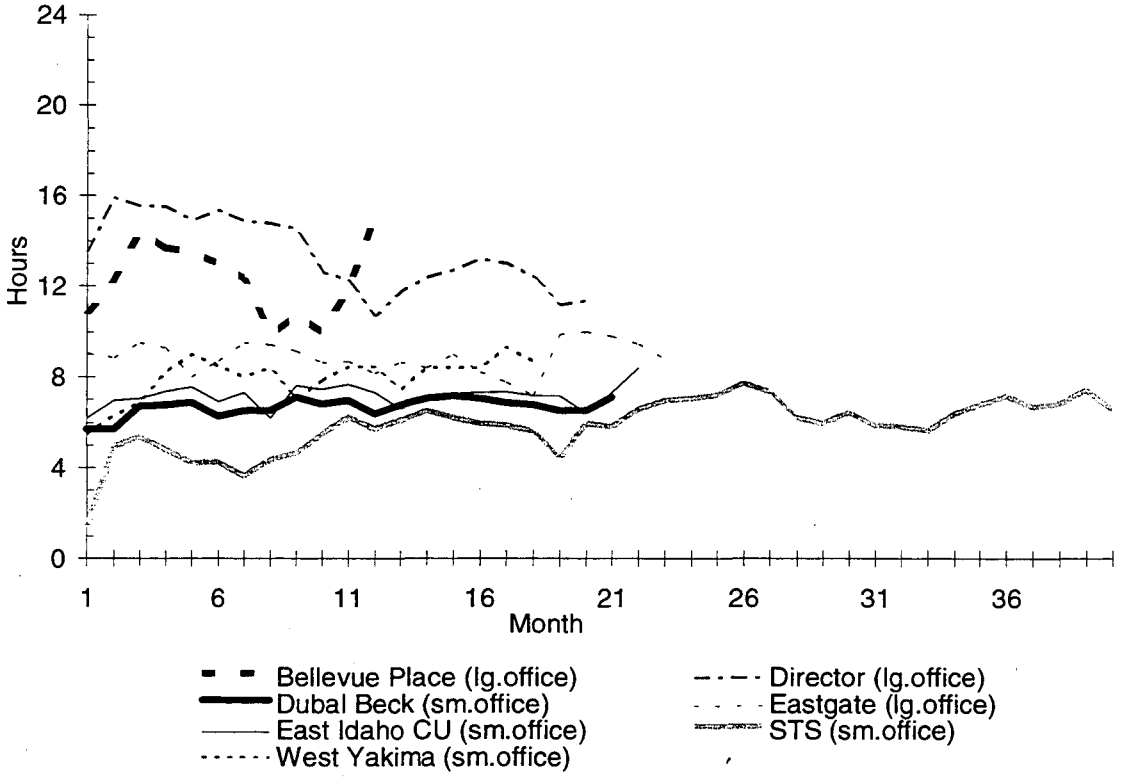
B.2 Changes in Building Hours of Operation over Time

A base for daily weekday and weekend hours of operation are sampled from two separate beta distributions. The resulting values serve as the mean for two normal distributions (calibrated to stochastic variability over time in the Energy Edge[1] data), used to generate 24 more values to represent daily weekend and weekday hours for each month. These 24 values are input to DOE2 as the daily hours of operation, specific to each month.

Energy Edge data on 13 buildings in the Pacific Northwest provide 6-30 month time series data on monthly full load hours of lighting equipment operation. From this small dataset, we can make a rough estimate of variability in hours of operation for a 'stable' office building and variability for a less stable office building. A summary of the Energy Edge time-series data is given in Figure B-1.

By applying 'stable' and 'unstable' hours of operation schedules to the buildings simulated by DOE2, we can estimate the effects of varying hours of operation on

Figure B-1. Weekday Hours of Operation for Energy Edge Offices

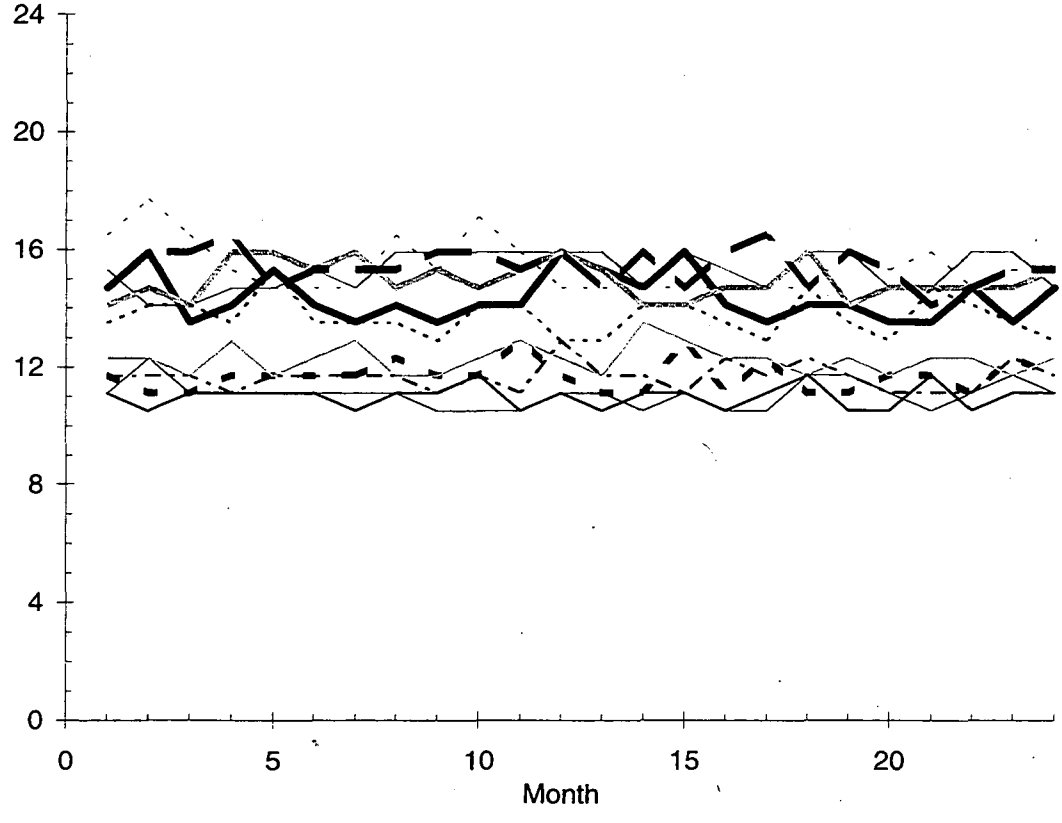


[1] Electric load data was supplied by Energy Edge project leader Mary Ann Piette of LBL. For information on the Energy Edge Project, see Piette, M.A., Diamond, R., Nordman, B., deBuen, O., Harris, J., Heinemeier, K., Janda, K., "Final Report on the Energy Edge Impact Evaluation of 28 New, Low-Energy Commercial Buidings, Lawrence Berkeley Laboratory, Berkeley, California, LBL-33708, February 1994.

each evaluation method's ability to estimate program savings. A range of effect sizes, from 3% to 15% electricity savings, will be used to see if hours of operation changes can prevent some evaluation methods from observing proportionally smaller savings. Figure B-2 depicts the simulated hours of operation for a handful of buildings in the simulation dataset. Hours are simulated as being distributed with a beta distribution around a base value for each building. The base values are normally distributed. Average hours are lower in Figure B-2 due to differences in the calculation of full load hours and simple hours.

There is a fundamental difference between the observed hours of operation in Figure B-1 and the simulated hours in Figure B-2: data in Figure B-1 suggests that hours could be better represented by a random walk, rather than a distribution around a mean value, as in Figure B-2. Future iterations of this analysis could incorporate a random walk function to simulate hours of operation that more accurately mimic the Energy Edge data.

Figure B-2. An Example of Simulated Hours of Operation



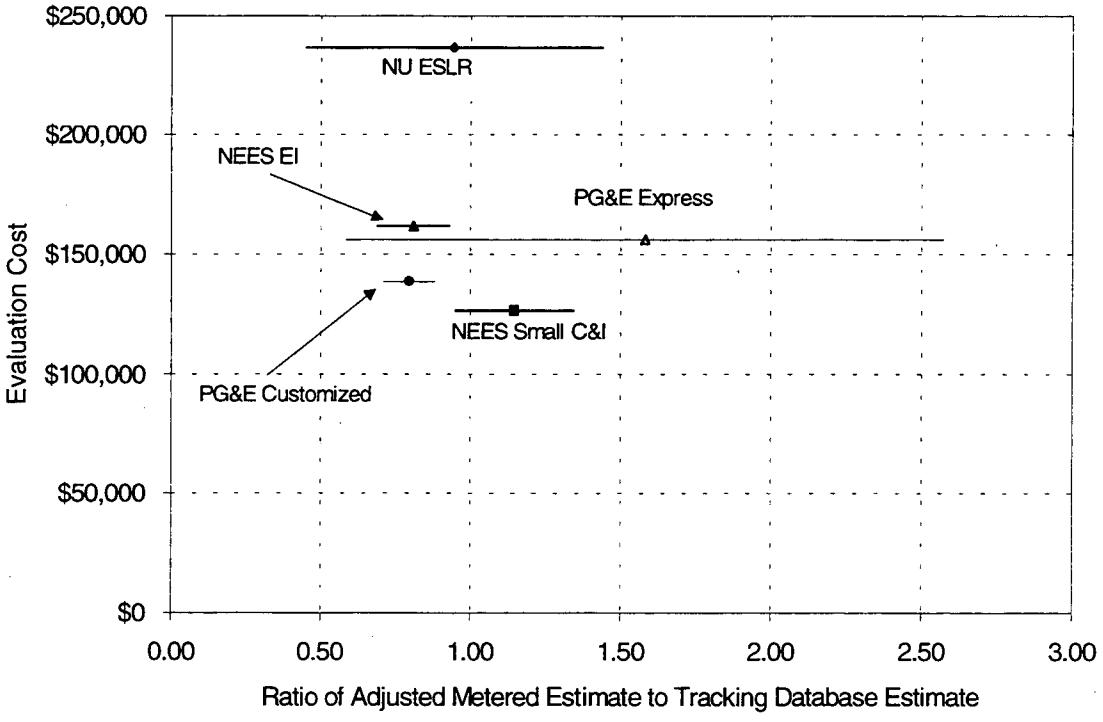
Comparing Annual Savings Estimates from Bottom-Up and Top Down Methods

Chapters 3, 4, and 5 discussed the accuracy and precision associated with bottom-up and top-down estimates of savings. In this appendix we compare these results and discuss some of the additional trade-offs which should be considered when selecting evaluation methods. We also discuss hybrid methods for calculating annual energy savings which combine results from multiple evaluation methods.

C.1 Comparing Costs and Results of Top-Down and Bottom-Up Methods

Based on the analysis in Chapter 4, we derived estimates of metering cost and precision for five different program evaluations. The results of that analysis are summarized in Figure C-1.

Figure C-1. Summary of Bottom-Up Method Results



The precision ranges from ± 66% for PG&E's Express program to ± 18% for PG&E's Customized program. If the metered sample in each evaluation is representative of the total participant population, we can assume these estimates are unbiased. The cost of these evaluations is between $100,000 and $250,000, and varies based on the metered sample size (which affects metering data

collection and analysis costs) and the total number of participants (which affects tracking database collection and analysis costs).

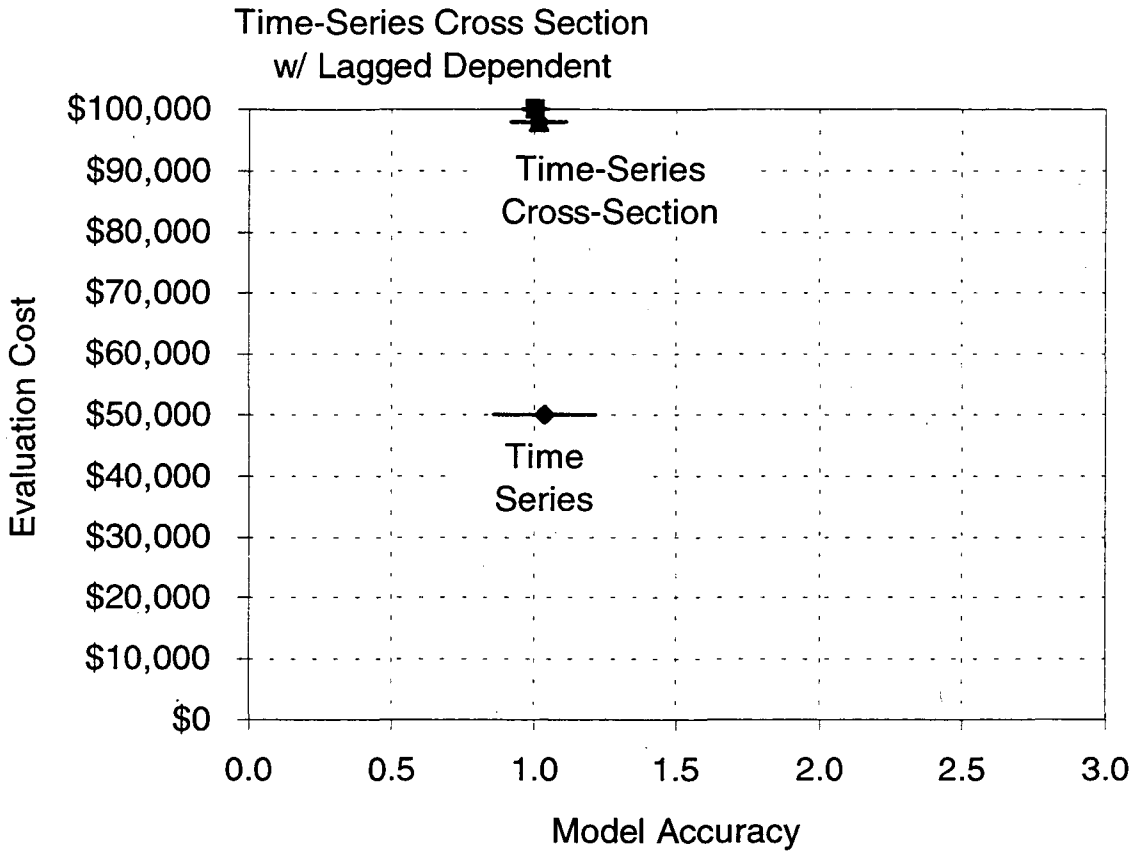
Because of our small sample of metering studies and tracking database information, we have insufficient data to confidently characterize the variability in the precision of metered estimates of annual savings. Differences in precision among the end-use metered studies appear to be due to two factors:

1. Quality in the program tracking databases. If the tracking database is imprecise or inconsistent with its estimate of savings at each site, a metering study which uses the tracking database estimates of savings to extrapolate to the entire population will also be imprecise. As demonstrated by the NEES and NU data, tracking database accuracy can vary dramatically from program to program. Utilities who review and compare their tracking databases to measured evaluation results, however, can iteratively improve their tracking databases' accuracy and precision.
2. Heterogeneity in the sample of participants, in their patterns of energy consumption, and in their selection of program measures. These factors complicate the extrapolation from a sample to the entire population. They also complicate the calculation of accurate tracking database estimates of savings for each site.

The average precision of the five studies we examined (measured as a standard deviation around the mean value after adjustments for interaction effects and limited duration metering) is about ±25%, the least precise metering studies result in precisions of ±50% or wider, and the most precise studies have precisions approaching ±10%. Based on the limited data available, these numbers represent our best estimates of the precision of end-use metering efforts to estimate annual savings.

Figure C-2 reviews the results of the econometric methods from Chapter Five. The costs we estimate for these methods are much lower than the estimated costs of the metering-based evaluation methods. Given our simulated set of buildings and monthly consumption data, the time-series and cross-sectional/time-series methods performed well, with almost little bias (i.e., close to 100% accuracy) and precision ranging from ± 18% for the time-series regression (utilizing only participant data) to ± 5% for the time-series, cross-section regression with a lagged dependent variable. Given our assumptions regarding the accuracy and precision of a tracking database (from data in Chapter Two), the SAE models require site inspection-based estimates of savings to perform well, which increases their cost to over four times the cost of the top-down methods in Figure C-2. Thus, we omit the SAE methods from this comparative analysis.

Figure C-2. Summary of Top-Down Method Results



If the simulated building and monthly consumption dataset accurately mimics a real-world dataset, the top-down methods provide more precise estimates of program savings than the bottom-up methods based on end-use metering results. Moreover, the top-down, billing data based methods are about half as expensive as the metering data based methods.

If cost, accuracy, and precision were the only factors worth considering when selecting an evaluation technique, billing data based top-down methods would be the obvious choice. However, there are several qualitative differences in bottom-up and top-down methods. These differences should also enter into the evaluation selection process. The first three are benefits of metering studies, and the remaining two describe benefits of billing analyses.

1. Metering can be performed just prior to and immediately following equipment installation, so that initial estimates of annual savings can be obtained in a matter of months. Billing data based methods require many months of consumption data (here we use 12) to estimate annual savings. If an estimate of program savings is needed in the very short-term (e.g., if a decision

regarding next year's program budget needed to be made, based on the success of this year's program), a metering study may be worth the additional cost.

2. Billing analyses provide no insight into the reasons for program success or failure. If a billing analysis detects an insignificant program effect, additional site inspections or communication with participants would be required to determine what aspect of the program resulted in a small estimate of savings For example, was it due to an inappropriate or defective technology being installed, a change in participant usage patterns as a result of participation, or an inability of the billing analysis to separate the true program effect from other factors?

3. Metering of equipment can provide not only total kWh savings, but time-of-use and kW load savings as well. This information can be valuable for demand forecasting and calculation of DSM program benefits.

4. The obtrusive requirements of metering studies (repeatedly visiting participant facilities to survey, install, maintain, and remove metering equipment) preclude their use in some programs. In addition, some programs' delivery mechanisms do not provide evaluators with an opportunity to meter equipment at a facility before the efficient, program-subsidized equipment is installed (e.g., a rebate program where participants apply for a rebate when they purchase and install the efficient equipment).

5. By incorporating data from a comparison group of nonparticipants, billing analyses can control for changes in electricity consumption based on non-programmatic factors, such as a downturn in the economy or changes in the weather. Metering studies do not usually make adjustments of this type, because the expense of incorporating metered data from a group of nonparticipants would dramatically increase the cost of the evaluation.

In the next section, we review some techniques which combine the results of multiple methods in an effort to address qualitative and quantitative shortcomings associated with using a single method to estimate annual program savings.

C.2 Hybrid Methods

Because no single method provides both an accurate estimate of program savings as well as a quantification of individual factors that affect savings, strategies that combine the results of multiple evaluation methods are quite useful. Such evaluation strategies enable evaluators to increase the statistical precision of their savings estimates and enhance their understanding of program strengths and weaknesses. The complexity of interactions among the utility, the program delivery, the program technologies, and the participants suggests that evaluation would benefit from holistic approaches incorporating methods from a multitude of evaluation perspectives. Different measurement and evaluation techniques can be used to verify each other and generate composite estimates with improved precision.

At this time, most utilities at least implicitly acknowledge the complementary roles of different evaluation techniques. For example, tracking database estimates of savings based on auditor inspections of installed equipment are used until end-use metering data are available. A combination of end-use metering data and tracking database estimates are used until a billing analysis based on monthly energy consumption data is available. Thus the savings estimate is continually refined based on the latest information. At issue here is the formalization of this process through explicit recognition and prioritization of various evaluation techniques over a multiyear time horizon.

C.2.1 Augmenting Billing Analyses with Verification Studies

In order to alleviate problems of credibility associated with billing analyses, a recent trend in utility DSM is to some utilities complement their econometric analyses with limited sample, short-term metering studies, called verification studies. The cost of these verification studies has been falling rapidly with the advent of technological innovations in metering and run-time data collection equipment.

For utilities that can afford both a billing analysis and a verification study, the combination of methods provides both top-down and bottom-up estimates of annual savings, which complement each other well. The billing analysis incorporates a wealth of information on participant and nonparticipant consumption characteristics and can control for HVAC/Lighting interactions. The metering study can verify installation and efficient operation of individual measures at a representative sample of sites.

C.2.2 Triangulation

While most utilities use multiple methods to verify savings estimates, a few utilities use weighting schemes to combine estimates of savings and generate a single, more robust estimate. Some utilities refer to these algorithms as triangulation. The simplest form of triangulation involves calculating a weighted average of the different estimates, weighting each estimate of savings by inverse of its variance.

A more complicated weighting scheme might combine estimates using Bayesian techniques, where the weight of each estimate is based on a (predetermined) subjective judgment of the evaluator's confidence in each method.

C.3 Taxonomy of Evaluation Methods and Utility Evaluation Strategies

The diversity of impact evaluation techniques used in current practice is illustrated in Table C-1. One of the most important distinctions demonstrated in this taxonomy is the distinction between methods that implicitly account for different factors that affect savings and methods that allow one to explicitly quantify the effects of those same factors. For example, site inspections allow evaluators to discover explicitly the number of sites at which efficient equipment was removed or malfunctioning. A billing analysis automatically (implicitly) accounts for

removed and malfunctioning equipment since this equipment does not contribute to savings. But the evaluators conducting the billing analysis are unaware of precisely *why* measured savings are lower than originally estimated; they only see the reduced estimate of savings (often in the form of a ratio of measured consumption and tracking database estimates of program savings).

Table C-1. Taxonomy of Impact Evaluation Methods Used in Commercial Lighting DSM Programs

Attribute ↳ / Evaluation method ⇩	Implicit Accounting of Attributes in Savings Calculations				Explicit Examination of Program Attributes		
	Adjusts for technology failure/ misuse[1]	Controls for exogenous factors[2]	Adjusts for take back effects	Adjusts for free riders and other selection biases	Identifies/ quantifies technology failure/ misuse	Identifies/ quantifies take back effects	Examines customer satisfaction and adoption process
Tracking estimate							
Tracking estimate with hours of use verification			Partially			Yes[3]	
Tracking estimate with site inspections	Yes				Yes	Yes[3]	Yes
Tracking estimate with short-term metering	Yes	Partially	Yes		Yes	Yes	
Bill comparison of participants / nonparticipants	Yes	Partially	Yes	Partially			
Billing analysis (regression of consumption data)	Yes	Yes	Yes	Yes[4]			
Statistically adjusted engineering analysis (SAE)	Yes	Yes	Yes	Yes[4]			
Logit model evaluating participation decision				Yes (explicitly quantifies)			

[1] Technology failure/misuse includes participant failure to install, participant sabotage

[2] Exogenous factors include weather, business and structure characteristics, and fuel prices

[3] If performed both before and after measure installation

[4] Only with the appropriate control group

Errata for *A Framework for Improving the Cost-Effectiveness of DSM Program Evaluations* LBL-37158

On page 70, Table 5-5 lists costs for 500 sites, not 250 as stated in the text.

On page 84, Table 6-6 should read as follows:

Table 6-6. Cost of Conserved Energy for a Hypothetical Commercial Lighting Program

Bottom-Up Precision	Mean (¢/kWh)	Median (¢/kWh)	Standard Deviation	90% Prediction Interval
Poor	4.8	4.0	± 44.	—
Average	4.4	4.0	± 1.6	± 60%
Good	4.1	4.0	± 0.67	± 27%

LAWRENCE BERKELEY LABORATORY
UNIVERSITY OF CALIFORNIA
TECHNICAL AND ELECTRONIC
INFORMATION DEPARTMENT
BERKELEY, CALIFORNIA 94720