

A Review of Commercial Sector Lighting Program Evaluation

Richard Sonnenblick and Joseph Eto
*Lawrence Berkeley Laboratory
Berkeley, CA*

The state of the art in evaluation tools is evolving rapidly as utilities, consultants, and academics apply techniques from economics, statistics, and engineering to the task of assessing demand-side management program methods and estimating net impacts. However, the state of the art is not mirrored by current practice. The 20 commercial lighting programs investigated in this article provide an opportunity to examine the recent practice of evaluation methods in the field. We begin with descriptions of the specific methods used by programs in our sample to evaluate program savings. The discussion of annual savings estimation is followed by a discussion of free ridership and of persistence of savings and its verification. We then introduce a taxonomy of evaluation methods that encapsulates the strengths and weaknesses of these annual savings evaluation methods. Finally, we discuss the need for evaluation strategies that use multiple evaluation methods and span multiple program years.

Evaluating the effects of a demand-side management (DSM) program on energy consumption is a daunting task. The goal is to measure how much energy would have been consumed by program participants if the program had not occurred. Because program savings can only be deduced and not directly observed, uncovering savings attributable to a program often uses quasi-experimental methods, which include information on both program participants and nonparticipants (a comparison group) both before and after program implementation. The state of the art in evaluation methods is evolving rapidly as utilities, consultants, and academics apply techniques from economics, statistics, and engineering to the task of assessing DSM program methods and estimating net impacts. In this article, we report on the impact

evaluation methods used by the 20 commercial lighting DSM programs listed in Table 1. The design, performance, and utility evaluation of these programs have been analyzed by Lawrence Berkeley Laboratory's Database on Energy Efficiency Programs (DEEP) project (Eto, Vine, Shown, Sonnenblick, & Payne, 1994). The 20 programs we assessed provide an opportunity to examine the recent practice of evaluation methods. More complete, technical descriptions of these evaluation methods can be found elsewhere (Hirst & Reed, 1991; RCG/Hagler Bailly, Inc., 1991).

Some bias may be present in our results in that utilities in some regions of the country were not willing to provide us with evaluation results from their DSM programs, making it impossible to present a complete assessment of current practice. In addition, we made no attempt to represent every state, utility, or region in our sample. Our assessment reports on the current practice of utilities that cooperated with our requests for information, and reported program cost and savings information. Whereas conclusive results from our nonrandom sample of 20 programs would be premature, our report represents a first step towards more comprehensive assessment of utility program cost and performance.

In our sample, program maturity ranged from 1st-year pilot programs to programs with 14 years of experience. Sixteen of the programs offered rebates to customers and four programs offered both the lighting equipment and installation at no cost to the customer. Programs included all manner of indoor lighting and lighting control equipment used in the commercial section, and a range of large and small commercial (and often industrial) customers were eligible for participation in the various programs. Complete information on the sample can be found in Eto et al. (1994).

In this article, we compare evaluation methods based on billing data used by 10 of the programs and end-use metering methods used by four of the programs. We examine the range of techniques used to estimate the proportion of free riders participating in each program. We also review the handful of programs that investigate the magnitude of market transformation effects. We analyze the effect of different measure-lifetime estimates on total resource cost. We then introduce a taxonomy of evaluation methods that encapsulates the strengths and weaknesses of these program evaluation methods in meeting different evaluation objectives.

CLASSIFYING EVALUATION METHODS IN THE SAMPLE OF 20 PROGRAMS

The distinction between "engineering" and "measured data" evaluation methods figures prominently in most discussions of program evaluation results. In recent years, conventional wisdom has held that engineering methods are inferior to methods relying on measured consumption data. We find this distinction misleading both in theory and in practice for the following reasons: (a) all methods of estimating energy savings rely on detailed data collection, which is always subject to some degree of stochastic and/or

TABLE 1
Summary of Commercial Lighting Programs in Sample

Utility	Abbreviation	Program Name	Program Year(s)	Evaluation Methods Used	Gross Post-Program Savings (GWh)
Boston Edison Company	BECO	Small C/I Retrofit	1991	TE, BA	8.3
Bangor-Hydroelectric Company	BHEC (pilot)	Pilot Commercial Lighting Rebate	1986-1988	TE	2.8
Bonneville Power Administration	BPA (pilot)	Industrial Lighting Incentive	1986-1988	TE	3.2
Central Hudson Gas and Electric	CHG&E	Dollar Savers Rebate	1990-1991	TE, SAE	16.1
Central Maine Power	CMP	Commercial Lighting Retrofit Rebate	1992	TE, EU, SI	15.7
Consolidated Edison	ConEdison	C/I Efficient Lighting Program	1991	TE, SAE	91.9
Green Mountain Power	GMP	Large C/I Retrofit	1992	TE	1.4
Green Mountain Power	GMP	Small C/I Retrofit	1992	TE	4.0
Iowa Electric Light and Power	IE (pilot)	Lighting Payback Plan	1990	TE	1.4
New England Electric System	NEES	Energy Initiative	1991	TE, EU, SAE	104.2
New England Electric System	NEES	Small C/I	1991	TE, EU, BA	23.5
Niagara Mohawk	NMPC	C/I Lighting Rebate	1991	TE	134.4
Northeast Utilities	NU	Energy Saver Lighting Rebate	1991	TE, EU, SI, SAE	149.8
New York State Electric and Gas	NYSEG	C/I Efficient Lighting Rebate	1991	TE	71.5
Potomac Electric Power Company	PEPCO	Commercial Lighting Rebate	1990	TE, BA	40.5
Pacific Gas and Electric	PG&E	C/I/A Rebate: Direct Rebate	1992	TE, EU, BA	130.0
Southern California Edison	SCE	Energy Management Hardware Rebate	1992	TE	96.6
Seattle City Light	SCL (pilot)	Commercial Incentives Pilot	1990	TE, BC	16.9
San Diego Gas and Electric	SDG&E	C/I Lighting Retrofit	1992	TE, BA	66.2
Sacramento Municipal Utility District	SMUD	Commercial Lamp Installation	1988	TE	2.6

Note. C/I = commercial/industrial; C/I/A = commercial/industrial/agricultural; TE = tracking database estimate; BA = billing data analysis using regression model; SAE = statistically adjusted engineering estimate; EU = end-use metering; SI = site inspection; BC = billing comparison.

systematic error; (b) all methods of estimating energy savings rely on the same basic engineering principles that postulate the existence of energy savings (thus, all methods could be considered engineering methods); and (c) a trend in utility regulation is encouraging evaluators to incorporate ex-post and participant information in their estimates of savings. Many evaluators now make use of both engineering-based assumptions and measured data to estimate lifetime program savings, blurring the distinction between pure engineering and measurement-based evaluation methods. Thus, to say that a clear distinction separates evaluation methods used in practice, or that measured data represent the truth whereas engineering methods are inherently flawed, is an oversimplification of both terms.

We distinguish between three general categories of postprogram impact evaluation methods, all of which incorporate some form of "measured" or observed postprogram information:¹ (a) tracking database and site inspection estimates, (b) consumption estimates using billing data, and (c) consumption estimates using end-use metering. These three categories are not entirely distinct; evaluation methods exist that span two or three of these categories. We believe these categories better describe the methodological distinctions among evaluations than do the terms of *engineering* and *measured* evaluation. The taxonomy of evaluation methods presented later in the article summarizes available methods and describes each method's ability to identify and control for different components of program savings. In the next three sections, we discuss tracking database, billing analysis, and end-use metering methods for estimating annual savings.

TRACKING DATABASE ESTIMATES OF PROGRAM SAVINGS

The most straightforward attempt to determine energy savings uses program tracking database information on participants' installed measures and three pieces of additional information: each measure's operating efficiency, the baseline efficiency of the measure to be replaced, and the annual hours of operation. The sophistication of the estimate is dependent on the sources of these three values. As noted, substantial amounts of postprogram information (short of measured consumption data) may be used in this postprogram evaluation method. In this regard, tracking database savings estimates are anything but unverified, preprogram, engineering estimates. The following subsections review the components of a tracking database estimate of annual savings.

¹Although we acknowledge the complementary nature of impact and process evaluations, the evaluations provide little evidence of formal information sharing between the two regimes. This puts utilities at a disadvantage because process evaluation information is often relevant in discussions of program savings.

Baseline Equipment Efficiency and Program Measure Efficiency

The efficiencies of both the new equipment and the equipment being replaced are 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. In San Diego Gas and Electric's (SDG&E) 1992 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. SDG&E revised their savings figures downwards for various measures by 18% to 48% to reflect more efficient base equipment. Other programs that relied on tracking database estimates such as those of Iowa Electric and Sacramento Municipal used similar assumptions to estimate the efficiency of existing equipment.

End-use metering studies by New England Electric System (NEES), Northeast Utilities (NU), and Pacific Gas & Electric (PG&E) inspected and metered both existing and new efficient equipment consumption at once verifying the quantity, type, and consumption of the new equipment and the equipment being replaced, but only for a small sample of program participants. These same program evaluations found that tracking database estimates of the number of program measures installed agreed favorably with site inspections: Between 97% and 103% of tracking database estimates of measures installed were verified by site inspections for a limited sample of sites in each program. Site inspections by Central Maine Power (CMP) also found that tracking database errors, on average, did not affect savings estimates significantly.

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 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 overestimate their own equipment's hours of operation. Table 2 lists the results of hours of operation studies performed by the utilities in our sample.

Three methods were used by evaluators to obtain hours of operation information. The most sophisticated evaluations relied on data collected by

TABLE 2
Summary of Hours of Use Studies in Sample

<i>Utility</i>	<i>Ratio of More Accurate to Less Accurate Estimate</i>	<i>Source of First Estimate</i>	<i>Source of Second Estimate</i>
CMP	0.70	Customer self-reports	189 fixture hours of use metering
BECo	0.73	Customer self-reports	On-site inspection of 18 sites
CHG&E	—	Assumptions by building type	Customer surveys of equipment hours
Con Edison	—	Assumptions by building type	Customer surveys of equipment hours
NEES EI	0.78	Customer self-reports	23 site end-use metering
NEES Small C/I	1.02	Customer self-reports	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
SDG&E	0.93	Assumptions by building type	Customer self-reports
SDG&E	1.18	Customer self-reports	88 site hours of use metering

light-sensitive data loggers or end-use metering equipment. Less sophisticated evaluations used program employees to conduct on-site visits and collect information from building managers and employees. Finally, some programs implemented mail or telephone surveys to obtain hours of operation information from participants.

A systematic bias in customer reports of hours of operation is apparent in our sample. Site inspections, hours-of-use metering, and end-use metering by CMP, NEES, NU, and PG&E found recorded hours were less than customer self-reported hours. In only one case—SDG&E's Energy Efficient Hardware program—end-use metering uncovered that customer self-reports substantially underestimated equipment operating hours.

Our review also indicates that, in most cases, hours of operation should be measured by building type. In the six evaluations in which 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 due to vacancy and usage characteristics).

MEASURED CONSUMPTION PROGRAM SAVINGS ESTIMATES USING BILLING DATA

There are limitless combinations of econometric and statistical techniques that can be used to estimate energy savings from customers' energy bills. These designs may perform simple comparisons or multivariate regressions

of energy consumption across groups or time periods. More rigorous designs also incorporate weather, demographic, dwelling, and end-use data. Table 3 summarizes the methods used along with some of the other characteristics of each model.

In evaluations of DSM programs, random selection of participants and nonparticipants from a pool of identical consumers is usually not possible; all qualifying customers are given equal opportunity to participate and customers self-select into the program. Thus, the comparison group and program group are not truly random, and methods to measure savings are almost always based on quasi-experimental designs.² Comparison of participant and nonparticipant energy consumption, before and after efficient measures were installed, is the simplest method of estimating program-induced savings. Statistical techniques that control for the differences between comparison and program groups and that adjust for changes in consumption due to weather and other exogenous factors are also often used. Many of the more thorough evaluations used billing analyses of both participants' and nonparticipants' energy consumption to estimate savings.

The importance of using a comparison group in an analysis of consumption records is exemplified by the experience of Bonneville Power Administration (BPA) evaluators. The BPA Industrial Lighting Incentive program evaluation included a regression of participant characteristics against pre- and postprogram energy consumption. The model was unsuccessful in detecting a program effect. This may have been a result of 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.

The simplest use of customer billing data involves comparisons of participants and (matched) nonparticipants' energy bills before and after program intervention. Comparison models may detect savings, but their inability to distinguish program effects from weather (hours of operation change seasonally in the northern areas of the country), price, and other exogenous effects puts them at a distinct disadvantage. Seattle City Light normalized consumption records for weather changes and compared participant and nonparticipant consumption to estimate savings.

Program evaluators use econometric models to regress factors thought to affect energy conservation against actual consumption data. Some of the variables used in our sample of evaluations are program participation, corpo-

²Quasi-experimental designs are used when study and sample characteristics make locating an identical control group difficult. The classic quasi-experimental design types were first explained by Campbell and Stanley (1963): (a) "one-group pre-test post-test designs" utilize program participant consumption data before and after program intervention, (b) "static-group comparison designs" use program participant and nonparticipant consumption data for the period after program intervention occurred, and (c) "nonequivalent comparison group designs" utilize program participant and nonparticipant consumption data from both pre- and postprogram time periods.

TABLE 3
Summary of Evaluation Methods Based on Billing Data

Utility	Type of Model Used	Comparison Group	Sample Size (Total Participants)	Notes (Time Series Data Used, Sample Stratification, etc.)
BECO	Δ Consumption _{participants} Δ Consumption _{nonparticipants}	Eligible nonparticipants	772 (919) participants; 5,826 nonparticipants	12 months pre, 8 months post; ^a 10 strata based on size and seasonal usage
CHG&E	SAE, facility type vars., ^b building characteristics vars., ^c 2 tracking estimate vars.	Eligible nonparticipants	54 (606) participants; 116 nonparticipants	4-5 months pre, 4-5 months post; verified hours of use with customer surveys
Con Edison	SAE, facility type vars.	Eligible nonparticipants and soon to be participants	n/a (2,276) participants; n/a nonparticipants	4 months pre, 4 months post; verified hours with customer surveys
NEES EI	SAE, self-selection var., ^d building characteristics vars., 1 tracking estimate var. ^e	Eligible nonparticipants	369 (4,114) participants; 611 nonparticipants	12 months pre, 12 months post
NEES Small C/I NU	Δ Consumption _{participants} adjusted for nonparticipants SAE, self-selection var., facility type vars., 1 tracking estimate var.	Eligible nonparticipants	831 (2,494) participants; 698 nonparticipants	12 months pre, 12 months post
PEPCO	Pooled cross-section regression, self-selection var.	Eligible nonparticipants	1,123 (5,967) participants; 1,271 nonparticipants	5 months pre, 5 months post; 7 strata based on size; weather adjusted kWh
SCL	Δ Consumption _{participants} Δ Consumption _{nonparticipants}	Eligible nonparticipants	341 (345) participants; 1,452 nonparticipants	12 months pre, 12 months post; 4 strata based on size; weather adjusted kWh
PG&E	SAE, self-selection var., building characteristics vars., 1 tracking estimate var.	Eligible nonparticipants	118 (128) participants; 229 nonparticipants	12 months pre, 12-36 months post
SDG&E	CDA, 12 end-use vars.	Eligible nonparticipants	724 (6,432) participants; 370 nonparticipants	12 months pre, 12 months post
		None	181 (789) participants	12 months pre, 12 months post; adjusted model based on end-use metering results

^aPre/post = number of months of billing data compiled before and after program measures were installed. ^bFacility type vars. = dummy variables used to indicate the type of facility (office, retail, school, etc.). ^cBuilding characteristics vars. = variables used to indicate changes in floorspace, participation in other DSM, recent renovation, upswing in business, and so forth. ^dSelf-selection var. = variable obtained from a logit model and used to adjust for self-selection bias. ^eTracking estimate var. = variable used to indicate the tracking estimate of savings for each customer.

rate characteristics (e.g., business type, changes in business climate/productivity, number of employees, whether their business expanded), structural characteristics (e.g., facility square footage, changes in hours of operation, participation in other DSM programs, recent renovations), energy price, weather, and measures installed. By including data on nonparticipants and participants both before and after the measures are installed, adjustments for factors such as free ridership, weather changes, energy price changes, and measure usage changes are implicit in the model.

One technique, used by a number of programs in our sample, involves regressing pre- or postprogram tracking database estimates of savings for each participant (among other variables) against consumption data. This method, called the *statistically adjusted engineering* (SAE) method, calculates the proportion of the tracking estimate verified by the regression model. If the tracking estimates included in the model are already fairly good estimates of program savings, the SAE method results in savings estimates with considerably higher precision than regressions of billing data alone.

Ratio estimates obtained using SAE models ranged from 0.53 for NEES's Energy Initiative program to 1.05 for Con Edison'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, Con Edison adjusted their tracking database estimate based on a survey of customers 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. These ratios, known as *realization rates*, have been reviewed elsewhere (see, e.g., Nadel & Keating, 1991).

ESTIMATING CONSUMPTION PROGRAM SAVINGS FROM END-USE METERING

Electronic current meters and data-loggers provide useful information for estimating both energy and peak-demand reductions. For a time-series analysis, 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 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 2 weeks before and 2 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 across evaluation methods, evaluators uncovered important information about the ratio of metered savings estimates to tracking database estimates. For example, in NEES's 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 end-use, 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.

Metering is able to provide accurate and detailed information on the electricity consumption of installed measures. Through the measurement of equipment electricity consumption, changes in customer use of program equipment are monitored in real-time. This real-time measurement alleviates many of the problems described in the preceding section on tracking database estimates: utility or manufacturer estimates of baseline and program equipment efficiencies and hours of operation are based on actual field measurements.

The main drawback of end-use metering is its high cost. Multiple site-visits are required to install, maintain, and remove the equipment. The cost of end-use metering prevents metering of all but a small sample of program participants. In none of these programs was every measure sampled at every site, so potential biases may result from sampling a nonrepresentative set of measures (e.g., those that are most accessible and easiest to connect to data loggers) at each site, and from sampling a nonrepresentative sample of participant sites. Other recent evaluations have used only light loggers and spot-watt meters, rather than end-use meters with current transducers, to record equipment hours of operation and operating loads at a substantially reduced cost.

Examining the ratios of measured consumption estimates (from SAE models and metering studies) with tracking database estimates of annual savings suggests a pattern of moderate overprediction. Where both measured consumption and tracking database savings estimates exist, the average ratio, weighted by each program's energy savings, suggests that measured consumption estimates of annual savings are approximately 75% of tracking database estimates.

This section concludes our review of annual savings estimates. In the following sections we discuss free riders, market transformation, and the persistence of savings over the lifetime of program equipment. We then conclude with a discussion of our evaluation method taxonomy.

FREE RIDERS

One of the key difficulties associated with the evaluation of DSM programs is the requirement of estimating only those savings directly attributable to

the program. Thus, savings of participants who would have implemented the same set of program measures on their own (known as free riders) are excluded. Measurement of free riders is difficult. Whereas 19 of the 20 programs had an explicit estimate of free riders participating in the program, the methods used to identify or control for free riders varied dramatically across programs. Table 4 lists the utility estimates of free riders for each program in our sample, along with brief descriptions of the methods used to obtain those estimates.

As shown in Table 4, the estimates of free riders varied dramatically across programs. Because the surveys used to obtain free-rider information (and the subsequent analyses) were unique to each program, we cannot automatically attribute variations in free-rider estimates to differences in each program's population, or to the different technologies offered by each program. The sophistication with which a survey approaches the question of free riders affects the resulting estimate of free riders. Some surveys based their estimate of free riders on a single question that asked "Would you have installed the same measure if the program had not been offered to you?" Other surveys approached the issue in a less direct way, offering several different questions to check for consistency of responses.

Another difficulty we face when comparing free-rider estimates is variation in the definition of what a free rider actually is. Some programs define free riders as anyone who would have installed the same measure at the time of program implementation. Other programs broaden this definition to include anyone who would have installed the measure at any time during the next few years. Some programs count those who answered free-rider survey questions with "don't know" or "unsure" as free riders, or as 1/4 or 1/2 of a free rider. To add to this confusion, several programs include multiple questions regarding free riders in their surveys and then use the results of only one of those questions (without detailed explanation) to calculate net savings. Table 4 describes only those questions that were used to generate utility estimates of free riders.

An evaluation based on billing data using an appropriate comparison group (i.e., customers who were not offered the program but are otherwise identical to program participants in that they would participate if given the chance) can implicitly control for free riders. Several of the utilities in our sample assume that because their billing analyses include comparison groups (usually a random group of nonparticipants matched to participants according to energy consumption patterns, as described in Table 3), they have controlled for free riders when estimating energy savings. But the proportion of customers installing program measures without a rebate in a random group of nonparticipants is likely to be lower than that proportion in a group of participants (who, by stating their willingness to participate, may be more inclined to install the measures without a rebate). Thus, the comparison groups used by the utilities in our sample may not accurately control for free riders (Train, 1993). We are unable to estimate the extent of this bias but expect that its effect would be to underestimate actual free riders.

TABLE 4
Free-Rider Estimates and Estimation Methods

Utility	Percentage Free Riders	Method Used - Survey Question	Response That Would Indicate a Free Rider or Partial Free Rider	Weighted Responses By
BECO	14.0	Surveyed participants: "Did you already plan to install measures?"	Yes.	Not weighted
BHEC (Pilot)	73.2	Surveyed participants: "Would you have installed . . . if this program had not been available?"	Yes, unsure.	Not weighted
BPA (Pilot)	0.0	Professional judgment.		
CHG&E	2.6	Surveyed participants: "Would you have installed equipment without a rebate?"	Very likely = FR; Somewhat likely = 1/4FR; Somewhat likely with less efficient equipment = 1/4FR.	Respondent savings
CMP	21.3	Surveyed participants: "Would you have purchased . . . without the rebate?" and "Did you first learn about . . . from CMP?"	Yes to the first question and no to the second question.	Respondent savings
Con Edison	4.5	Surveyed participants: "How likely is it that equipment would have been replaced in the absence of the rebate program?"	Very likely in 3 months = FR; Somewhat likely in 3 months = 1/4FR; Very likely in 3 to 6 months = 1/4FR; Somewhat likely in 3 to 6 months = 1/4FR; Very likely in 1 to 2 years = 1/4FR; Somewhat likely in 1 to 2 years = 1/4FR.	Respondent savings
GMP (Large C/I)	12.5	Collaborative.		
GMP (Small C/I)	0.0	Collaborative.		
IE (Pilot)	44.0	Surveyed participants: "Suppose you were not offered this cash incentive allowance program?"	"I would have bought the same efficiency equipment this year."	Not weighted
NEES (EI)	6.5	Surveyed participants: "If EI had not been offered in 1991, would your company have spent this amount, in addition to any costs you already paid to install . . . at that same time?"	Yes.	Measure/respondent savings
NEES (Small C/I)	7.0	Surveyed participants: "What action would you have taken without program?"	Installed same efficiency equipment this year.	Measure/respondent savings
NMPC	12.7	Discrete choice model based on participant/nonparticipant characteristics.		
NU (ESLR)	10.0	Estimated from billing analysis.		
NYSEG	22.0	Surveyed participants: "What would you have done if the rebate had not been available?" and "How much did the rebate influence decision to purchase?"	Installed same efficiency equipment and strong or some influence.	Respondent savings
PEPCO	21.0	Surveyed participants: "Which statement best characterizes your actions?"	Basically did what I had planned to do anyway.	Not weighted
PG&E	23.0	Discrete choice model based on participant/nonparticipant characteristics.		
SCE	15.0	Participant survey; no further information.	Unknown.	Unknown
SCL (Pilot)	N/A			
SDG&E	18.1	Vendor and contractor surveys; no further information.	Unknown.	Unknown
SMUD	0.0			
Average	16.2	Professional judgment.		
Standard Deviation	17.0			

When billing analyses with comparison groups are not used, surveys of participants and nonparticipants are generally used to estimate free riders. The most sophisticated use of survey data is illustrated by Niagara Mohawk and PG&E, who used logit models calibrated with participant and nonparticipant survey responses to provide an estimate of the proportion of free riders. Although logit models are sophisticated statistical techniques, they are equally dependent on selection of an appropriate control group, appropriate explanatory variables, and quality survey data.

MARKET TRANSFORMATION

Utility DSM programs can result in additional energy savings for participants and nonparticipants over and above those directly targeted by those programs (e.g., if the program influences customers to undertake additional energy efficient equipment investment on their own or encourages nonparticipants to install program measures). We broadly classify these spillover and free driver effects as market transformation.³ Estimating the extent to which DSM encourages participants, nonparticipants, and dealers to install or stock efficient equipment without a rebate requires extensive surveys of all customers and dealers regarding program awareness and their decisions to purchase efficient equipment. Alternatively, aggregate sales data for efficient equipment can be compiled and analyzed. Both techniques are difficult and not considered part of the normal practice of utility program evaluation. Only four programs attempted to estimate the magnitude of participant spillover—the number of additional efficient measures later installed, without rebates, by utility customers who were educated through their initial participation in the program. One program also asked survey questions aimed at verifying the existence of free drivers—nonparticipants who install efficient equipment as a result of hearing about the program or about program measures from those customers with firsthand program experience. The results of these studies are summarized in Table 5.

Whereas none of the programs estimated the additional energy saved through spillover or by free drivers, the survey results suggest that the effects of the programs on customer behavior and perceptions of efficient technologies could drive, and eventually transform, the market for efficient equipment. Free drivers and spillover effects represent a new resource that, when properly measured, can affect utility and total resource cost results significantly. This is in contrast to free riders, who do not reduce actual resource savings (free riders do save energy), but instead represent a transfer of capital from the utility, and thus ratepayers, to the free riders.

³A detailed discussion of the many facets of market transformation can be found in Feldman (1994).

TABLE 5
Evidence of Free Drivers and Spillover From Evaluation Surveys

Utility	Affirmative Responses		Survey Question
	Participants	Nonparticipants	
CHG&E	25%		Influenced by program to buy efficient equipment on your own?
NEES EI	65%		Would you now install equipment without a rebate?
NEES Small C/I	51%		Would you now install equipment without a rebate?
NU	51%	13%	Influenced by program to buy efficient equipment on your own?

PERSISTENCE OF SAVINGS AND MEASURE LIFETIMES

Neither initial billing analyses nor end-use metering methods can verify the long-term persistence of program savings. 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. Consequently, none of the utilities in our sample have performed studies that address the long-term persistence of program savings.⁴

Current estimates of savings are often based on the assumption that equipment will operate for the duration of the manufacturers' estimate of the equipment's useful life.⁵ Measure lifetime varied widely for identical measures from program to program. In some programs, lifetimes were based only on manufacturers' estimates of product longevity. In a few cases estimates were adjusted downwards to account for some premature retirement due to the predicted frequency of building renovations. Whereas several utilities (CMP, NEES, Seattle City Light [SCL]) used site inspections and billing analyses to estimate savings persistence 1, 2, and 3 years after installation, in no cases were measure life estimates based on a complete longitudinal set of data from past program participants. The average measure lifetimes, weighted by the proportions of program equipment, for each program in our sample are given in Table 6. Because each program installed different quantities and different types of lighting equipment, we expect variation in the average lifetimes across programs.

Explicit persistence of program savings is best identified using site visits. On-site inspections in the Boston Edison Small C&I Retrofit program uncov-

⁴Utility DSM programs and DSM program evaluation are too nascent to have long-term studies of persistence: Measures from the earliest large-scale DSM programs (from the early 1980s) are just reaching the end of their manufacturer's rated lifetimes.

⁵Alternatively, for the American Society of Heating, Refrigerating, and Air Conditioning or the Association of Home Appliance Manufacturers estimate of measure life.

TABLE 6
Summary of Measure Life Estimates Used to Calculate Lifetime Savings

<i>Utility</i>	<i>Measure Life Estimate (Years)^a</i>	<i>Source of Estimate</i>
BECo	15	IRT report ^b
BHEC	10	Utility report ^c
BPA	15	Utility report
CHG&E	15	Utility contact
CMP	7	Utility report
Con Edison	11	Utility contact
GMP Small	6	Utility report
GMP Large	15	Utility report
IE	12	Utility report
NEES EI	18	Nordax database ^d
NEES Small C/I	15	Nordax database
NMPC	13	Utility contact
NU	17	Utility contact
NYSEG	10	Utility contact
PEPCO	10	Utility contact
PG&E	16	Utility report
SCL	16	Utility report
SCE	13	Utility report
SDG&E	15	IRT report
SMUD	5	Utility contact

^aAll estimates represent program averages weighted by the specific equipment installed. All measure life estimates, regardless of original source, have been verified with utility representatives. ^bIRT report = program summary sheet from the Results Center, Aspen, CO. ^cUtility report = evaluation report from utility. ^dNORDAX = Northeast Region Demand-Side Management Data Exchange, documented by Synergic Resources Corporation, Philadelphia.

ered a 13% rate of measure removal for lighting measures after 18 months. CMP evaluators discovered that up to 15% of all lighting measures had been removed due to theft, dissatisfaction, and equipment failure within 2 years. As an upper bound, 30% of all compact fluorescent lamps CMP had installed were stolen (primarily from hotel rooms) or removed due to dissatisfaction with light levels.

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 2-year period. SCL evaluators used comparisons of participant and nonparticipant billing data to estimate savings persistence over a 3-year period. Whereas NEES found almost 100% persistence, SCL found a gradual degradation of savings, where approximately 95% and 88% of original savings remained after 2 and 3 years, respectively. However, the cause of such a degradation is not limited to measure removal. Degradation of savings, as evidenced by a billing comparison, could be the result of exogenous changes in participant and nonparticipants' equipment efficiency, poor maintenance of measures, or increased consumption due to take back.

TAXONOMY OF EVALUATION METHODS AND UTILITY EVALUATION STRATEGIES

The diversity of impact evaluation techniques used in these 20 programs is illustrated in Table 7. 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 because 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).

The taxonomy described in Table 7 is a valuable tool for program evaluators. With some knowledge of which sources of evaluation error are most relevant to their programs, program evaluators can select evaluation methods that reduce or eliminate those errors in the evaluation results. The taxonomy also demonstrates the numerous tradeoffs between evaluation methods. Not every method can identify and control for every possible facet of evaluation error, so evaluators, and the regulators who guide their choices, must identify those methods that can reduce a savings estimate's most significant biases.

Because no single method provides both a reliable estimate of program savings as well as a quantification of individual factors that affect savings, overall evaluation strategies that combine the results of multiple evaluation methods are quite useful. 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 between the utility, the program delivery, the program technologies, and the participants suggests that evaluation would benefit from holistic approaches that incorporate 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 multi-year time horizon.

TABLE 7
Taxonomy of Impact Evaluation Methods Used in Commercial Lighting DSM Programs

Evaluation Method	Implicit Accounting of Attributes in Savings Calculations			Explicit Examination of Program Attributes			
	Adjusts for Technology Failure/Misuse ^a	Controls for Exogenous Factors ^b	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 with hours of use verification			Partially			Partially ^c	
Tracking estimate with site inspections	Yes				Yes	Yes ^c	Yes
Tracking estimate with short-term metering	Yes	Partially	Yes				
Bill comparison of participants/nonparticipants	Yes	Partially	Yes	Partially			
Billing analysis (regression of consumption data)	Yes	Yes	Yes	Yes ^d			
Statistically adjusted engineering analysis	Yes	Yes	Yes	Yes ^d			
Logit model evaluating participation decision				Yes (explicitly quantifies)			

^aTechnology failure/misuse includes participant failure to install, participant sabotage. ^bExogenous factors include weather, business and structure characteristics, and fuel prices. ^cIf performed both before and after measure installation. ^dOnly with the appropriate control group.

NEES uses an iterative process in which program savings for the current program year are estimated based on billing analyses from evaluations of previous program years. They use a number of methods, including end-use metering and billing analyses, to estimate energy savings. NU also augments estimates of savings based on the program auditors' tracking database with on-site equipment assessments, end-use metering, and analysis of billing records. SDG&E relies on tracking estimates until hours of operation information are available from participants, at which point tracking estimates are adjusted based on the new hours of operation information. When billing analyses become available, usually 1 or 2 years after program implementation, tracking estimates are adjusted based on billing analysis results. PG&E has improved the precision of their savings estimates significantly by leveraging the smaller sample results from end-use metering against results from the tracking database and from regression models based on billing records.

Eventually, refinements in our understanding of the factors that affect program savings may make extensive evaluation unnecessary and allow us to adjust tracking database estimates using measured consumption information from a small sample of participants. Evaluation methods could be selected that address key program uncertainties, as identified by previous evaluations. If the cost of each evaluation technique were known beforehand, then the cost of the evaluation could be traded off against the probable increase in precision associated with each evaluation method (Hummel, 1993; Wirtshafter & Baxter, 1991).

CONCLUSIONS

Current practice in DSM program evaluation is evolving quickly. Five years ago we would have been hard pressed to find even a handful of programs with evaluations incorporating multiple measurement methods. Focusing only on commercial sector lighting rebate programs, we found almost a dozen programs with both tracking database and measured consumption savings estimates.

Our review of free-rider evaluation methods suggests that there is little consensus among utilities about the definition of a free rider. Although the absence of consensus is a secondary concern for the total resource cost of energy efficiency programs, free riders have important consequences for the rate impacts of programs on utility ratepayers. We note, with some irony, that comparatively little attention has been devoted to measuring free drivers and spillover effects, which both reduce total resource cost of energy efficiency and mitigate the rate impacts of these programs.

Current practice in evaluation, and in evaluation of the 20 programs examined, uses a diverse set of methods. With the exception of the methods used to perform end-use metering, no two programs implemented their evaluations in the same way. The diversity of evaluation methods used, coupled with our limited sample size, makes it difficult to draw general conclusions about the efficacy of particular methods.

ACKNOWLEDGMENT

The work described in this article was funded by the Assistant Secretary of Energy Efficiency and Renewable Energy, Office of Utility Technologies, Office of Energy Management Division of the U.S. Department of Energy under Contract No. DE-AC00-76SF00000.

REFERENCES

- Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Palo Alto, CA: Houghton Mifflin.
- Eto, J., Vine, E., Shown, L., Sonnenblick, R., & Payne, C. (1994). *The cost and performance of utility* (Rep. No. LBL-34967). Berkeley, CA: Lawrence Berkeley Laboratory, Commercial Lighting Program.
- Feldman, S. (1994). Market transformation: Hot topic or hot air? In *Proceedings of the 1994 American Council for an Energy Efficient Economy Summer Study on Energy Efficiency in Buildings*. Asilomar, CA: American Council for an Energy Efficient Economy.
- Hirst, E., & Reed, J. (Eds.). (1991). *Handbook of Evaluation of Utility DSM Programs*. Oak Ridge, TN: Oak Ridge National Laboratory.
- Hummel, P. (1993). Resource allocation and DSM program evaluation planning. In *Proceedings of the 1993 Energy Program Evaluation Conference* (pp. 637-642). Chicago: National Energy Program Evaluation Conference.
- Nadel, S. M., & Keating, K. M. (1991). Engineering estimates vs. impact evaluation results: How do they compare and why? In *Proceedings from the 1991 Energy Program Evaluation Conference* (pp. 24-33). Chicago: National Energy Program Evaluation Conference.
- RCG/Hagler Bailly, Inc. (1991). *Impact evaluation of demand-side management programs* (Rep. No. CU-7179). Palo Alto, CA: Electric Power Research Institute.
- Train, K. E. (1993). A review and critique of statistical techniques for estimating net kWh and kW impacts. In *1990 Southern California Edison Energy Management Services and Hardware Rebate Program Evaluation (Volume 8)*. San Dimas, CA: Southern California Edison.
- Wirtshafter, R., & Baxter, L. (1991). Establishing priorities for future evaluation efforts. In *Proceedings of the 1991 Energy Program Evaluation Conference* (pp. 137-142). Chicago: National Energy Program Evaluation Conference.

ENERGY SERVICES JOURNAL, 1(1), 55-65
Copyright © 1995, Lawrence Erlbaum Associates, Inc.

Model Specification and Treatment of Outliers in the Evaluation of a Commercial Lighting Program

Michael T. Ozog and Ronald E. Davis
Hagler Bailly Consulting, Inc.
Boulder, CO

Donald M. Waldman
University of Colorado

Dorothy A. Conant
New England Power Service Company
Westborough, MA

This article addresses several modeling and estimation issues that confront evaluators every time they examine data from a demand-side management (DSM) program. Particular attention is paid to commercial lighting programs in which participants are likely to be heterogeneous (e.g., different industries and different energy consumption patterns). The state of data analysis overall and DSM evaluation in particular is such that there are no generally agreed on answers to questions like: How should the variables in the energy use or savings equation be specified? What is the correct functional form of the energy savings equation? Do we discard observations that seem to be far from the norm in the sample (outliers)? Is least squares multiple regression appropriate or is a less restrictive statistical model better? Is heterogeneity of users important? We examine the data from one commercial lighting program and wrestle with these issues using various regression diagnostics and model specifications. Where we can, we generalize. Where we cannot, the suggested tests and comparisons may help decide what is best for the particular DSM program and resulting data.

