Evaluating the Impacts of an Energy Study Program

A Case Study

Dakers Gowans, Nexant, White Plains, NY

ABSTRACT

This paper is a case study of the challenges that were addressed while conducting an impact evaluation of the Technical Assistance Program (TAP), a systems benefit charge program operated by the New York State Energy Research and Development Authority (NYSERDA). The TAP provides funding to conduct energy feasibility, rate analysis, and operational studies as well as walk-through audits of small facilities, and it accounts for nearly a third of the annual kWh savings reported by NYSERDA's portfolio of programs. The evaluation was conducted in 2004 and then updated in 2006; prior to this period impact estimates were based on responses to a telephone survey of program participants.

The purpose of the evaluation was to obtain the best possible estimate of the energy (kWh/year, MMBtu/year) savings and demand (kW) reductions that have resulted from the program's operations. TAP participants are not required to implement any recommendations made in their feasibility studies, nor are they required to report installation activity back to NYSERDA if they do elect to act on the recommendations. A detailed research plan was developed in 2004 specifically addressing this absence of a continuing relationship between the TAP and its customers. The plan called for sampling the program's 827 completed studies (over 1,027 in 2006) by stratifying the population between large projects with recommended savings greater than 1,000,000 kWh/year, and all smaller projects. In order to focus resources on studies that contribute the greatest risk or value to the program's reported savings the sample was weighted so that 80% of the sampled reports were drawn from the stratum of large projects.

In 2004, a significant challenge for the sampling plan was the lack of a comprehensive tracking data base for the program. The solution was to draw samples of studies in three rounds using the best information about the population that was available at the time of the draw, while simultaneously populating a tracking database with information pulled from hard-copy file records. To ensure that the sample design criteria were maintained, sample strata were balanced at each draw by using the latest version of the database.

A key element of the research plan developed in 2004 to address the voluntary and un-reported implementation of the recommendations made in the studies was to disaggregate the standard realization rate into two components. These are: the Measure Adoption Rate (MAR), the ratio of self-reported installed measures to study-recommended measures; and the Savings Realization Rate (SRR), the ratio of installed kWh/year as determined through engineering reviews and site visits, to the study-estimated kWh/year for the measures reported by customers to have been installed. Developing the MAR, which was based on customer response to a telephone questionnaire, allowed investigators to identify projects for site visit review. Finally, project and program realization rates were the product of MAR times SRR.

Introduction

A comprehensive impact review of the Technical Assistance Program (TAP) operated by the New York State Research and Development Authority (NYSERDA) was completed in March of 2005,

and again in March of 2007. While the purpose of the reviews was to obtain the best possible estimate of the program's savings, the challenges faced by the investigators and the solutions that they devised will be of interest to the evaluation community. The review's conclusions will also be of interest to other energy efficiency program administrators faced with the problem of estimating the impact of an energy audit or energy operations study program that does not require customers to adopt any of the study's recommendations.

NYSERDA is the administrator of **New York Energy $mart**[SM], New York's Public Benefits Charge (PBC) Program, which is funded though a system benefit charge initiated in 1996 by an order of the New York State Public Service Commission. NYSERDA began operating the PBC Program in 1998, and the TAP was one of the inaugural offerings.

The TAP provides cost shared funding to end-use customers for feasibility studies, energy operations management reviews, rate analysis and aggregation studies. Funding in most cases is up to 50% of the total study costs, with a maximum contribution of $50,000. The program also funds walk-through energy audits for small facilities (<$75,000 annual electricity costs), though these were not included in this evaluation review due to their small (<5%) contribution to the program's impacts. As of the end of 2006, the program had supported 1,024 studies (excluding the walk-through energy audits) and had contributed approximately $16,000,000 towards their cost.

The TAP does not require end-use customers to adopt any of the recommendations that are made in their study. However, about 60% of these customers have implemented at least one recommended measure; when considering studies that were completed 3 or more years ago this ratio rises to 81%. NYSERDA believes that the cost share requirement is key to assuring customer commitment to the study process and to implementing cost-effective recommendations that support their business goals.

Evaluation work prior to 2004

Between 1994 and 2004[1] NYSERDA estimated TAP impacts based on 4 rounds of telephone surveys of program participants. Participants, who were randomly selected from a population of studies completed one to three years before the survey start date, were asked which of their recommended measures they had implemented. The survey data were collected entirely from the self-reported responses to the telephone survey; no 3rd party confirmed the accuracy of the responses by conducting on-site verification visits. Based on the participant-reported results, NYSERDA developed kWh and BTU per NYSERDA-incentive-dollar ratios for all implemented measures. NYSERDA counted measures under contract for construction as completed, giving the indices a forward-looking component. In 2003 NYSERDA estimated that 27 kWh of electric energy savings was realized for every $1 of TAP funding.

NYSERDA significantly increased its evaluation funding and hired a team of evaluation contractors in early 2003. The measurement and verification (M&V) evaluation of the TAP began in 2003, with file reviews and site visits for 16 randomly selected completed feasibility studies. The sample was drawn from a population of 84 NYSERDA-telephone survey respondents who reported that some or all of their feasibility study recommendations had been implemented or were under construction. However, the sample frame was flawed having been drawn from a subset of telephone survey respondents rather than the entire population of nearly 700 completed studies, and so the results were not statistically valid. The error in defining the sample frame was due to a misunderstanding; the M&V evaluation contractor asked NYSERDA for a list of all completed projects and NYSERDA responded with the 84 telephone survey customers who reported having implemented. The contractor, not fully understanding the scope of the program, failed to grasp that there were more than 600 studies

[1] The TAP has been in existence in one form or another for more than 20 years. Public Benefit Charge funding started in 1998.

unaccounted for. Nonetheless, the work completed in 2003 did have value in that it indicated the need for a comprehensive review because there were significant discrepancies between measures customers reported as operational and what was found as installed during site visits.

Design and execution of the 2004 TAP evaluation

Based on the 2003 review and the experience of NYSERDA's staff, the evaluation team identified a number of TAP characteristics that would require consideration for the 2004 M&V evaluation review. This section discusses them and describes how they were included in the design of the evaluation review.

The program did not have a comprehensive program tracking database, but instead had a collection of data sources, none of which had been continuously maintained and which varied in terms of data quality and completeness. (The exception to this was the NYSERDA funds contributed to each study, which were accurately recorded and had been the basis for past estimates of the program's impacts.) Thus a significant portion of the 2004 evaluation work involved populating an Access™ tracking database that had not been maintained for several years but when completed could be used to draw samples and to aggregate savings impacts according to different criteria such as measure type, customer type, study type, year of study, and facility location. The data were drawn from the hard-copy file record, which fortunately had been carefully archived and was comprehensive for most studies.

Since impact savings were to be based on the estimated savings for measures that were recommended for implementation, rules-based procedures needed to be established to identify the recommended savings amounts in a consistent manner. The following is a discussion of the most significant of these rules:

- Savings values for recommended measures were taken from the final report that was approved by NYSERDA
- In cases where a study examined the potential impact of a planned or contemplated expansion or building addition, the potential savings value for the existing facility was entered into the database.
- For a study with multiple and mutually exclusive options, the value for the shortest payback option was entered into the database, unless another option was clearly recommended in the study for other reasons.
- Only implemented measures or measures recommended for implementation were to count toward the program's impact potential. Measures were classified as recommended, recommended for further study, not recommended, or implemented[2].

Populating the tracking database was a multi-month task, but the review schedule could not afford such a long delay. To overcome this obstacle, the evaluation team used a rolling sampling plan in which one third of the sample was selected during three separate draws over a period of twelve weeks, with the sample adjusted at each draw using the latest available information from the database to ensure that the sample design criteria were maintained. Eventually data for 827 projects and over 5,200 measures were entered and reviewed by the evaluation team, a process that paralleled the evaluation review.

The sample frame was the universe of all completed studies and the task was to quantify the gross kWh, kW and non-electric fuel savings impacts that were realized as a result. The study used the realization rate evaluation procedure where the ratio of the evaluator's adjusted savings impact for the sampled studies was divided by the value recorded by the program. The program's total reported kWh, kW and non-electric energy savings for all projects are then multiplied by the realization rates. The

[2] Occasionally a customer would report having implemented a measure while the study was still in progress.

sample size was designed to meet an 80/20 confidence/precision reporting goal. The coefficient of variation for the 2003 realization rates was 1.02, resulting in a sample size of 43, which was rounded up to 45 of the 827 completed studies that had resulted in one or more of the recommended measures being implemented.

An analysis of the data records that were available at the beginning of the evaluation review demonstrated that 80% of the program's potential impacts were contained in just 20% of the completed studies. An early decision in the sample design was to weight the sample to favor these large contributors by first dividing the population into large and small strata and to then draw 80% of the samples or 36 studies from the stratum of large contributors and the balance from the small contributors. In the 2006 review, the sample included a project that accounted for 12% of the TAP recommended savings. Because this single project dominated the analysis of the large stratum, the evaluation team assigned it to a unique super-large stratum, and developed a rule that any project with recommended savings greater than 4% of the TAP savings would be assigned to a unique stratum of one project.

Because the program's contractual relationship with its customers ends with the completion of the study, the evaluation team had no knowledge about which of all the recommended measures might have been implemented. The sample plan therefore had to find a way to identify 45 completed studies that had resulted in one or more of their recommended measures being constructed. The evaluation team also realized that implementation would be a function of time, with older studies more likely to have adopted a recommendation due to a facility's need to consider, then plan, fund and implement. Earlier program survey results suggested that it took 5 years for the adoption rate to reach saturation, after which little or no TAP study-motivated construction took place. Thus the evaluation team's realization rates needed to include a time dimension that measures adoption rates as a function of years since study completion.

As with any impact evaluation, realization rates include a technical estimate of a project's true savings, the verified savings. For the TAP, this technical component can only be determined for facilities that report having constructed at least one recommended measure from their study.

To capture the adoption rate over time as well as the verified savings impact for each study in the sample, the evaluation team split the realization rate into 2 components; a Measure Adoption Rate (MAR) and a Savings Realization Rate (SRR).

MAR calculations were based on customers' responses to phone interviews conducted by the evaluation team. The goal of the phone survey was to identify facilities that had adopted a recommendation from their study, and to determine what percentage of recommended measure savings each customer had implemented. The telephone survey sample size had to be greater than the design goal of 45 studies that had resulted in savings because some of the facilities would be classified as non-respondents and some respondents would report that they had not adopted any of their study's recommendations. In the end the sample size for the MAR surveys was 170 TAP studies.

The MAR was calculated as follows:

$$MAR_n = \frac{kWh_{Installed}}{kWh_{\mathrm{Re}commended}}$$

Where,

MAR = Measure Adoption Rate

n = Years since study completion, an integer ranging from 0 to 4

$KWh_{Installed}$ = Total kWh savings estimated in the TAP studies for all customer-reported installations

$KWh_{Recommended}$ = Total kWh savings for all measures recommended for implementation in the TAP studies, excluding non-respondents

A MAR of 1.0 would indicate that customers had installed all recommended measures in their TAP studies. If customers reported implementing measures that accounted for only 50% of the total savings for the recommendations in the study, the MAR would be 0.5. In practice, interviewers queried the respondent about each recommended measure in their study, and MARs were calculated at the measure level and then reported as a weighted average for the study.

The SRR, which represents a project's technical realization rate, was calculated using field-verified and engineer-reviewed savings for the phone survey respondents who reported having implemented at least one recommended measure. The SRR captured any corrections to the original TAP study estimates, and corrected for bias or error in the customer-reported adoption rate.

The SRR for the entire field-verified sample was calculated as follows:

$$SRR = \frac{kWh_{Field-verified}}{kWh_{Phone-verified}}$$

Where,

SRR = Savings Realization Rate

$KWh_{Field\text{-}verified}$ = Field-verified kWh savings for all installed measures from TAP study

$KWh_{Phone\text{-}verified}$ = kWh savings estimated in the TAP study for all measures that customers reported as installed, during the MAR phone survey

An SRR of 1 indicates a 100% observation for all three factors. Note that unlike the MAR, the SRR is independent of time.

Finally, the MAR and SRR results were combined to calculate the PRR for each MAR year bin. The PRR was calculated as follows:

$$PRR_i = MAR_i * SRR$$

Where,

PRR_i = Program Realization Rate for studies completed i years ago

MAR_i = Measure Adoption Rate for studies completed i years ago

SRR = Savings Realization Rate calculated from the field-verified sample

Table 1 summarizes the steps used to calculate the TAP realization rate.

Table 1: Steps for PRR calculation

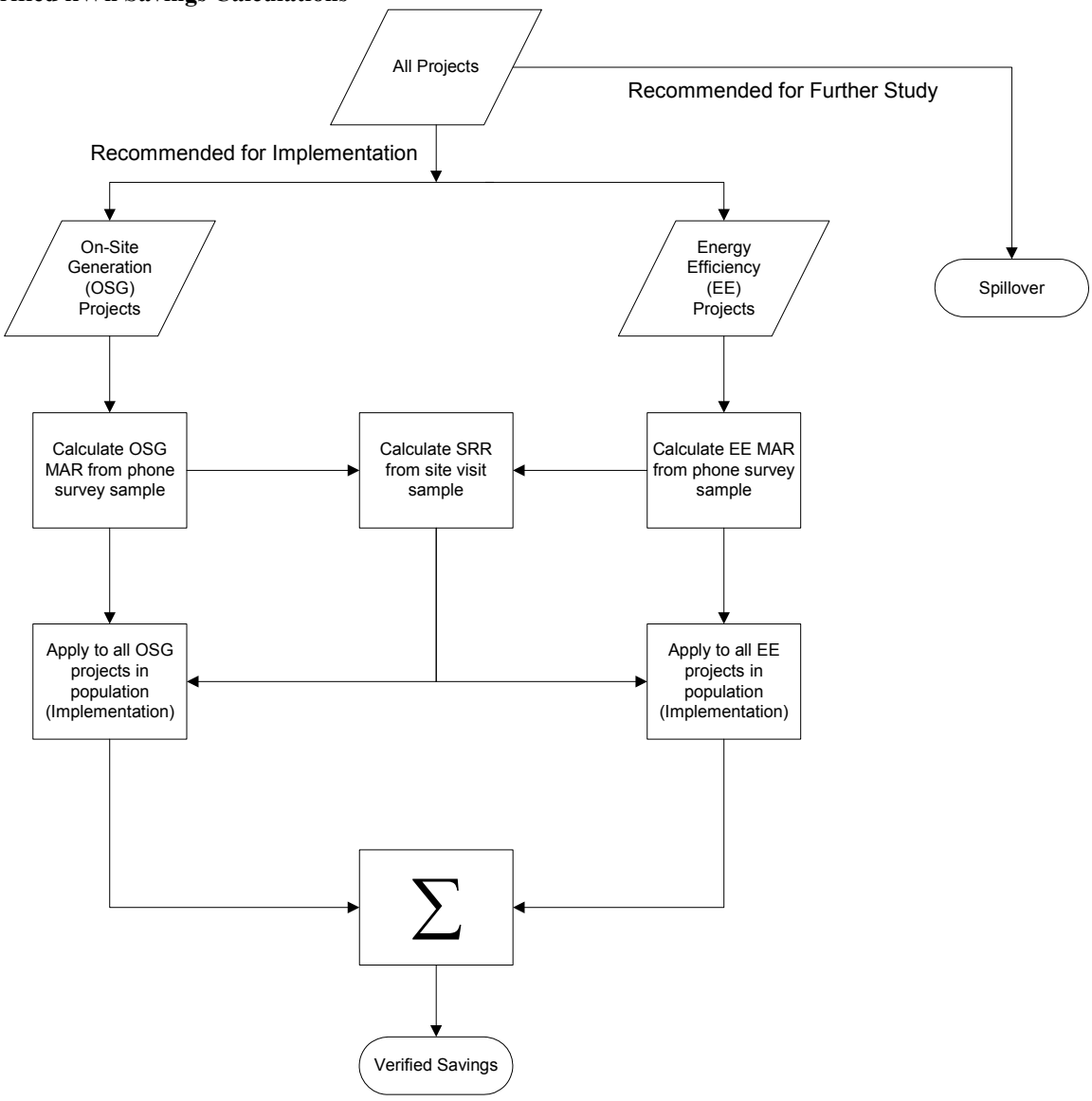
Calculation Step	Years since project completion				
	1	2	3	4	5+
(1) $kWh_{Recommended,\ Phone\ survey\ sample}$					
(2) $kWh_{Installed,\ Customer\text{-}reported}$					
(3) $kWh_{evaluation,\ M\&V\ sample}$					
(4) Measure Adoption Rate = (2) ÷(1)					
(5) Savings Realization Rate = (3) ÷ (2)					
(6) Program Realization Rate = (4) x (5)					

After the 2004 review was underway, the evaluation team realized that TAP studies that investigated the potential for on-site generation, usually co-generation, have distinctly different characteristics than energy efficiency projects, and have significantly larger potential impacts. These are

a result of the complexity, high cost of construction and long planning cycles that are typical of generation projects. Consequently, most on-site generation studies had very different measure adoption rates than energy efficiency studies. The evaluation team's approach for capturing this distinction was to consider OSG and energy efficiency projects as two sub-populations and conduct separate but parallel reviews for both, resulting in separate realization rates.

The procedure for calculating program realization rates is illustrated in Figure 1.

Figure 1: Verified kWh Savings Calculations



Results, 2004

Following the framework described earlier, the TAP evaluation team concluded that as of the end of 2004, the program's participants were saving approximately 454,895 MWh/year[3], which

[3] The 0.51 realization rate was reported with ±0.09 precision at the 80% confidence interval.

represented approximately 51% of the potential savings of measures recommended for implementation in the program's studies. The MWh/year savings represented about 26% of the **New York Energy $mart**[SM] total savings, making the TAP the 2nd largest single contributor to the portfolio, before discounting for participation in another implementation program. Approximately 40% of the TAP savings came from projects that received implementation assistance from another NYSERDA program. The evaluation team also reported that the TAP had achieved approximately 85 MW of summer on-peak coincident demand reduction and 2,236,800 MMBtu/year in non-electric savings[4].

While the 2004 TAP evaluation review successfully increased the reliability of the program's reported savings, the team realized that behind the scenes were both good results and room for improvement.

One of the positive results, and a significant finding, was an SRR of 1.01 ±0.07. The SRR, which represents the percent of realized savings for recommended measures that were reported by customers as being implemented, is a measure of the accuracy of the program's engineering savings estimates. The SRR was based on data collected from 34 site inspections at facilities where program participants reported having implemented one or more measures recommended in a study. These data were augmented with results from an additional 25 TAP studies where the recommended measures were implemented through the Commercial / Industrial Performance Program (CIPP), NYSERDA's standard offer program. The CIPP had been evaluated in 2003 and was shown to be accurately reporting savings (realization rate of 1.02 ±0.03). In addition, each CIPP project was required to perform M&V with results reviewed by a 3rd party overseer, who also conducted baseline and post construction inspections. Knowing that the reported savings were reliable, the evaluation team used the CIPP results in the SRR calculations, avoiding the cost and duplication of additional site visits.

A second good result was the validation of the sampling method that was weighted towards the large contributors. Approximately 68% of the program's total recommended MWh savings were surveyed and included in the MAR calculation. Site visits were conducted at facilities accounting for approximately 13% of recommended MWh savings, and the results used in the SRR calculation. Significant representation like this increases confidence in evaluation results in a way that is more easily conveyed than through statistical analysis, though the team of course depended on statistics to weigh the significance of the findings.

While the basis for the MWh realization rate and adjusted savings calculations was sound, there was a need for additional data and methods to better capture the program's true MW reductions and non-electric (MMBtu) savings. The majority of the TAP studies focused on electric energy savings and often paid less attention to the consequent summer on-peak coincident demand reduction. In many reports, potential kW impacts were often omitted. An exception was a subset of studies focusing on summer on-peak coincident demand, but these comprised only a small portion of the total population. To solve the dilemma of the need to report MW impacts while handicapped by sparse data, the evaluation team developed a kWh/kW ratio of 6,028 based on the field-verified impacts from the 34 site visits. This was understood to be an interim solution, and the program staff intends to improve the reporting, tracking and calculation of this critical performance metric.

Because the TAP is funded with a systems benefit charge on customer's electric bills, by design and necessity it targets electric efficiency investigations. Despite this, the program has resulted in significant non-electric benefits and accounts for approximately 75% of all **New York Energy $mart**[SM] non-electric savings. Furthermore, these savings are net of any increase associated with on-site generation studies, which account for a significant subset of the program's studies. While the 2004 evaluation review resulted in statistically significant results and reported the program's realized non-

[4] Non-electric savings include net reductions in fuel oil, and natural gas, and coal, as well as steam purchased from an off-site plant such as a district steam system.

electric savings, the sample by design was biased towards electric impacts and was not representative of the non-electric measures.

A conclusion of the 2004 evaluation was that the program's savings impact is a function of the rate at which recommended measures are implemented. The SRR is known with good certainty to be essentially 1.0, so once completed, a measure will deliver the savings predicted by the study. The MAR becomes, then, the key metric for reporting impacts. During the design of the TAP evaluation project and prior to analyzing the results, it was assumed that the MAR would increase smoothly with time, though not linearly, and that it could be used in future years to predict the savings achieved by the program. The smooth curve assumption turned out to not be the case, as shown in Figure 2.

Figure 2: 2004 measure adoption rate (MAR) for electric energy efficiency measures

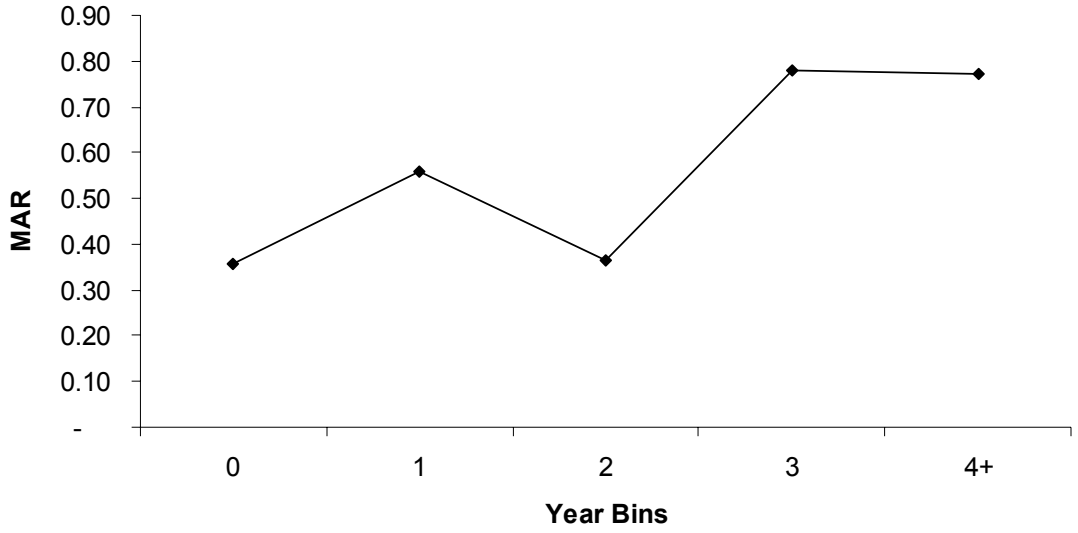


Figure 2 is built from the 2004 survey of 139 program participants. The horizontal axis represents the years since study completion; studies were binned by year. In 2004, year bin 1 would represent studies completed between December 31, 2001 and December 31, 2002. On average adoption does increase with time, but there is an anomaly in the curve; the rate decreases from 56% in year one to 36% in year two. If the curve had been applied in 2005, one year after the evaluation review, the savings recommended in 2002 would have moved from year bin 1 to year bin 2, with the absurd prediction that adoption had decreased from 56% to 36%. This illogical quirk not only called into question the proposed application of the MAR to report program impacts after 2004, doing so risked the possibility that a year with an abnormally large quantity of kWh savings for recommended measures could result in the program reported impacts decreasing as the abnormal group moved from year bin 1 to year bin 2. One of the conclusions of the 2004 evaluation team was that with repeated surveys the curve would smooth and reveal an underlying steady state adoption. The team recommended continuing the MAR sampling and survey work.

Additional research conducted in 2006

During the 2005 reporting cycle the TAP continued to use the MARs and SRRs developed in 2004 and further research in improving the realization rate was postponed while higher priority evaluation research tasks were carried out. In 2006 the evaluation team planned to update the MAR curve for MWh, and to attempt to collect enough data to develop year-bin MARs for OSG projects and non-electric savings[5]. The sample frame was all studies completed since December 31, 2004 up to the date of the sample draw in August of 2006. As in 2004, the study population was stratified into large and small contributors, with 80% of the sample drawn from the 20% of all reports that accounted for 80% of the recommended MWh savings. In all, 32 reports out of a population of 200 were randomly selected for review, with the sample size calculated to match the sampling fraction achieved in 2004.

In order to build on the 2004 survey record and use the earlier findings, the MARs for each 2004 sample element were included in the 2006 analysis. There were 2 classes of MARs from 2004 to be considered; those that had reported implementing 100% of all recommended measures, i.e. a report with a MAR of 1.0; and those that potentially may have implemented additional recommended measures during the 2005-2006 interim, i.e. a report with a MAR less than 1.0. Studies with a MAR of 1.0 retained the 1.0 MAR in 2006 while all others were resurveyed to see if additional measures had been implemented with a resulting increase in the MAR. Each study from the 2004 sample was represented twice in the 2006 analysis. Thus a 2002 study with a MAR of 1.0 contributed a MAR of 1.0 to both the year 2 and year 4 bin MARs. A 2002 study with a MAR less than 1.0, say 0.73, contributed 0.73 to year bin 2, and contributed a MAR of 0.73 or higher to year bin 4, depending if survey responses showed that additional recommended measures had been implemented.

Additional decisions that helped shape the 2006 evaluation plan were to use the 2004 SRR, and to draw the sample based on potential MWh savings. The evaluation team had good confidence in the accuracy of the 2004 SRR and this decision avoided the expense and time required for site inspections. Similarly, using MWh savings in the sample design was a pragmatic recognition of the need to focus limited resources on the most important reporting metric, despite the 2004 observation that non-electric impact reporting would likely be improved through targeted sampling.

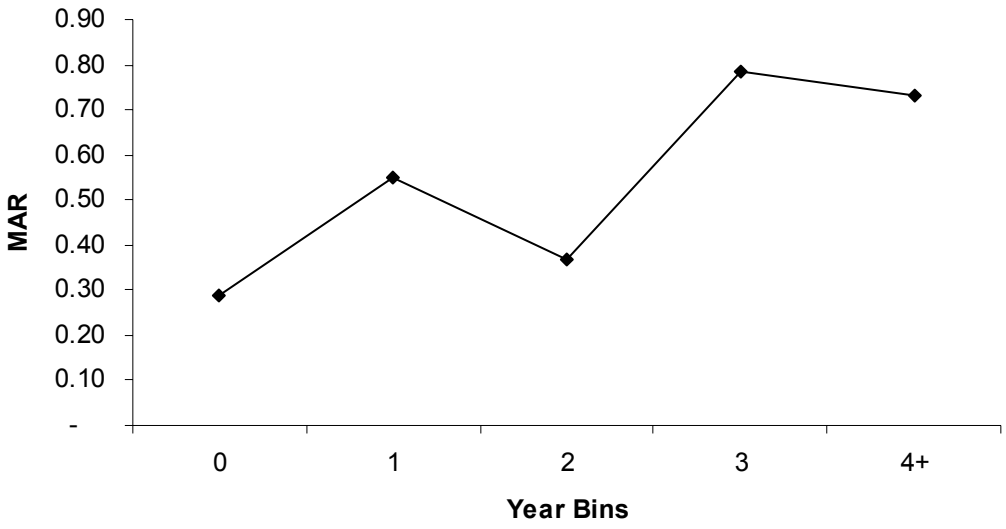
When completed, the 2006 MAR survey had contacted 104 program participants and, combined with the 2004 survey results, was able to report on the implementation status of 1,498 individual measures representing 25% of the TAP potential MWh savings. The average 2006 MAR for all program years was calculated as 0.59, compared to 0.54 calculated in 2004. The program's realized savings were reported as 652,312 MWh/year[6], 114 MW summer on-peak coincident demand reduction, and 2,308,100 MMBtu/year. As in 2004, the MWh impacts are considered reliable, while the MW reduction continues to be based on an MWh/MW factor, and the non-electric savings are based on a sample biased toward electric impacts.

The MAR curve based on the 2006 results continues to contain discontinuities as shown in Figure 3.

[5] The 2004 MARs for OSG and non-electric savings were not statistically valid at the year-bin level and were reported as single values covering all program years. This was largely due to the sample design which was not stratified for non-electric savings.

[6] The 2006 results based on a 2006 realization rate of 0.50 reported with ±0.08 precision at the 80% confidence interval.

Figure 3: 2006 measure adoption rate (MAR) for electric energy efficiency measures



The continuing existence of discontinuities in the MAR curve has implications for evaluation design. At the close of the 2004 evaluation study, reviewers felt that by increasing the sample size and adding observations to each of the 5 year bins that the curve would show an ever-increasing adoption rate with time. But the 2006 curve, representing 25% of the potential MWh savings of the TAP's studies, argues against the existence of a universal curve and probably reflects the actual state of all recommended measures. In positing the ever-increasing adoption rate theory, the evaluators most likely overlooked significant external factors that influence the decision to carry out a study's recommendations, including general economic conditions, energy prices, and energy price trends. In addition, because the TAP casts such a wide net and funds studies on a range of diverse energy management options, the studies themselves may focus on different project types in different years. From an evaluator's perspective, it is probably not necessary to make a final determination of the nature and impact of these external factors. Rather, the conclusion that there is no single MAR curve that represents the TAP's activities for any given program year suggests that the MAR should be continually updated, i.e. that surveys should be conducted on a periodic basis, preferably in conjunction with site inspections as done in 2004.

Conclusion

The TAP impact evaluation began with a comprehensive plan that addressed the unknown status of recommendations made in participant's studies, and had as a goal the development of a universal adoption rate curve of the recommendations. In carrying out the plan, the evaluation team has continually modified the original design and identified new research topics, which when completed will better represent the comprehensive nature of the program, and more accurately report its savings impacts, particularly for non-electric energy savings and demand reductions. Notwithstanding the need for improvement, the method has provided reliable estimates of electric energy impacts attributable to the TAP, the reporting framework has been incorporated into the program tracking database, and future

studies will likely continue to follow the same approach. Highlights of the approach include the following:

- Approximately 20% of the projects deliver 80% of the program impacts. The methodology focuses on these large contributors by weighting the sample in their favor in proportion to their contribution; i.e. 80% of the sample is drawn from the 20% of projects that make up the large stratum.
- Tracking projects by year bins results in a time-series of program realization rates that account for external variables such as market conditions or energy prices that fluctuate from year to year.
- Studies frequently present customers with a range of options, rather than clear go, no-go recommendations. As part of the TAP evaluation review, the evaluation team uses a rules-based approach to examine these ambiguous projects and then select measures that best match the "recommended" or "recommended for further study" classifications.
- Some combined heat and power studies are so large that their influence can dominate the savings analysis. Super large projects, those whose impacts are approximately 4% of recommended program savings, are treated as a unique stratum of one element.

References

Osei-Antwi, D., Gowans, D. 2005. *M&V Evaluation, Technical Assistance Program*. New York Energy Research and Development Authority.

Gogte, S., Gowans, D. 2007. *M&V Evaluation, Technical Assistance Program*. New York Energy Research and Development Authority.