DSM EVALUATION METHODS: QUANTIFYING COST AND ACCURACY (IDENTIFYING COST-EFFECTIVE ALTERNATIVES TO END-USE METERING)

Marc Hoffman and Andrew Parece
XENERGY, Inc.
Burlington, Massachusetts

Introduction

A distinguishing feature of sample design and research planning for DSM impact evaluation is the complexity introduced by the range of accuracy and cost associated with measuring the parameter of interest. Traditional sampling studies account for bias implicitly with a valid sample design. Where significant discrepancies between measured and actual impacts are identified for some evaluation methods, the application takes on a decision dimension: how do the alternatives compare on the basis of sampling costs and attainable accuracy, and how are these weighed to arrive at the most cost-effective research plan?

A recent scoping study, sponsored by eight New York utilities and performed by XENERGY, examined alternative evaluation methods to quantify their cost and accuracy characteristics. Costs were defined by marginal sampling cost, and accuracy by the identifiable bias uncontrolled for with the sampling process. In this study, four impact evaluation methods (end-use metering, closed-form engineering models, building simulation models, and hybrid statistically-adjusted engineering models) were analyzed on the basis of these criteria. Quantification of cost and accuracy was obtained for each method by building type, producing scatterplots depicting the underlying tradeoffs. This paper presents a review of the analytic methodology and a discussion of major findings.

The following are the evaluation (impact estimation) methods under study:

1. ***End-use metering:*** hourly end-use metering of facility electric usage. Several levels of metering can be pursued (single point, multiple points, multiple end uses, and so on). At a minimum, we refer to metering all loads for significant end uses.

2. ***Engineering building simulation models:*** an algorithmic model of hourly end-use loads based on engineering judgment, equipment stock, operational/behavioral parameters, and weather data. We make the distinction between engineering simulation models, which generate hourly loadshapes by end use and closed- form savings models, which identify *impacts* based on delta watts, full load hours, and other inputs (see 4 below). The former type models can be structured to develop *before* and *after* end-use loadshapes, and thus can be considered equivalent to savings models in this sense.

3. ***Hybrid Statistical/Engineering (HSEM) Methods:*** an engineering model that is adjusted for biases inferred in end-use models by analyzing regression coefficients for hourly total building loads (on end use load estimates). The statistical adjustment is based on whether loads are systematically over-estimated/under-estimated as inferred by regression by building type. The adjustment requires hourly total building loads to be known (see Ref. 1).

4. ***Engineering Tracking Models:*** a closed-form single-equation model of end-use load impacts based on engineering judgment. Significant parameters of models are delta watts, full load hours, diversity and coincidence factors. Other factors are often developed to account for persistence, free ridership, snapback, and so on.

Upon review, concerns among several diverse interests have focused on the conspicuous absence of billing or statistical analyses among the methods considered here for assessment. We address these concerns prior to providing details of the study. Both our understanding of the merits and efficacy of billing/statistical analysis and their exclusion from the framework to be presented are based on considerable experience and review of ongoing research in this area. In our experience, billing analysis categorizes a range of analyses whose common thread is the use of metered billing information.

This common thread has borne several types of analysis that rely on the explanation or modeling of load levels over time. Use of billing data can be as elementary as tests for significant differences (over time) on weather-normalized loads within an experimental design. These and more elaborate methods (conditional demand, constrained change, or HSEMs), all fundamentally attempt to explain or model known load levels (bills). In terms of relative explanatory power and in the purest sense, these methods can be considered equivalent. A linear constrained change model could be developed to emulate an engineering or HSEM model, given identical input data and information on the functional form of the models; conversely, an engineering model could incorporate significant independent variables to emulate the change or conditional demand models.

Practically, however, the types of models described rarely resemble one another and the lack of abundant, high quality data leaves the assessment of relative efficacy indeterminate. We have found hybrid statistical/engineering modeling to be a sensible, reviewable, and reproducible approach to utilizing billing data for a variety of purposes. This is not the sole approach the authors condone or use, but, in designing for propitious use of billing data for evaluation, we find that it meets some important criteria absent in other methods.

Typically, billing analyses are performed with monthly bills rather than hourly load data by type of day. Numerous studies have confirmed patterns in variance of end-use loads by hour of day. A requirement for the study under review and one common to evaluation research is the assessment of estimation techniques which specifically considered coincident peak hour impacts. Modeling monthly billing demand by any method provides only estimated non-coincident demand impacts, neglecting the explanatory power of calibration to hourly load data.

Defining the Frontier: Cost and Risk

A starting point in quantifying the tradeoffs presented by the various measurement methods involves specifying the criteria for assessment. Among the candidates for quantification are: sample sizes, precision/confidence, bias, total absolute error, total sampling costs, total evaluation program cost, and total or marginal cost as a percent of avoided cost.

Cost

The above list in fact defines a relatively limited universe of criteria. For instance, total sampling costs are a function of precision and confidence requirements, the sample sizes these requirements entail, and the sampling cost per unit. Assuming that marginal sampling costs per unit are fixed, imposing some precision/confidence constraint results in a range of total cost associated with the evaluation methods. Conversely, if total sampling cost is the constraint, the methods represent a range of attainable precision. In effect, the precision/confidence and sampling cost criteria mirror one another—choosing the constraint defines the criteria. (See Table 1).

In order for cost to be defined as the constraint, some conventions are necessary. Are costs defined as total evaluation program cost, total cost per kWh/kW saved, or cost per kWh/kW saved relative to avoided cost? To avoid the ambiguities in using such "subjective" criteria, precision/confidence has been defined as the constraint (90/10) with marginal *per unit* sampling costs applied to resultant sample size requirements to derive a measure of "variable" program evaluation costs.

Risk

With precision/confidence and sampling costs taken into consideration, a further aspect requiring quantifica-

Table 1. Hypothetical Sampling Cost/Precision Tradeoff

Constraint		Criteria	
Precision = ±10%	Cost:	Engineering (Tracking)	$10/MWh
		Engineering (Simulation)	$20/MWh
		HSEM	$25/MWh
		End-use Metering	$50 MWh
Cost = $30/MWh	Precision:	Engineering (Tracking)	±5%
		Engineering (Simulation)	±10%
		HSEM	±15%
		End-use Metering	±35%

tion is the bias, or identifiable measurement error associated with the evaluation methods. The bias component is significant in this context for two reasons. First, statistical sampling implicitly relies on unbiased measurement to produce unbiased inferences about the population under study. Second, many evaluation methods' impact estimates have been shown to exhibit systematic biases, typically corresponding to discrepancies in reported/actual operational or behavioral characteristics of facilities being modeled. The biases cannot be controlled for with increased sample sizes, and represent a substantial potential risk associated with some measurement methods.

The methodology employed in the scoping study quantified this aspect of evaluation planning by examining empirical evidence on the magnitude of these biases. This examination requires an analytic framework that allows for estimation of systematic biases of evaluation methods' hourly load estimates by end use. Specifically, some estimate of the bias needs to be developed for each impact estimation method, such as:

$$E\ [|1-\frac{\tau\ actual_{ijt}}{\tau\ estimated_{ijt}}|]$$

where:

i = *index of customers*
j = *index of end uses*
t = *index of hours*
E = *the expectation operator*
τ = *impact level*

The methodology used in this study relies on assumptions of an SAE application (see Ref. 2), but expands on this framework to derive estimates of relative costs and risk of the methods. Table 2 summarizes the input requirements and presents hypothetical scenarios of cost and risk parameters. A brief description of the data and SAE methodology, along with application to sample sizes and risk parameters follows.

Assessing Individual Methods

The preceding discussion imposes several data development tasks to estimate sampling requirements, resultant sampling costs, and bias levels accurately. The New England Electric System has generously permitted XENERGY use of a set of data collected for a stratified sample of its commercial customers spanning seven building types, or sectors.

Engineering estimates of end-use loads have been developed for each of the buildings under study using XENERGY's LOAD PLANNER load shape modeling software. Further, the engineering "simulation" estimates are improved upon using statistical inferences about the biases of the engingeering models by end use and hour. These inferences are drawn from an accepted technique whereby hourly engineering estimates for all end uses are regressed upon known hourly whole-building load data. The process that generats the HSEM estimates is shown in Figure 1.

Examining the β_{jt}, from the primary, unadjusted sector level regressions, one can discern which loads are

Table 2. Calculating Risk and Cost

Risk = (Bias + Precision) Estimated Savings				
Method	Bias	Precision	Estimated Savings (GWh)	Risk
A	0%	±10%	±20	±2
B	16%	±10%	±20	±5
C	35%	±10%	±20	±9

Cost = Unit Cost of Method x Sample Size			
Method	Unit Cost per observation	Sample Size[a]	Total Cost
A	$54,000	35	$1,890,000
B	$2,400	37	$88,800
C	$20	46	$920

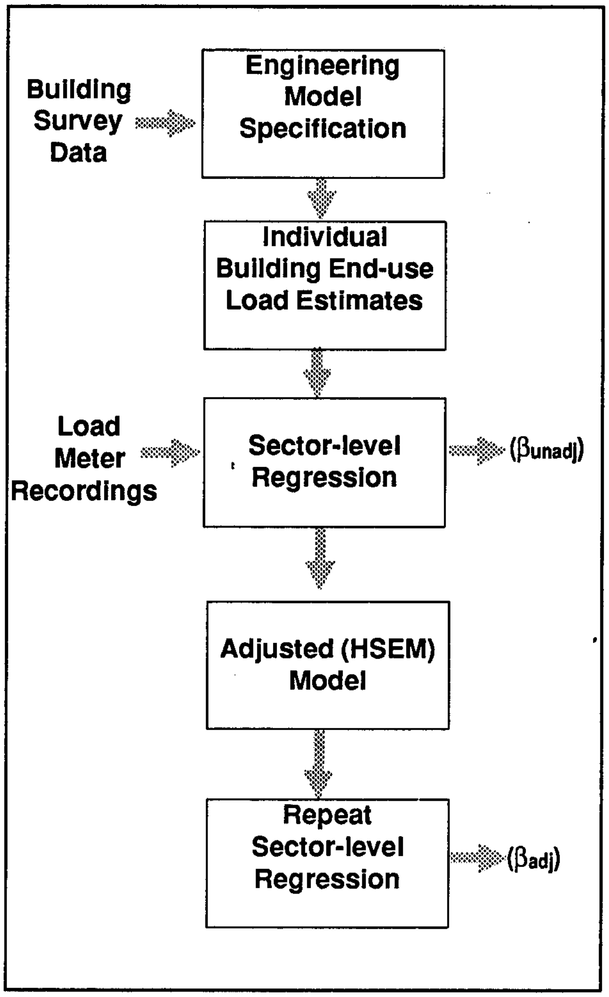
Building Survey Data
Engineering Model Specification
Individual Building End-use Load Estimates
Load Meter Recordings
Sector-level Regression
(βunadj)
Adjusted (HSEM) Model
Repeat Sector-level Regression
(βadj)

Figure 1. The HSEM Methodology

systematically over-estimated or under-estimated by building type and hour, as these coefficients represent estimates of the ratio or actual to estimated load. Examples of documented bias phenomena of engineering models include over-estimation of miscellaneous loads due to the use of nameplate ratings, and lighting estimate biases related to under-estimation of load factors and operating hours.

The β at the first stage are estimates from the equation:

$$L_{it} = \sum \beta_{jt} E_{ijt} \, err_{it}$$

where, for end use *j* and hour *t*:

i = index of customers
j = index of end uses
t = index of hours
E = the engineering estimate
β_{jt} = *the regression coefficient.*

Understanding of these biases is incorporated into the engineering models, resulting in an adjusted or hybrid statistical/engineering model (HSEM). It is important to note that the adjustment process is iterative and performed by building type, resulting in the set of adjustments to the engineering models that best accounts for the hourly difference between estimated and actual load for all buildings in the sector. Thus the process does not completely eliminate the bias found among end uses. In fact, a repetition of the regression phase of analysis identifies any remaining bias after the adjustments are made.

As mentioned previously, translating the estimates into sample sizes required by each method for some desired level of accuracy depends on assumptions about the relationship between load and savings impacts. Although the engineering load shape model employed for this study predicted end use *loads*, and not before/after impacts, these models are useful proxies for an algorithmic engineering loadshape impact model.

Quantifying Cost—The Measurement Variance

For the HSEM and engineering simulation methods, the variance of the β_{jt} were assumed to represent the variance of the measurement error associated with these methods. Consider that engineering or HSEM estimates could be compared directly to end-use metered data for the entire population. Then

$$E\,[\,|\,1 - \frac{\tau_{metered_{ijt}}}{\tau_{estimated_{ijt}}}\,|\,]$$

where:

i = index of customers
j = index of end uses
t = index of hours
E = the expectation operator
τ = impact level

represents the mean relative error in the estimates. For this study, we assume that this quantity is equal to $|\,1 - \beta_{jt}\,|$ (*i.e.*, $\tau_{metered}/\tau_{estimated} = \beta$ From this assumption, it follows that the variance of the regression coefficients provides a reasonable estimate of the variance of the ratio of actual to estimated impact, or the variance of the measurement error. To summarize, the regression coefficients are estimates of the ratio of actual to estimated usage. Their variance is taken as the variance in measurement error, while the absolute difference of the coefficients from unity represents the bias of the method.

Lacking end-use metered data to compare with the engineering or hybrid estimates, we feel that this is the most precise measure available. In essence, our reliance on the regression coefficients drives the quantification process. In our judgment, this provides a more sound basis for specifying sample sizes and bias estimates than methods that rely entirely on assumptions about these parameters or models.

Quantifying Cost—The Population Variance

When sampling for impacts with individual estimates, the measurement variance referred to above is only one component of the variance necessary to specify sample sizes completely. The other variance component, of equal magnitude across all methods, is the underlying population variance of the parameter of interest. This is the variance estimate familiar in traditional sampling studies where accurate measurement is presumed. To illustrate, for end-use metering, where no *measurement* variance is evident in the impact estimate produced (another assumption), a sample designed to estimate total program impacts requires an estimate of the variance of the true program impacts among the population of participants.

Estimates of this variation were arrived at separately by two methods. The first involved identifying the coefficients of variation (CVs) of an engineering parameter thought to significantly affect impacts, namely operating hours. These CVs were obtained by utility service territory and building sector from a merged New York Power Pool audit database. Another measure of underlying impact variance was taken from the NEES dataset used to estimate end use hourly loads. The variance of unitized hourly end-use loads (kWh per square foot) was examined as a proxy for DSM impacts. In most cases, the results were consistent. Results from the NEES dataset were used for population CVs to provide a consistent data source for all parameters. For each measurement method, the measurement and population variance are used together through calculated CVs to derive sample sizes.

Quantifying Risk

An implication of our assumptions regarding the regression coefficients as estimators of the ratio of actual to estimated usage relates to the bias, as noted previously. Specifically, a direct result of this assumption is the availability of $|1 - \beta_{jt}|$ as an estimate of the method's estimated biases. The magnitude of this quantity is taken as a measure of the method's expected relative error. When multipled by the total estimated population impact, we have a measure of the identifiable risk associated with the method for the given population. Recall that precision requirements were constraints for assessing costs. The relative precision presents another element of risk insofar as it bounds the expected sampling error for a given confidence level.

Lacking regression coefficients for engineering tracking estimates for the NEES sample data, some assumptions were applied. First, for tracking estimates, the bias was taken to be equal to that of the engineering simulation models. Second, the measurement variance was taken to be twice that of the engineering simulation models. We fell these are appropriate assumptions relative to other methods. In addition, end-use metering is assumed to arrive at the true impact, so that the measurement variance is zero.

Results—Individual Methods

Table 3 presents results of the analysis performed for the office sector over the peak hour of average summer days. The peak hour of summer average day types was selected as an indicator of the relative variance and bias for the alternative methods. Other groupings of hours (all peak hours, all off-peak hours, all shoulder hours) and sectors were analyzed with results that are generalizable from those presented. Summer was selected for demonstration because the heating parameters in the winter days were not significant probably because of low electric heat saturation. The peak hour was selected because the HSEM adjustment process focused on calibrating the loads specifically for this hour. Results for each sector were similar and the office sector was selected for illustration because of its size.

Table 3 shows that the individual methods cover a spectrum of cost and risk. Using the lighting end-use as an example, on one extreme with regard to cost is end-use metering, with an estimated $1.8M required to cover a sample size of 35. In contrast, an engineering tracking model covers the required sample sizes with costs orders of magnitude lower. The tradeoff, of course, is the risk element. The susceptible risk evident with the tracking model is over four times that seen with end-use metering. Lying between these extremes on both scales is the HSEM method.

Table 3. Sampling for DSM Impacts Individual Methods

Individual Methods	Sample Size[a]	Cost[b]	Risk (MWh)
End Use Metering			
Lighting	35	$1,890,000	4,040
Cooling	121	6,534,000	88
Miscellaneous	161	8,694,000	122
Engineering Tracking			
Lighting	46	920	18,253
Cooling	179	3,580	311
Miscellaneous	181	3,620	575
HSEM			
Lighting	37	88,800	10,427
Cooling	163	391,200	88
Miscellaneous	166	398,400	456

[a] Sample sizes required for ±10% precision, 90% confidence. Engineering tracking standard error is assumed 50% greater than HSEM. Hybrid/engineering tracking standard error is assumed 30% greater than HSEM. Risk uses predefined target levels, assuming 20% of impacts are among offices.

[b] Costs are Marginal Unit cost based: Tracking, $20; HSEM, $2,400; End-use metered, $54,000.

Combinations of Methods for Estimating DSM Technology Impacts

A common tactic used in sampling applications is to leverage some knowledge about a given parameter's relationship with another parameter about which accurate information is available for the population. For example, in sampling for load research we often make use of the relationship between billing kWh and kW demand, perhaps within a stratified design, to lower sampling costs. The sampling (*i.e.*, cost) efficiency results from knowing more about the relationship, or ratio of the two parameters, than we know about, say, kW demand, on its own.

Sampling to *validate* some impact measurement affords the same efficiencies by virtue of this concept. The estimates produced by some model other than end-use metering have been shown to exhibit some measurement error and bias relative to the true impacts. We can think of these as *good guesses* of the true impacts. Taking for granted that an engineering *tracking* model will be used to produce impact estimates for all participants, the issue arises: how good are the estimates, and how can they be improved in a cost-effective manner to desirable precision/ confidence levels?

Ratio Estimation Applied for Impact Measurement Validation

The ratio estimation concept applies because we are fundamentally interested in the ratio, on average, between the tracking estimate and the true impact level. If one were to sample all participants and measure DSM impacts via end-use metering, a ratio could be developed for each participant (actual impact/estimated impact) by which the tracking estimates can be adjusted. This presents an extreme case of overkill, as we would no longer need thė tracking model, and the detrimental cost implications are obvious. Theoretically, some sampling of participants with end-use metering could provide a mean ratio by which to adjust tracking model estimates, the underlying assumption being that there is some ascertainable relationship between the estimates and the true impact. Deviations of this (mean) ratio from unity indicate the extent of the bias of the estimates, while the dispersion or variation of the ratio indicates the extent to which validation sampling efficiencies exist. The concept of validating estimates in this manner is shown in Figure 2 with hypothetical data for the purpose of demonstration.

Estimating Sample Sizes for Validation

The preceding discussion assumes that some *a priori* knowledge of the variance of the ratio is required in order to develop sampling plans. Intuitively, the sample size required is dependent on the strength of the relationship between the actual and estimated impact. Some basis for the magnitude of the ratios and their variances must then be developed. Since we have already argued that the regression coefficients from the NEES HSEM model estimation process can be taken as estimates for $\tau_{metered}/\tau_{estimated}$ these coefficients can lend some perspective on the magnitude of the ratios and their variances.

If we take the standard error of these parameters to be a measure of the dispersion of the ratio we can use it for purposes of determining validation sample sizes. The population variance is not required in estimating sample sizes because there exists a tracking model estimate for all participants. We can sample to identify the ratios of these estimates to actual. In addition, the validation approach implicitly derives unbiased estimates, as any number of end-use metered readings used in this framework will eliminate the bias. This method for improving upon estimates with smaller *validation* samples is currently being

	Average Estimated Savings				Ratios		
	N	Engineering Tracking Model	HSEM	End-use Metering	Hybrid Tracking	Metered Hybrid	Metered Tracking
Tracking Model	1,000	150					
HSEM Sample	13	140	125		.89		
End-use Metered Sample	2	130	120	76		.63	.58

Validated Estimate I (Two-tiered):

Ratio of Metered/Tracking = 76/130 = 0 .58
Improved Estimate = Tracking • 0 .58 = 150 • 0.58 = 87

Validated Estimate II (Three-tiered):

Ratio of Hybrid/Tracking - 125/140 = 0 .89
Ratio of Metered/Hybrid = 76/120 = 0 .63
Improved Estimate = Tracking • 0 .89 • 0.63 = 150 • 0.89 • 0.63 = 84

Figure 2. Validating Engineering Tracking Model Estimates with End-use Metering, and HSEM Estimates

employed as the basis for impact estimation at Northeast Utilities. As with the analysis of independent methods, we assume that end-use load estimates are appropriate proxies for sample size requirements for DSM impacts.

Table 4 utilizes the results of the two-tier method sample sizes with predefined impact estimates (statewide program targets) to show that equal precision can be achieved, for the same risk level, with fewer end use meters thanif end-use metering is used alone. Table 3 shows that 35 end-use meters would be needed to achieve ±10% precision given the estimate of lighting population CV. For the lighting end use, these results indicate that with 11 end-use meters the engineering-tracking estimates can be validated with ±10% precision. This result and the result that would be obtained with 35 end-use meters would both be within 10% of the true gross savings. However, the validation framework implies dramatic cost reductions.

Multi-tiered Methods

By extension, if we can realize cost savings in sampling for validation of a tracking model with some end-use metering, it may prove possible to layer more than one validation stage for further cost savings. For example, if we presume that the variance of the ratio of successively more precise estimates was smaller than that assumed in a one-stage validation, this would entail even fewer high cost end-use metering sample points at potentially significant cost savings. The effect is demonstrated in Figure 2. An additional estimate needs to be produced here, specifically:

$$H_{it} = \sum \psi_{jt} E_{ijt} + err_{it}$$

where, for end use *j* and hour *t*:

i = index of customers
j = index of end uses
t = index of hours
E = the engineering estimate
H = the HSEM estimate
ψ_{jt} *= the regression coefficient*

Here, the regression coefficients represent the ratio of a hybrid estimate to an engineering tracking estimate, the first tier of the validation adjustment. As illustrated in Table 4, the estimate produced with 19 HSEM and 3 end-use meters when applied as ratios to the other tier's estimates produces a population estimate with ±10% precision. Again, the precision of this estimate is the same as would be obtained with 35 end-use meters alone (see Table 3) or 11 end-use meters in a two-tier approach. Figure 3 shows these results relative to independent methods.

Table 4. Sampling for DSM Impacts Combinations of Methods

Combination of Methods[a]				Cost[b]	Risk (MWh)
Two-tiered					
Metering/Eng tracking	*Tracking*		*End-use meters*		
Lighting	All Participants		11	$614,000	4,040
Cooling			21	1,154,000	88
Miscellaneous			20	$1,100,000	122
Three-tiered					
Metering/Hybrid/Eng Tracking	*Tracking*	*Hybrid*	*End-use meters*		
Lighting	All Participants	19	3	$227,600	4,040
Cooling		36	3	$268,400	88
Miscellaneous		34	5	$371,600	122

[a] Sample sizes required for ±10% precision, 90% confidence. Engineering tracking standard error is assumed 50% greater than HSEM. Hybrid/engineering tracking standard error is assumed 30% greater than HSEM. Risk uses predefined target levels, assuming 20% of impacts are among offices. Assumes 1000 program participants.

[b] Costs are Marginal Unit cost based: Tracking, $20; HSEM, $2,400; End-use metered, $54,000.

Conclusions

Figure 3 illustrates the cost-risk tradeoff of the three individual methods and the two combination methods for lighting for the office sector. Risk and cost are increasing on their respective axes. Therefore, the method with the lowest risk and cost is the one closest to the origin. Metering and the two combination methods have low risk but significantly different costs. Engineering-tracking has extremely low costs but extremely high risk and HSEM has low cost but medium risks. While XENERGY recommends the lowest cost, lowest risk strategy this information will allow others to perform their own cost-risk tradeoffs.

It is important to recognize that these results are based solely on the analysis of one data set. Therefore, we believe they should be used to consider selections of methods but should not be extended to mean that these sample sizes guarantee the target precision. Further investigation and quantification of these relationships is currently being pursued by XENERGY with several utility clients. These efforts are being designed to test the assumptions presented here in order to verify the efficacy of the approach and its value in evaluation research.

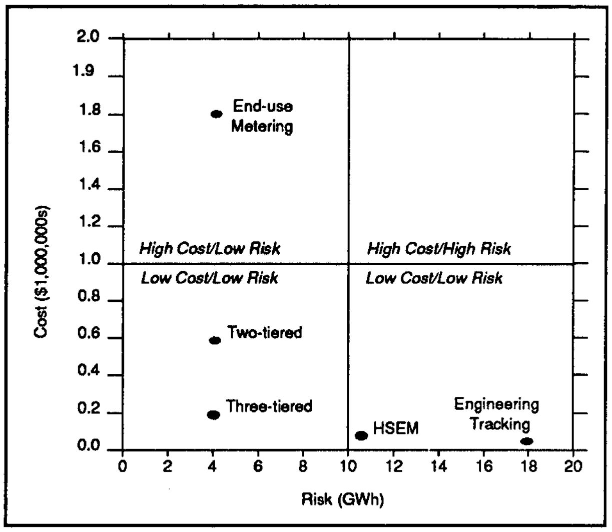

Figure 3. The Cost/Risk Trade-off Comparing Individual and Combined Methods:Lighting Impact Evaluation, Peak Hour/Summer Average Day

References

(1) Andrew Schon and Karen Hamilton. "Commercial Sector Alternatives to End-Use Metering," *End Use Load Information and Its Role in DSM* (Conference Proceedings) Irvine, CA, 1990

(2) Kenneth Train. "Statistically Adjusted Engineering (SAE) Models of End-use Load Curves," *Energy*, Vol. 10, No. 10, 1985.

Bibliography

Commonwealth of Massachusetts DPU Petition 89-260, June 1990.

Michaels, Harvey, Cary G. Bullock, Jr., and David L. Wang. "Techniques for Estimating Energy Use in Residential, Commercial, and Industrial Buildings," XENERGY Inc., March 1984.

RCG Hagler, Bailly Inc. New Jersey Conservation Analysis Project: Final Contractor's Report to the NJCAT, New Jersey Conservation Analysis Team, August 1990.

Train, Kenneth E. "Correcting Self-Selection Bias in the Estimation of Audit Program Impacts," Cambridge Systematics, Berkeley, CA.

Violette, Daniel, and Michael Ozog. "Use of End-use Load Research Data in Statistical/Econometric Evaluations of DSM Programs," RCG/Hagler, Bailly, Inc.

Wright, Roger L., and Curt Puckett. "Conservation and Load Management Impact Evaluation Plan 1990-1994," Northeast Utilities Service Co., 1989.

Wright, Roger L., and M. W. Townsley. "Measuring DSM Impacts: End-Use Metering and the Engineering Calibration Approach," *End Use Load Information and Its Role in DSM* (Conference Proceedings), Irvine, CA, 1990.