The Comprehensive Approach to Commercial New Construction Program Impact Evaluations - Lessons Learned in California

Douglas Mahone and Catherine Chappell, Heschong Mahone Group, Fair Oaks, CA
Roger Wright and Edward Erickson, RLW Analytics, Sonoma, CA
Pete Jacobs, Architectural Energy Corp., Boulder, CO
Andrea Horwatt and Martin Morse, Southern California Edison, San Dimas, CA
Valerie Richardson, Pacific Gas & Electric, San Fransicso, CA

ABSTRACT

Two major utility companies have twice conducted impact evaluations of their nonresidential new construction programs, following California's Protocols for measurement and evaluation. Several important methodological issues are discussed, including: sampling of participant and nonparticipant buildings, collection of both on-site and decisionmaker data, calibration of building simulation models, and determination of net-to-gross savings ratios. The paper presents lessons learned from these state-of-the-art impact studies.

Introduction

This paper discusses lessons learned from four demand-side management (DSM) impact studies performed cooperatively by the Southern California Edison Company (Edison) and the Pacific Gas and Electric Company (PG&E). The studies evaluated the two utilities' nonresidential new construction (NRNC) programs for their program years 1994 and 1996. During these program years, the utilities' NRNC programs included incentives for a mix of prescriptive and performance measures, and design assistance.

Background

In California, the state's four investor-owned utilities (IOUs) have a regulatory framework which permits shareholder earnings from their DSM programs. To ensure that regulatory reporting needs for these programs are adequately addressed, in 1992 a collaborative of the IOUs, regulatory community, and various stakeholder organizations jointly developed the *Protocols and Procedures For The Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs* (M&E Protocols). The M&E Protocols prescribe rigorous methodologies to be used by the IOUs to document and verify the costs and benefits of their DSM programs.

The evaluation of nonresidential new construction has been significantly impacted by evolution of the M&E Protocols. In 1996, the Protocols were modified to change the unit of analysis for NRNC impact evaluations from measure-level to whole building-level. As a result of this change, measures recommended or rebated by the utility programs as well as measures not recommended or rebated by programs can be captured. This whole building approach also enables evaluators to accurately capture the overall energy savings of buildings which exceed California energy-efficiency building standards (Title 24) in one end use, but are below standards in another.

Objective

The prismary objective of these studies (see References) was to measure the first year gross and net impacts (both energy anddemand) of the programs, and to do it accurately and economically.

These studies also attempted to go beyond previous NRNC impact studies of PG&E and Edison in capturing better information on indirect program effects, particularly participant and non-participant spillover.

The authors used fundamentally similar methodologies for both the 1994 and the 1996 studies, but made improvements the second time around which produced better studies and more accurate results. This paper will not be reporting the findings of the studies, but rather will highlight and discuss the most important methodological improvements.

Overview of Methodology

The methodology for all four of the impact studies (both utilities, both years), relied on extensive data collection, detailed energy simulations to produce gross savings estimates, and econometric analysis to arrive at net savings estimates.

The studies followed these basic steps:

1. Draw a representative sample of the program participant buildings, and develop a comparable sample of nonparticipant buildings.
2. Recruit the sampled buildings to cooperate with the study.
3. Collect decisonmaker data by phone for all sampled buildings - information on how decisions to implement energy efficiency were made, characteristics of the owners, building types, etc.
4. Collect detailed on-site survey data for each building - physical dimensions, building envelope, lighting system, mechanical system, schedules, etc.
5. Construct DOE-2 energy simulation models of each building and calibrate to billing or metered data.
6. Calculate gross savings at the whole building and end-use levels for each building by comparing as-built to baseline efficiencies of all measures (both incented and non-incented).
7. Estimate gross savings for the populations of participants and non-participants, using statistical expansion techniques.
8. Estimate net program savings based on comparisons of participant and non-participant populations.

The key areas of discussion for this paper, based on the most important methodology improvements are:

- Sampling Issues - how to most effectively draw the participant and nonparticipant samples
- Data Collection - improvements in accuracy and effectiveness for new buildings, both with on-site data and with decisionmaker data.
- Modeling and Calibration - efforts to improve model accuracy through calibration techniques
- Net Savings and Spillover - two different ways to estimate: difference of differences, and econometric techniques to identify free-ridership and spillover

Sampling Issues

This section will discuss the key issues in selecting the sample of buildings to be studied. The foundation of these studies was the sample design and sampling execution The distinction is that the sample design sets the goals for recruiting the sampled projects into the study, while the sampling execution determines how well these goals are met.

Sampling vs. Census

In evaluating a program with a relatively few number of participants it is necessary to decide between a sampling approach and an attempt at a 100% census.[1] A census has the apparent advantages of simplicity and statistical precision. However, an attempted census may actually yield poorer information if (a) the census is incomplete, especially if some of the largest projects are omitted, or (b) there are fewer resources available to assure accurate and detailed data collection.
By contrast, sampling can be used to deliberately limit the number of projects in the study. The sampling approach focused on a smaller number of projects, but included a high proportion of the largest projects. With sampling, both recruiting and data collection could be more careful and detailed. This helped control bias from non-response, self-selection, and measurement error.

In the case of PG&E, for example, the 1996 program had 392 participants. Only 141 of these were in the final sample, but we achieved excellent response rates and statistical precision from the sample. Moreover, a census of the 392 participants would probably have doubled the cost of the study unless we had seriously compromised the quality of our recruiting, data collection, and analysis. So we feel that a census approach would have seriously jeopardized the study.

Importance of the Sampling Frame

The 1994 evaluation initially tried a new approach to sampling, because it was thought that the program penetration rate was high. An attempt was made to utilize the Dodge new construction database[2] as a sampling frame[3] for the participants as well as the nonparticipants. The idea was to draw a single sample, and then determine after the fact whether each sample project was a participant, partial participant or nonparticipant. The objective was a representative sample of all construction in the state.

That approach was not actually followed because it appeared likely that there would be too few participants in the sample; penetration rates turned out to be much smaller than assumed. So in 1994 we followed a modified approach in which we tried to identify the program participants in the Dodge data base, and then use a single sampling plan to select both the participants and nonparticipants. This allowed us to control the number of participants in the sample, but the process of matching program participants to the Dodge database was tedious and error prone and led later to confusion in the field.

There was an equally serious but more subtle problem. In 1994, both the nonparticipant and participant samples were stratified by building type and an estimate of square footage developed from the information in the Dodge data base (reported square footage data was incomplete). The stratification resulting from this sample design was relatively ineffective leading to poorer than necessary statistical precision, especially for the participants.

The root of the problem was that the Dodge new construction data base lacked any stratification variable[4] comparable in effectiveness to the estimate of savings found in the program tracking data base. With the ratio estimation methods used in this project, the primary purpose of stratification is to provide appropriately high sampling fractions for the small number of very large participants and low sampling fractions for the numerous smaller participants. Here the size of a project refers to its actual savings. So the ideal stratification variable would be highly correlated with the actual savings of the

---

[1] Table 5 of the California Evaluation Protocols specifies that a census should be attempted for any nonresidential new construction program with fewer than 450 participants. These projects waived that requirement in favor of a rigorous sampling approach and greater attention to data quality.

[2] For the 1994 study, we used Dodge data listing all major construction projects in California in 1992-94.

[3] A sampling frame is the dataset from which the sample is drawn.

[4] A stratification variable is used to segment the sample frame into smaller groups. Stratification is generally done to create more homogenous groups that require smaller samples to estimate parameters.

project. The Dodge data base usually included an estimate of the cost of the new construction, and sometimes an estimate of the square footage of the project. The 1994 sample design was stratified for both participants and nonparticipants using this Dodge-based estimate of square footage. But for program participants, our Dodge-based square footage was found to be very poorly related to the actual energy savings due to the program.

By contrast, the tracking estimate of savings was much more highly correlated with the actual savings. Building on what we learned from the 1994 study, the sample design for the 1996 program participants used the program tracking system directly as a sampling frame. This eliminated the problem of matching participants with the Dodge data base and improved the field work. Equally important, the 1996 participant sample design was stratified by the estimate of savings found in the tracking system. This greatly improved the statistical precision of the savings estimated from the participant sample, with little impact on the cost of the project. The improvement in statistical precision was roughly comparable to increasing the sample size four-fold.

Choosing the Sample Size

When a sampling approach is taken, an important issue is specifying the sample size. On the one hand, if the sample is too small, the results are not statistically reliable. On the other hand, if the sample is too large, resources are wasted.

In these evaluation studies, the sample sizes were estimated using Model-Based Statistical Sampling (MBSS™) techniques.[5] We adopted the following criterion: the participant sample should be large enough to estimate the gross savings of participants within ±10% at the 90% level of confidence. Under the MBSS approach, a statistical model is specified to relate the target variable – the gross savings found in the evaluation of each sample site – to the stratification variable – the tracking estimate of savings of each site. The variance in this relationship is measured by a parameter called the error ratio. For example, an error ratio of 50% means, roughly speaking, that the standard deviation of the errors in the relationship is about one-half of the expected value of the measured gross savings.

Given the assumed value of the error ratio – and of an even more esoteric but fortunately rather stable parameter that we will not discuss here – and given the distribution of savings in the tracking system, it is straightforward to develop an efficient sample design and to calculate the sample size required to provide the expected statistical accuracy. This methodology adequately addresses several important issues including: (a) the need to estimate savings rather than use, (b) a way to factor in the efficiency of the sample design, and (c) the known information from the program tracking system.

MBSS uses the error ratio[6] to estimate the variance of unobserved variables. Often the error ratio can be judged rather accurately from consideration of each specific project. Objective estimates of error ratios can often be developed from the data of prior studies. In planning the 1996 evaluation studies, we drew on the error ratios estimated from the 1994 sample data. Even though, as discussed above, the 1994 sample was not well stratified, we were able to estimate the underlying error ratios. From these, we were able to quantify the reduction in the sample size due to efficient stratification using the tracking data base, and to choose the sample sizes required in the new studies for both PG&E and Edison. In both of these studies the achieved precision was very close to our predictions.

---

[5] *Methods and Tools of Load Research, The MBSS System, Version V.* Roger L. Wright, RLW Analytics, Inc. Sonoma CA, 1996.

[6] The error ratio is a measure of the variance around the trendline. It is analagous to the coefficient of variation.

Matching Nonparticipants to Participants

In order to obtain good estimates of net savings, it is generally thought to be useful to match the participant and nonparticipant samples. The size of the nonparticipant sample was chosen to be roughly equal to the participant sample. We also wanted the type of sites in the nonparticipant sample to be comparable to the participant sites. However, since the participant sample was stratified by the tracking estimate of savings, we could not use the same sampling plan for the nonparticipants.

We developed the nonparticipant sampling plan to match the participants by following the suggestion of an outside reviewer. The nonparticipant sampling plan had to be limited to the information available in the Dodge data base – building type and an estimate of square footage. We used the program-participant tracking system to develop an efficiently stratified sampling plan based on building type and square footage. Then we classified the Dodge sites into these strata. Finally we randomly selected the nonparticipant sample following the allocation determined by our sampling plan. This effectively addressed the essential challenge – that the nonparticipant sample should closely represent the participant population.

Data Collection

The data collection procedure varied between the 1994 and the 1996 study. The various data collection and modeling responsibilities in the 1994 study were assigned to separate contractors. An integrated approach to data collection was used in the 1996 study. The recruiter, the auditors, and the modeling analysts worked together to ensure that all of the necessary data was collected efficiently, and that errors were quickly caught and corrected.

The key difference between the field data collection process used in the 1994 and 1996 studies was that, for the 1996 study, the building models were constructed and reviewed by the auditor shortly after the on-site visit. This process dramatically improved the team's ability to produce models that accurately reflected the building's actual operating conditions. It also allowed for quick feedback from the modeling to the on-site data collection effort, allowing for quick resolution of any data collection problems. By contrast, after the 1994 on-site data was collected, it went through a lengthy process of data entry and cleaning. By the time it was handed over to the analysts, several weeks and many additional on-sites had passed, and the data had to be digested by a new group of modeling analysts, before errors were identified. This made it difficult and expensive to correct the errors.

The overall quality of the simulation models developed for the 1996 studies improved dramatically. The error ratio, a statistical measure of scatter between the model and the expected results decreased from 30% in 1994 to ~8% in 1996, thus improving the overall precision of the savings estimates. The improved surveying, modeling, and quality control procedures implemented for the 1996 studies are thought to be responsible for this improvement.

In addition to the on-site engineering data collection process, improvements were made to the decisionmaker survey process.

For the 1996 study, the decisionmaker data was all collected by a single interviewer who was knowledgeable about NRNC and familiar with the target population of decisionmakers. By contrast, the 1994 study data was collected using telemarketing procedures and less knowledgeable interviewers. The 1996 procedures produced more consistent decisionmaker data, because one interviewer was responsible for all of it. This interviewer was also able to achieve a 100% response rate (a completed decisionmaker survey for each site in the sample), which is difficult with nonresidential building developers and owners. By contrast, there was only a 66% response rate in the 1994 study. This reduced the accuracy of the entire 1994 study net savings results.

The 1996 decisionmaker survey also made use of a scaled response to questions, rather than a yes/no response as was done in 1994. For example, when asked if the utility program representative had had a significant influence on building efficiency, in 1994 the response was yes or no; in the 1996 study, the response was given on a scale of one to seven. This scaled response provided better information to the econometric net-to-gross analysis, and was less conservative than the 1994 approach.

The authors believe that the 1996 study's approach to data collection both improved the quality of the field data used in the engineering analysis, reduced the time required to conduct the data collection as compared to the 1994 study, and also produced more complete and consistent decisionmaker data for the net analysis. The strengths of this approach were:

- The data collection timeframe for any single site was much shorter than in the 1994 study. This minimized the burden on the customer.
- Experienced DOE-2 modelers performed the on-site audits. They were also the users of the data, so they knew what they needed.
- The as-built DOE-2 models were generated and examined soon after the on-sites. This allowed any issues to be addressed quickly, when the auditor was likely to remember details about the site and when the customer was more likely to be receptive to follow up contacts.
- Every site had decisionmaker data, gathered by the same knowledgeable interviewer, with more informative, scaled responses to questions.

Modeling and Calibration for Gross Savings

The following subsections describe how building energy simulation techniques were used in this study. The gross savings estimates were developed using engineering analysis techniques based on the on-site survey data that was collected.

Generating DOE-2 Models

Engineering models were developed for each building in the on-site survey sample using the DOE-2.1E building simulation program. This program was used because it is widely accepted as providing reasonable whole building analysis results, and it has the capability to model virtually all of the energy features of the buildings we were studying.

An automated process was used to develop basic DOE-2 models from data contained in the on-site survey, program information and other engineering information. The "deck generation" software took information from these data sources and created a DOE-2 model. This process, besides being quicker and more economical than building models by hand, allowed for greater consistency in the modeling, and allowed us to make changes to the analysis process in a fairly automated fashion.

Once the models were calibrated and quality checked (see next section), an automated batch process was used to create a series of parametric simulation runs. These runs were used to estimate gross savings for participants and nonparticipants on a whole-building and measure class basis. The runs included:

- An "as-built" model representing the building as found by the surveyors.
- A "baseline" model representing the building with equipment and envelope efficiencies as specified by Title 24 energy code requirements.
- A series of parametric runs to isolate the savings attributable to motors, lighting, HVAC, shell/daylighting and refrigeration end-uses.

The models were developed using an automated BDL[7] generator, developed by AEC and RLW Analytics. This method ensured that all of the models were consistent, thus eliminating a potential source of bias in the results.

Calibrating the Models

A persistent question, in using building simulation techniques, has to do with the reliability of the models and their ability to accurately replicate building energy use magnitudes and patterns. In order to assure accuracy, the models were calibrated and the results they generated were reviewed for reasonableness.

Over the course of these studies, a variety of calibration techniques were tried.

Billing Data. The primary calibration technique involved the use of utility billing data for the sampled customer sites. Monthly energy consumption and demand from the DOE-2 models were compared to billing data (using weather data and billing data for the same historical time period) to assess the reasonableness of the models. Adjustments were made to a fixed set of calibration parameters until the models matched the billing data. The goal of the calibration process was to match billing demand and energy data within ±10 percent on a monthly basis.

In order for the comparison between simulated electricity consumption and billing data to be meaningful, there needed to be a good match between the surveyed space and the space served by the revenue meter. For some buildings, the space treated by the program was less than the space served by the meter. An example of such a mismatch is a major tenant improvement or tenant finish in a multi-tenant building.

During the on-site survey, the surveyors collected meter number information, and assessed the match between the space served by the meter(s) and the surveyed space. For the 1994 study, less than half of the surveyed sites had usable billing data for calibration. The lack of useable billing data was attributed to poor matching between metered spaces and surveyed spaces, errors in meter number transcription, meters not accessible during on-site surveys, and technical problems with the data received from the billing system. For the 1996 study, the billing data capture rates were significantly improved. This improvement is largely attributed to improved training techniques for meter number identification. Thus, the modelers were able to successfully calibrate more sites, as shown in Table 1.

Table 1. Model Calibration Success

Study Year	Calibrated	Unable to Calibrate	No Data
1994	35%	10%	55%
1996	60%	20%	20%

In the 1996 study, about a fifth of sites that had useable billing data resisted reasonable attempts at calibration. There are many combinations of input parameters that can be altered in a model to force the results to match billing data. Improper calibration actions, however, can degrade the ability of the model to predict energy savings, thus the modelers were instructed to not make unreasonable changes to the models during the calibration process.

[7] BDL is DOE-2's Building Description Language

During each study, a major level of effort was expended in the model calibration process. For the 1996 study, hourly weather data for 1996 and half of 1997 were obtained for 54 weather stations throughout California. Billing data were extracted for the same periods for all sites with meter numbers. The modelers ran separate simulations for each weather year, and compared results at each calibration step to the available billing data.

To assess the impact of this effort on the energy savings predictions, the calibrated models' results were compared to their uncalibrated results. The impact of the calibration process on the overall results was fairly small (approximately 7%), and not significant with respect to the precision of the overall savings estimates. In the end, the calibration process served primarily as evidence that the modeling effort produced realistic results.

Short Term Metering Data. In the 1996 study, for the sites where the surveyed space and the metered space did not match, short-term metering data were used instead of billing data to calibrate the DOE-2 models. The short-term metering equipment was installed on the circuits feeding the surveyed space only, thus serving as a temporary "proxy" meter for the surveyed and modeled space. The surveyors assessed the feasibility of installing short-term metering equipment during the survey. The electrical panels serving the surveyed space were identified, and sites with fairly "clean" circuitry were identified. If the site appeared to be a reasonable candidate, the surveyor recruited the site for short-term metering.

The results of the short-term metering (STM) for the 1996 study were somewhat disappointing. The overall results of recruiting and equipment installation is shown in Table 2.

Table 2. Short-term Monitoring Installations for 1996 Study

	Successful Installation	Dropped During Installation	Dropped During Recruiting	Total STM Candidates
PG&E	7	3	15	25
SCE	7	5	15	27

About 60% of the sites identified during the survey were dropped during recruiting. Most of the dropouts were due to customer refusal or conditions on site that made installation infeasible. Several sites dropped out during installation, due to severely mixed circuiting or unsafe installation conditions. Overall, the impact of seven additional calibrated sites on the overall study results for each utility was insignificant.

In the 1994 study, the primary objective of the monitoring was to define building performance parameters that were important for creating the DOE-2 models, but which were not observable during an on-site survey. These parameters tended to be concentrated in buildings with complex HVAC, refrigeration, and/or control systems. Short-term end-use monitoring was done on a sample of 30 buildings. The data were then used to calibrate the engineering models of each monitored building. Since buildings with complex HVAC and/or refrigeration systems tended to be larger sites with large expected savings, the measurements and calibrated models were thereby targeted at a large portion of the total expected savings.

We believe that the short-term monitoring approach followed in the 1994 study was more effective. This approach, while more expensive, yielded good information about building operations and equipment performance that was used to improve the simulation models. It essentially increased our engineering knowledge about parameters that are important in the simulations, but which cannot

readily be determined through survey techniques. It is still difficult to generalize these findings to the population of buildings, but when applied to the major savings sites, the results are made more accurate.

Calculating Gross Savings

The estimates of gross program savings were made by comparing the as-built simulated building energy consumption to a baseline level of energy consumption. The baseline energy consumption for all buildings was defined to be the energy consumption of the building as if all of the equipment was specified to be minimally compliant with Title 24 and the building was operated on the schedule found during the on-site survey.

A gross savings estimate was calculated for each building in the sample. The savings estimated were projected to the population of participants using the results of the model-based statistical sampling procedures. Gross savings estimates were developed for both the participant and the non-participant population.

Net Savings and Spillover

The net savings of the program are defined to be the true savings of the participants due to the program, relative to what they would have done in the absence of the program, plus any spillover savings among the nonparticipants. The challenge of estimating net savings is to determine what both the participants and nonparticipants would have done without the program – since this is not directly observable.

One simple approach to estimating net savings is to ask participants directly or indirectly what they would have done without the program. Under some conditions this seems to be a pretty good approach but it is discouraged under the California M&E Protocols. Instead, two other approaches are available, the difference of differences approach and econometric estimation.

Difference of Differences vs. Econometric Estimation

The difference of differences approach involves three steps. (a) estimate the energy efficiency of the participants as a fraction of the baseline use, (b) estimate the energy efficiency of the nonparticipants as a fraction of the baseline use, and (c) estimate the net savings as (a) – (b). The idea is to use the nonparticipant sample as a comparison group to shed light on what the participants would have done without the program.

The difference of differences approach can provide an unbiased estimate of net savings if (a) participants and nonparticipants are similar, (b) there is no self selection bias among participants, and (c) there is no spillover savings among the nonparticipants. More generally, the difference of differences approach can be unbiased if the two samples are well matched and the amount by which participant savings is overestimated due to self selection is exactly offset by the spillover savings among the nonparticipants.

To address the potential bias in the difference of differences approach, econometric methods are often used to estimate net savings. In our studies, we used a two-step modeling approach. First, a logistics model was formulated to estimate the probability that each site in the joint participant/nonparticipant sample did participate in the program. This predicted probability of participating was used to calculate Mills and Double Mills variables.

Second, we formulated and estimated a linear regression model predicting the efficiency choice of each project in the joint sample. The dependent variable was the efficiency of the site relative to the

baseline energy use of the site. For example, a value of .05 indicated that the as-built energy use of the site was 5% lower than the baseline use. The explanatory variables included the following types of factors: (a) characteristics of the project and the designer not affected by the program, such a the type of building, the financial criteria used to choose between features, etc., (b) factors directly reflecting the program, such as an indicator for program participation, and the response to questions about the degree of interaction with the utility and the amount of influence by the utility on the project, and (c) the Mills and Double Mills variables. Four models were developed - for both energy and demand savings for each of the two programs.

Under appropriate statistical assumptions, these models can yield unbiased estimates of the true net savings among the participants after adjusting for free ridership, and the spillover savings among the nonparticipants. The approach is fairly simple. We estimated the impact of the program on each sample site by comparing the efficiency of the site predicted by the model to the efficiency that would have been predicted in the absence of the program. In effect, we plug into the regression equations the actual values of the explanatory variables for the site and then substitute the adjusted values of the explanatory variables assuming the program did not exist. We multiplied the baseline use of the site by the estimated change in efficiency due to the program in order to convert the results to kWh and kW savings. Finally we used our standard statistical methods to expand the results from the sample to the population of all program participants and nonparticipants.

Nonparticipant Spillover

All four of these regression models indicated substantial spillover savings among nonparticipants. This modeling result was due to a positive, statistically significant relationship between the efficiency of the nonparticipant site and the degree of influence of the utility on the design of the project as reported in our decision-maker survey. Holding other variables fixed, the efficiency of the project was found to be about 0.10 higher for a nonparticipant who reported being very strongly influenced by the program compared to a nonparticipant who reported not being influenced at all. Remarkably, we found this same result for both energy and demand savings in both programs.[8]

The actual estimate of the spillover savings reflected the degree of influence reported by each nonparticipant. No spillover savings was claimed for nonparticipants who reported that the influence was small. Substantial savings could be achieved only if large numbers of nonparticipants reported being influenced. In fact, the spillover component was a substantial component of the net savings in both programs. These variables were statistically significant at the 10% level or better.

In the two 1996 impact evaluations, completely independent samples were analyzed using the same data collection and analysis methodology. Essentially the same results were obtained in the two studies. Since independent replication is the strongest form of validation, we consider these results to be strongly validated.

The estimates of spillover savings carried relatively poorer statistical precision than the estimates of participant net savings. This was due to the limitation of sampling nonparticipants using the Dodge data base and the associated large variances of these data, as discussed in the Sampling Issues section above. Even so, the econometric estimates of total net savings were quite accurate, in the ±20% to ±30% range. Because of the spillover savings, the econometric estimates were somewhat higher than the difference of difference estimates, suggesting that the later might tend to understate the true impact of these programs.

[8] These variables were statistically significant at the 10% level or better. The same steps were followed to develop all four of these models.

Conclusions

To summarize the primary lessons derived from the efforts to improve the methodology between the 1994 and 1996 studies:

1. Sampling - The foundation of an impact study is the sample, since we can't revisit every project in the program. The sample needs to be cost-efficient, it needs to represent the entire program population, and it needs to include nonparticipants which are comparable to the participants. We have learned that most reliable starting point for sampling is the program tracking database and its estimates of savings for participant buildings. This allows a very good sample of participants to be drawn. Then, a comparable sample of nonparticipants can be drawn from the Dodge database, using the same characteristics as the participants in the database.
2. Data Collection - The next most important element of an impact study is good data. Since data collection is expensive, the challenge is to get the best data you can within the time and money constraints of the study. Based on improvements in our data collection methods between the 1994 and 1996 studies, we have learned some important lessons in how to meet this challenge. We learned that it is worth using experienced building model engineers to collect and clean the data, rather than lower cost surveyors. It is also important to avoid time delays between data collection and its use in modeling, as this makes it easier to catch and correct errors. For decisionmaker data, used to estimate free-ridership and spillover, the simple yes/no approach to questions used in the 1994 study was not able to produce significant effects, but the scaled responses used in 1996 study were able to do so.
3. Model Calibration - Since our gross savings estimates are calculated with simulation models, we initially believed that calibrating those models was going to be important. We learned, however, that if high quality data and models can be obtained, the need for calibration of those models diminishes; the additional accuracy to the savings estimates is not significant. Nevertheless, calibration increases confidence that the engineering analysis was done accurately. We ended up going through a calibration step anyway, but it had little effect on the outcome of the study.
4. Net-to-Gross Analysis - The final step in the analysis attempts to account for non-engineering factors such as free-ridership and spillover. There is still disagreement in professional circles on the best way to calculate the net savings, so we tried two approaches. The simple difference of differences approach underestimated overall savings and did not provide strong statistical significance in its estimate. The econometric approach explicitly captured free-ridership and significant spillover effects, with acceptable statistical precision. The Double Mills Ratio approach, favored by the California M&E Protocols, did not appear to be a reliable method for improving estimates. Using the two approaches hasn't settled the professional disagreements as to the best approach, but calculating the results both ways did give us confidence that we had arrived at sound conclusions about the program impacts.

Based on the experiences gained in conducting these studies, the authors feel they have refined a reliable and accurate methodology for estimating the net impacts of nonresidential new construction programs. These programs, due to the comprehensive nature of the energy efficiency improvements they promote in a wide range of new building types, are especially challenging to evaluate. The methodology discussed here may not be the lowest cost approach, but it builds on widely accepted evaluation Protocols, is grounded in sound engineering analysis, uses high quality data as the basis for estimation, and allows for good confidence in the reliability of the findings.

References

Pacific Gas & Electric Company 1998. *Impact Evaluation of Pacific Gas & Electric Company's 1996 Nonresidential New Construction Program.* Study ID Number 389, prepared by RLW Analytics, Inc.

Southern California Edison 1998. *1996 Non-Residential New Construction Evaluation Final Report.* Study ID Number 543, prepared by RLW Analytics, Inc.

Pacific Gas & Electric Company 1997. *Impact Evaluation of Pacific Gas & Electric Company and Southern California Edison 1994 Nonresidential New Construction Programs.* PG&E Study ID Number 323, SCE Study Number 522, prepared by RLW Analytics, Inc.