To Do or Not to Do: Is It Time for Another Impact Evaluation?

Susan Haselhorst, ERS, North Andover, MA
Erik Mellen, Northeast Utilities, Westwood, MA
Chad Telarico, DNV KEMA, Syracuse, NY

ABSTRACT

Program administrators of energy efficiency programs and their regulatory consultants inevitably must decide whether to conduct another round of impact evaluations. Custom C&I program impact evaluations are particularly expensive, requiring on-site measurement and verification to confirm the applicant's savings estimates. The decision to repeat a study is influenced by available funds, program size, and the perceived stability of the program, and often just because an arbitrary period of time has passed since the last evaluation.

This paper presents a novel approach for developing objective criteria to aid in deciding whether to proceed with an expensive full-scale evaluation. The criteria consist of different measurements of the quality of the applicant savings estimates and subsequent program administrator (PA) engineering reviews, comparing past program activities (the benchmarks) to the present program on an application-by-application basis. The inference is that if the present program is measurably different from the benchmark, it is prudent to proceed with the full-scale impact evaluation. The incremental cost to complete an M&V impact assessment is about $10,000 per site, while a desk review of the same site is about an order of magnitude less expensive.

This approach applies to programs where savings are estimated using custom algorithms or site-specific parameters, and where the quality of analysis and the PA review can significantly contribute to the outcome of the results.

Introduction

In 2012, the MA Gas Working Group was faced with a dilemma: Should they move forward with a third consecutive impact evaluation with the hope of boosting the program realization rate, or postpone it to conserve resources, but potentially under-report savings?

The Working Group composed of the gas energy efficiency PAs of Columbia Gas, National Grid Gas, NSTAR Gas, Berkshire Gas, New England Gas, and Unitil, the evaluation contractors ERS and DNV KEMA, and the Massachusetts Energy Efficiency Advisory Council (EEAC) consultants was responsible for the direction and execution of evaluation of the natural gas CI programs. Programs were designed and evaluated jointly and statewide; although each PA individually administers the program, with unique processes for outreach, savings estimation oversight, and tracking. Over a 3-year period, the PA's gas programs had been transformed from a small-budget, moderate technical review model to a rapidly expanding program with more rigorously reviewed savings estimates. The program ambition had increased as well, expanding the portfolio to include a wide array of measures, such as high efficiency heating equipment, heating systems, heating controls, EMS, boiler combustion controls, building shell measures, and a variety of high efficiency gas industrial process equipment. As illustrated in Table 1, the programs had doubled in savings for 3 consecutive years.

Table 1. Massachusetts Gas Energy Efficiency Program Accomplishments

All Program Administrators	2008	2009	2010	2011
Number of participants	~200	339	335	369
Total tracking savings (therms)	~1,000,000	1,978,536	4,427,361	7,915,793
Total evaluated savings (therms)	No prior evaluations	1,410,696	2,985,423	?
Statewide realization rate	N/A	71%*	67%	?
Relative precision at 80% confidence	N/A	±11.1%	±9.0%	?
Sample size	N/A	43	48	?

* Controlled for outlier

The PAs had sponsored two prior evaluations for PY2009 and PY2010. The approach in both of the previous studies had been based on on-site M&V of a representative sample of participants. The realization rate each year was about 70% statewide. It was hypothesized that the low realization rates reflected the earlier implementation model and that as procedures became more rigorous the realization rate would increase to be on par with the electric programs, which are typically in the 90% range for gross savings.

The past evaluations had concluded that administrative errors and factors that could have been identified in a more rigorous technical review contributed to variances in realization rates. The PAs were taking steps to improve the technical review. However, since each PA independently administers a statewide common program, process improvements across PAs were not uniform. Some of the PAs were convinced that significant improvements to the process had been made, while other PAs concluded that their process improvements were barely underway. It was not clear whether the PY2011 projects of the third year, reflected enough improvement to warrant another impact evaluation, either statewide or for any particular PA.

New Approach

Rather than embark on a full impact evaluation or postpone an evaluation entirely, the Working Group tried a new approach. The group agreed to test, through a systematic review of a sample of PY2011 projects, whether the engineering estimation process had changed sufficiently to warrant one or more of the PAs proceeding to a full M&V impact evaluation. In concept, significant program changes, including changes to engineering methods, should be the primary trigger for an impact evaluation because a stable program should produce stable realization rates. This proposed method offered a way of testing a key element of program delivery – the measures savings estimation process.

In commissioning this task, the Working Group agreed first to a framework and then to one key ground rule. In the framework, a statistically selected sample of PY2011 sites would undergo desk reviews (the desk review sites) to characterize the current state of savings estimate quality. These results would be compared to similar reviews of sites that underwent M&V in the last two evaluations (the benchmark sites) to determine if there was a measurable improvement in the PY2011 methods.

The key ground rule was that objective criteria had to be determined prior to the completion and presentation of the PY2011 desk review results to avoid inadvertent tilting towards a preferred outcome. These were dubbed the decision criteria. It was also agreed that a decision whether to proceed to a full impact information would be made independently for each of the three PAs with the largest savings (PA1, PA2, and PA3 for the purposes of this paper) and statewide.

Methodology

The method for implementing the framework is outlined in Figure 1 and described in some detail in the subsequent five sections.

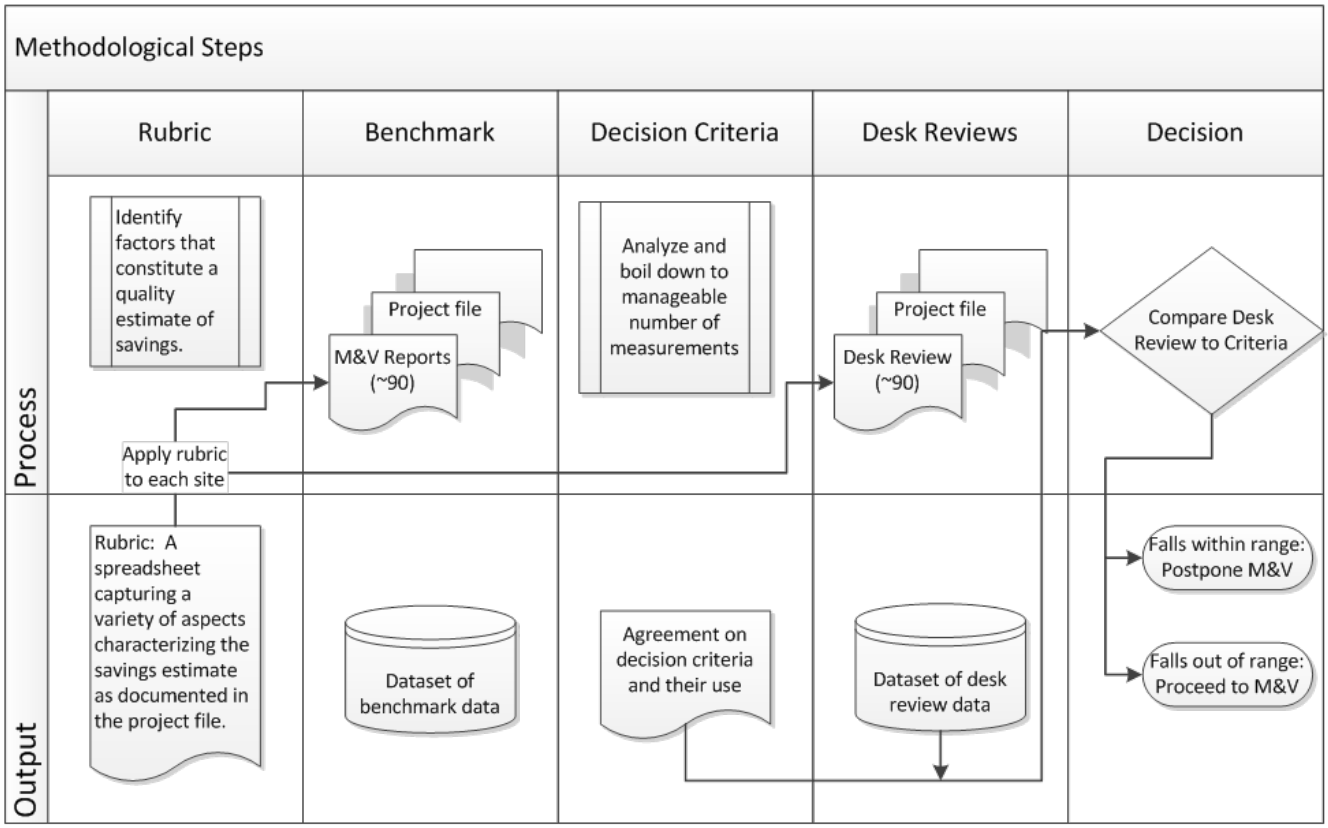

Figure 1. Framework Methodology

Step 1: The Rubric

As a first step, the impact team had to create a rubric for assessing the quality of a project savings estimate. This rubric had to capture the judgments made by an engineer during a review of applicant savings and had to be based on the material available to a reviewer prior to the installation of the measure and stored in the project file. For example, the results of the review could not rely on post-installed gas usage, as that information would not be available for an administrator reviewing an estimate as part of an application approval process.

The team focused on characterizing aspects of the project that could be reviewed from the project file alone and, when done properly, would lead to a better estimate of savings. The characteristics considered important could be summarized as follows:

- Was the baseline correct for the measure?
- Was an appropriate savings methodology employed?
- Was there evidence that customer billing had been consulted in reviewing the savings estimate?
- Was the savings fraction (savings as a percentage of total pre-installation gas usage) reasonable?
- Were all the documents present in the file (application, invoice, savings estimation description, native spreadsheets, or models)?
- Were the savings reproducible?
- What was the quality of the overall savings estimate?

These questions were translated to a spreadsheet designed to capture the reviewer's responses systematically and consistently from site to site. Table 2 represents the rubric showing some of the more important fields, although altogether there were about 70 parameters entered by the engineer for each site.

Table 2. Excerpt of the Spreadsheet-Based Rubric

Item	Notes
Customer and Measure ID	
Site ID	Evaluation-assigned site ID
Measure ID	Evaluation-assigned measure ID
Customer type	Examples: retail, office, K-12, hospital, etc.
Tracking measure category	Measure type: boiler replacement, EMS, etc.
Quantitative Savings Analysis	
Tracking savings, therms	Per tracking data
Evaluator desk review savings, therms	For benchmark: actual evaluated savings. For desk reviews, evaluator estimates.
Evaluator pre-install weather normalized billed gas use	Best available weather normalized billed use. Used for calculating savings fractions.
Applicant savings fraction	Tracking savings / pre-install gas use
Evaluator savings fraction	Evaluator savings / pre-install gas use
Documents Checklist	
Document list: application, offer letter, TA study, calculations, invoice, inspection report	The engineer checks off each document type found in the project file.
Tracking and Billing Review	
Status of tracking and billing data in the file.	Checklist inventory and quality assessment
Baseline Assessment	
Baseline specified by the applicant	Indicates whether a retrofit or code baseline was used. The engineer also specified whether the baseline choice was clear, implied, or not clear.
Baseline determined by the evaluator	Evaluator judgment whether a retrofit or code baseline is appropriate.
Assessment of Savings Estimation Methods and Quality	
Building simulation	The engineer indicates which method was best suited to the measure and which method was used by the applicant.
Proprietary method	
8760 or bin spreadsheet	
Factor driven, one-line calcs	
Estimation quality	Evaluator judgment of quality of estimate overall.

Pick lists, predefined descriptors embedded in the spreadsheet, were defined for qualitative parameters to permit comparison across projects. For example, the "overall quality of the estimate" reflects the reviewer's overall judgment about the estimation method documented in the project file. A higher quality savings estimate provides appropriate assumptions supported by site-specific information with transparent methods of calculation. Table 3 tabulates the reviewer's five pick-list choices in the rubric with illustrative examples.

Table 3. Evaluator Assessment of Quality as an Example of Pick-List Based Parameter

Reviewer Pick-List for Quality of Estimate	Example
Native files, reasonable, some field measurements, clear documentation	S361. This measure was an installation of combustion controls on a process boiler and included PHAST runs with related input and output files. Included extensive measurements of the boiler.
Evidence of good estimation, but no native files to verify	S248. The measure was a boiler installation in a new building. eQuest was used as the model. Project files included a TA report, which appeared to be reasonable and thorough; however, no native electronic eQuest files were included. The recent billing data appeared incomplete and did not corroborate the model.
Algorithm with some site-based information, but poor assumptions	S304. This measure was an installation of a 100% OA direct-fired unit. The estimate uses a one-line heat load calculated with actual building envelope data and schedules but doesn't adequately account for the outdoor air component.
Use a fixed savings fraction with no site-based data	S327. This is a combustion controls measure with savings estimated as a fraction of the facility-billed use without consideration of other potential gas uses on-site.
No calculations apparent	S281. This is a steam trap repair measure. The file includes information about other measures, but no steam trap inventory or other information, although a total count was provided.

Step 2: Creating the Benchmark Dataset

Once the rubric was designed, the engineering team went back to the site reports and project files selected for M&V sites from the two previous evaluations and applied the rubric to each site.

For the comparisons to be meaningful, it was important that the judgments of each engineer (a team of six) were similar across projects and year to year. Engineers underwent training on the intent and use of the tool and each of the final completed rubrics was reviewed by the same senior engineer to ensure consistency.

A total of ninety-one benchmark sites were reviewed using the rubric. The results from each of the spreadsheet templates were compiled into a single dataset. One of the more useful outcomes at this stage was a comparison of the applicant and evaluated savings fraction by measure. The savings fraction is the savings in therms divided by the total site billed use (in therms) prior to the implementation of the measure. These results are presented in Table 4 on a site-weighted basis. This table shows that PY2009–2010 applicants generally overestimated the fraction of the existing gas bill that would be saved by the measure – which corresponds in turn to the lower program realization rate. These fractions could be useful as a sanity check of applicant savings.

Table 4. Savings Fractions for Select Measures Using PY2009 and PY2010 Evaluated Results

Measure	Number of Measures	Average Evaluated Savings Fraction	Average Tracking Savings Fraction
Boiler/DHW replacement	37	6.9%	12.3%
EMS	13	8.1%	8.7%
Boiler burner/controls	11	2.5%	5.0%
Heat recovery	9	7.0%	11.4%
Insulation roof	8	18.5%	22.6%
Windows	6	1.3%	2.6%
Insulate walls/attic/ducts	5	10.7%	10.2%

Step 3: Defining the Decision Criteria

With the benchmark dataset in hand, the decision criteria had to be conceptualized and quantified in such a way that that the pass/fail test would be unambiguous once the desk review sites' results were in. As noted previously, these criteria had to be established before the desk review step to ensure objectivity. In addition, all the PAs and the EEAC consultants had to agree to the Decision Criteria even though some of the PAs hoped for opposite outcomes.

There were multiple options for how to proceed. How many criteria should there be, and what should they be? Should they be based on a simple site count or weighted in a manner reflecting the site's impact on program outcomes? Should individual criteria be weighted or each counted the same? How should non-numeric parameters, such as the quality of estimate, be translated into an objective score? How should the margin between passing and failing be defined?

Ultimately, the Working Group agreed to seven decision criteria, as shown in Table 5. Each criterion was presented as a percentage of total program tracking therms meeting the criterion. Thus, the baseline criterion can be interpreted as indicating that the baseline was appropriate at benchmark sites representing 75% of the program therms. The criterion also specified the range of values (No Action Range) considered close enough to the benchmark to show process changes are insufficient to warrant an impact evaluation. Thus, if the baseline was appropriate for sites representing between 60% and 89% of the program therms, evaluation would not be warranted per the baseline criterion. Likewise, if the desk review sites had incorrect baselines representing greater than 89% or less than 60% of program therms, an impact evaluation was warranted.

Table 5. Statewide Decision Criteria Summary

Criterion	Benchmark Value	No Action Range	Weighting Factor
Baseline is appropriate. This criterion captures how often the applicant identified the correct baseline (retrofit or replacement at end of life). Inappropriate baselines were a major source of discrepancies in previous evaluations.	75% of the time	60%–89%	40%
Savings method was appropriate. This criterion captures how often the applicant used an appropriate savings calculation method. For example, often a vendor estimated savings as a fixed percentage of the gas bills, when a bin analysis was more appropriate. Some sites had no savings calculations.	47% of the time	38%–57%	10%
Savings fraction. This is the average program savings as a percentage of the average pre-installed bills. The savings fraction should be a stable indicator of actual measure savings and therefore is useful as an independent comparison of the applicant savings estimates from year to year.	8.2%	6.6%–9.8%	10%
Document inventory. This criterion represents the frequency of certain documents observed in the project files. This was intended to be an objective measure of administrative consistency.	44% of documents found	35%–53%	10%
Evidence of bills in the file. This criterion captures how often bills appeared in the project files since gas bills are so useful in estimating or benchmarking gas savings.	35% of the time	28%–42%	10%
Savings were reproducible. This criterion indicates how often there was sufficient information for the reviewing engineer to reproduce the applicant savings.	54% of the time	43%–65%	10%
Quality of the estimate. This is an overall assessment of the quality of the savings estimate. Table 3 specifies the five choices.	67% reasonable quality	54%–81%	10%
Threshold standard	20%		

To finalize the criteria, the Working Group had to finalize the range of values for each criterion where no M&V would be required (No Action Range).

The degree of change in the criterion value considered significant enough to warrant proceeding to the on-site work (the "threshold standard") was 20%. This threshold is somewhat arbitrary. Finding that gas billing is factored into the savings analysis 20% more of the time, for example, shows an improvement in the estimation process, but it does not follow that savings will increase 20%. That being said, a 20% change in a criterion is likely to be large enough to rise above the noise in the results, indicating that more systematic changes have occurred and yet not so large as to preclude the identification of any improvements. The Working Group also agreed to weight the individual criterion, as shown in Table 5 into a single score.

An attempt was made to develop an analytical model relating the individual benchmark criteria scores to the site realization rate using regression analysis. The model only weakly explained the realization rate. It is speculated that a better model would have to account for measure mix, project size, and other factors not directly related to the savings estimation process. However, the model did consistently show that the baseline was the most significant criterion; therefore, the baseline criterion was assigned the highest weighting.

A detailed example of how one criterion value was calculated follows.

Example of Baseline Criterion

Changes to the baseline reference from the preexisting equipment to building code (or the equivalent) accounted for about 5% of the 30% discrepancy in realization rate observed in the previous evaluations. A frequent source of baseline changes occurred when a failed large capital piece of equipment, such as a boiler, was claimed as the baseline, when code would have been more appropriate. Identifying the correct reference baseline (pre-existing vs. code) is a crucial decision the administrator must make in the review of the application that can have a large impact on the savings. The selection is often a technical decision and it must prevail against both customer and in-house pressures to claim more savings; therefore, it is an excellent indicator of the robustness of the review process.

Table 6 compares the applicant and evaluator identification of the baseline. In some cases, the applicant baseline was not documented at all or was ambiguous. The cases where the applicant baseline was not clear or different from the evaluator's are shaded red, while agreements are shaded green.

The "Savings in agreement" value of 74% in Table 6 is the portion of desk review estimates, by savings, where the evaluator and applicant baselines are in agreement. According to the table, about a quarter of the program savings were subject to a baseline adjustment in the benchmark studies. The disagreement is represented by the three red cells where the applicant baseline was not clear or where the applicant incorrectly indicated that the baseline was the preexisting conditions.

Table 6. Benchmark Result: Baseline Agreement

STATEWIDE Applicant Assessed	Evaluator Assessed Clearly code or equivalent	Clearly Preexisting Conditions
Clearly code or equivalent	343,047	
Apparently code or equivalent	711,221	
Not clear	423,921	
Apparently preexisting conditions	554,288	537,781
Clearly preexisting conditions	900,235	3,850,840
Savings in agreement (shaded green)	5,686,297	
Agreement savings	74%	
No action range	> 59%	< 89%

Step 4: Desk Reviews of Current Projects

Once the decision criteria had been defined and agreed upon, the engineering team commenced the desk reviews of a statistically selected sample of PY2011 projects applying the rubric to each. The sites were

selected using an on-site M&V sample strategy. If the results proved a site M&V impact evaluation was warranted, the engineering team could quickly and efficiently convert the desk reviews to a site M&V plan.

A total of eighty-five sites were reviewed using the rubric.

Step 5: Compare Desk Reviews to the Benchmark

The criteria values were calculated and compiled from the PY2011 desk reviews. The criteria scores are presented in Table 7 for the state as a whole and also for the three largest PAs. Color coding is used to show where a criterion was out of the No Action Range (coded red) indicating that the savings estimation process had improved or regressed and an M&V impact evaluation was warranted. Criteria that remained within range are color-coded green. Results are presented for the state and also for the three largest PAs. As noted previously, the Working Group had agreed that the results would be examined statewide and by each of the three largest PAs.

Table 7. Desk Review Results Compared to Decision Criteria

Benchmark	Statewide Benchmark Value	State	PA1	PA2	PA3
Baseline is appropriate.	75% of the time	79%	74%	78%	87%
Savings method was appropriate	47% of the time	61%	85%	47%	72%
Savings fraction.	8.2%	6.8%	6.8%	6.7%	7.6%
Document inventory.	44% of docs found	42%	47%	43%	48%
Evidence of bills in the file.	35% of the time	45%	71%	38%	42%
Savings were reproducible.	54% of the time	39%	27%	47%	72
Quality of the estimate.	67% reasonable quality	71%	80%	65%	78%

These findings indicate that a significant change in practice is not indicated broadly enough to warrant another statewide impact evaluation (only three of seven criteria are out of range). However, when the results are examined by PA, a different conclusion is reached for PA3. Both PA1 and PA3 did stray outside of the range more often than not. PA1 showed both improvements in three categories and an erosion in reproducibility. However, when the criteria are considered on a weighted basis, they indicate that only PA3 showed sufficient change to warrant another impact evaluation with all criteria indicating the same trend towards improvement.

These conclusions are aligned with the PA reports of process changes. PA3 reported that a significant and definitive change occurred in the late 2010 timeframe. Prior to the change, the gas program manager conducted the savings estimate review; after that date, staff engineers were assigned responsibility to review custom estimates of savings. The other PAs did not identify any such sharp change in practice.

Based on the evidence of this process and the confirmatory information from the PAs, the Working Group decided to proceed with an impact evaluation of PA3's program only.

Results of the Impact Evaluation

The evaluation for PA3 proceeded with on-site M&V of a sample of sixteen sites. The final evaluated PA3 realization rate showed substantial improvement over previous evaluations. The PA3 realization rate

history is summarized in Table 8. Clearly, the realization rates have improved from the PY2010 evaluation. While the PY2009 evaluation showed a high realization rate, the results were much less precise, beyond the effect of the small sample size.

Table 8. PA3 Realization Rate Trend

Program Year	Realization Rate	Relative Precision	Sample Size
PY2009	84.9%	±29.2%	7
PY2010	47.3%	±11.2%	13
PY2011	84.4%	±6.9%	16

Figure 3 compares the projected desk review and evaluated savings against tracking savings. Both desk review and evaluated savings are well correlated with tracking savings, although the desk review savings are biased downwards with an unweighted realization rate of about 69%, whereas the unweighted evaluated realization rate is 80%. While desk review estimates of savings are not available for the PA3 PY2009 or PY2010 studies, we suspect the reviews would have shown a low realization rate with a high rate of variance because the paperwork was so scant in those years.

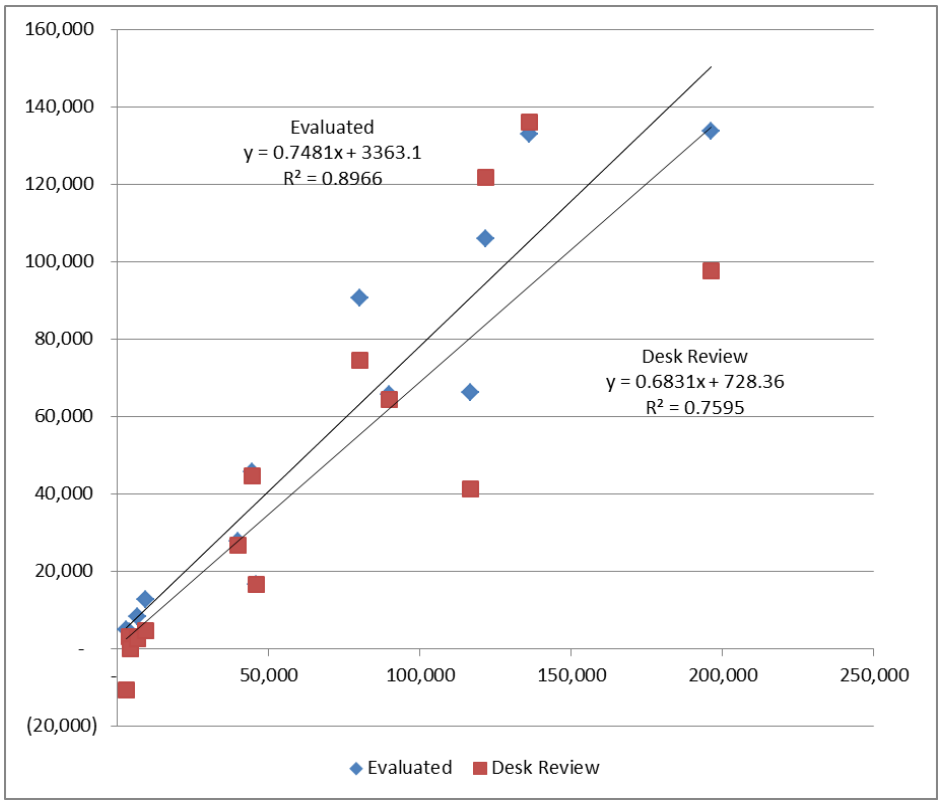

Figure 3. Desk Review and Evaluated Savings vs. Tracking Savings

In most cases, the desk review projected and evaluated realization rates were reasonably close. In a few cases, where both saving estimates relied on billing data, the results were the same. The divergences occurred for the most part as follows:

- Smaller sites, which tended to have lower quality estimates with less site-specific information incorporated into the applicant analysis.

- Sites where the billing data was incomplete, leading the desk reviewer to incorrect conclusions.
- The most divergent case occurred where the desk review had concluded the technology had been incorrectly applied, producing negative savings, where the on-site M&V concluded the technology was correct. Multiple gas meters at the site also confounded the desk review conclusions.

These results are not surprising, given that less effort is expended on smaller projects, and so the estimates tend to be less site specific. Also, billing analysis can be an excellent method for verifying savings, but it is important to have all the affected bills and also confirmation that other factors such as occupancy or production are not affecting the outcome.

Figure 4 compares the evaluated realization rates to a desk review quality index listed in Table 5, where a 5 rating is the highest quality estimate and where a rating of 1 indicates that no savings methods were documented in the files. In the graph, the shaded bar represents a ±20% band around the 85% realization rate line (the PA3 program evaluated realization rate). There were eleven sites with quality indices of 4 or better, seven of which fell within the band, and there were five low-quality sites, two of which fell within the band. While these are small numbers, the trend indicates that the sites judged to have higher quality estimates produced evaluated realization rates that were better aligned to the final program realization rate.

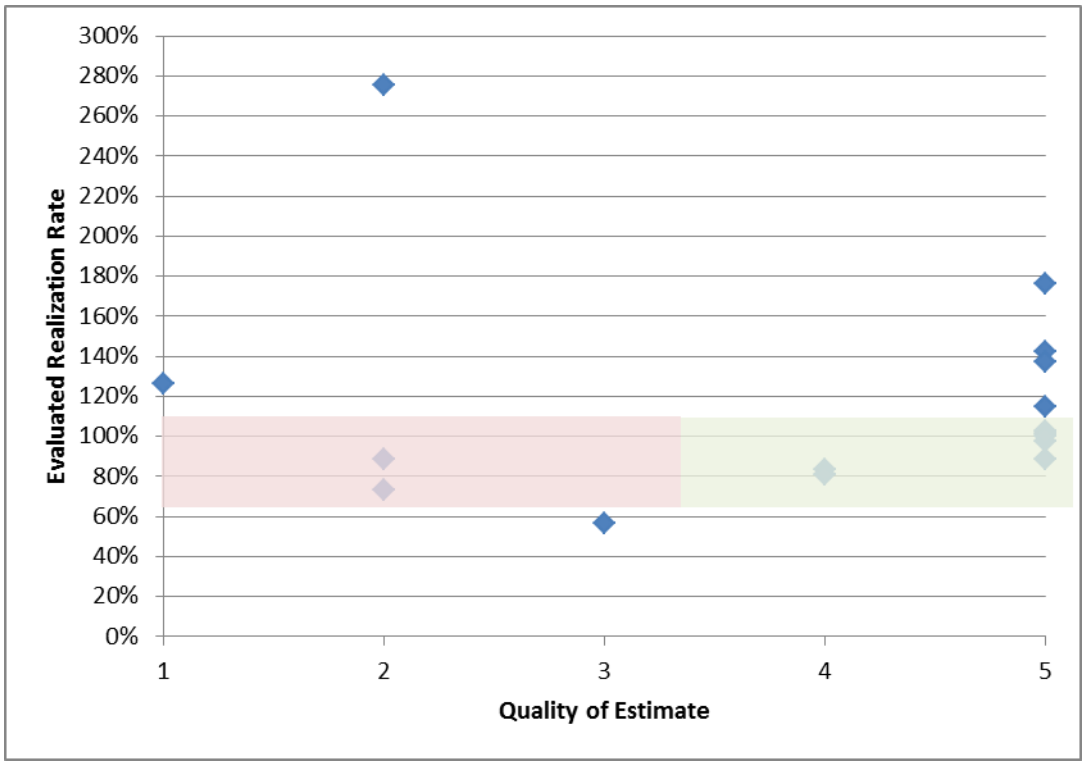

Figure 4. Quality Index vs. Realization Rate

The correct assessment of baseline was considered the most heavily weighted of the criteria. There was one site where the evaluated baseline differed from the applicant baseline. The desk review had also concluded the applicant baseline was incorrect.

Conclusions and Recommendations

The intent of comparing savings estimates against benchmarks is to provide a reproducible and systematic method for measuring the quality of an energy savings estimate in a custom program. This

process could be useful for both implementers and evaluators. While it is not highlighted in this paper, the method can provide implementers an actionable indicator of where savings estimates are weak. Also, the savings fractions in Table 4 provide a mechanism for flagging applicant estimates that may warrant a closer look during the application review.

The largest potential benefit, however, is for evaluators. Current practice for deciding when to conduct another round of evaluation often relies on subjective judgment that "it's been long enough" or on a framework that requires regular evaluations, whether they are needed or not. Evaluating a custom program is expensive and a more objective basis for timing evaluations could free up resources for other activities or trigger a necessary evaluation earlier than might have been otherwise. The benchmarking can also provide a basis for charting long-term progress or for program-to-program comparison of best practices.

While the method shows promise, it cannot be considered validated with PA3's single data point. The PA3 evaluation results did corroborate the findings of the benchmarking showing significant improvements in the program realization rates, confirmation of baseline findings, and alignment with findings about quality. However, the realization rates of the other PAs were not evaluated; therefore, it is unknown if their realization rates would have remained stable. Furthermore, while it stands to reason that better estimating processes lead to better realization rates, the method of estimating savings is not deterministic. For example, a robust savings estimate based on detailed models and pre-installation metering will be impacted by an unexpected addition of a shift subsequent to the measure installation. Likewise, a 'guess-timate" is sometimes correct.

The resources required to implement the method includes defining a rubric, compiling site reviews to form the benchmark, and regular desk reviews of new program application folders to determine if changes against the benchmark have occurred. The site-by-site input of the rubric as part of an on-site evaluation is trivial if it is incorporated at the outset of an evaluation. Development of the rubric and compilation of the results is an additional, but relatively small, increment to an overall evaluation. Compiling benchmarking data from past applications or evaluation on-sites is more substantial scope, requiring in the order of about a day per site to review, quality control, and compile the data. Finally, implementation requires organizational commitment to sustain benchmarking year to year.

The method was useful to the Working Group in deciding on how to proceed, where there had been fundamental disagreement before the analysis was presented and led to an outcome. The Working Group has informally discussed that should a full impact evaluation commence next year, it may be worthwhile to complete the desk review step with the rubric, building another data point for the method. Potential improvements that might be tested at that time include refinements to the rubric and also incorporating calls to customers to identify any major operational changes that might impact savings.