energy & resource solutions

120 Water St., Suite 350
North Andover, MA 01845
Phone: (978) 521-2550
Fax: (978) 521-4588
Web: www.ers-inc.com

# MEMO

**DATE:** **NOVEMBER 27, 2012**

**TO:** **CUSTOM GAS EVALUATION TEAM**

**FROM:** **HASELHORST**

**RE:** **REVISION: DECISION CRITERIA FOR PROCEEDING TO M&V**

In response to the discussion of November 7, the Decision Criteria have been revised as follows:

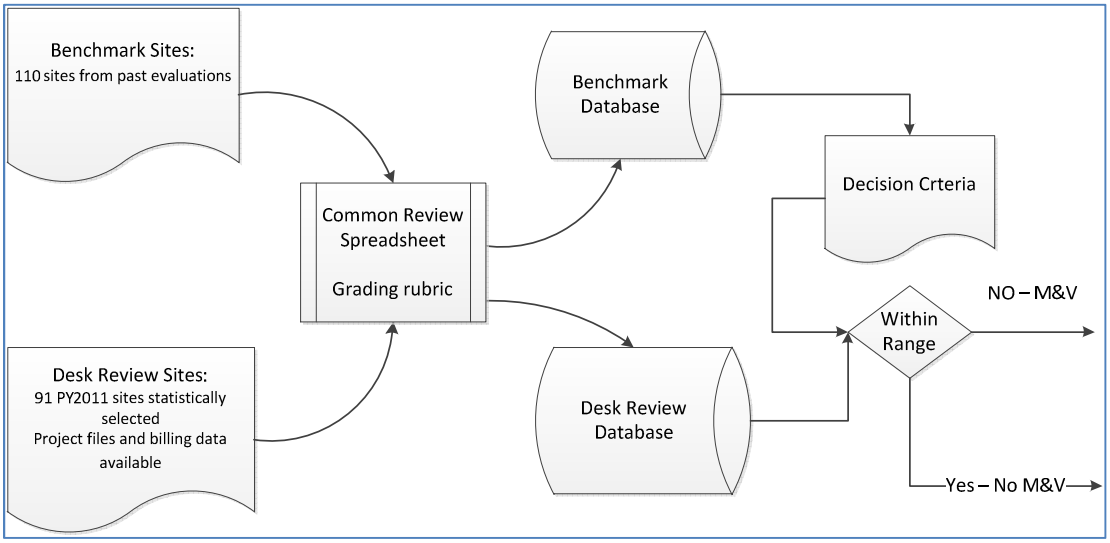
- Revised figures include the results of all sites from the 2009 and 2010 impact evaluations.
- The desk review ratio has been dropped as a criterion because it is not an apples-to-apples comparison with the evaluated realization rate and may be misleading.
- We propose adding the hypothesis testing to determine whether the benchmark populations and the PY2011 are statistically different.

## BACKGROUND

The primary purpose of the desk review is to assist the Evaluation Team (the PAs, the EEAC consultants, and the KEMA team) in deciding whether to proceed with site M&V for the PY2011 Custom Gas Program or not. Because of the high cost of on-site M&V, it is only prudent to proceed if there is evidence of improved tracking savings estimation methods in the PY2011 population.

A statistically selected sample of PY2011 sites will undergo desk reviews to characterize the current state of savings estimate quality. These results will be compared to similar reviews of sites that underwent M&V in the last two evaluations (the benchmark sites) to determine if there is a measurable improvement in the PY2011 methods. The Evaluation Team has agreed that this comparison must include objective criteria to aid in the decision making ("Decision Criteria"). It was also agreed that the Decision Criteria should be determined prior to the completion and presentation of the PY2011 desk review results to avoid inadvertent tilting towards a preferred outcome. This process is illustrated in the following figure.

**Figure 1. Decision Making Process**




The memo presents the benchmark review findings and a straw man proposal for Decision Criteria indicative of changes in estimate quality. A final section in the report describes the process for determining benchmark results.

## BENCHMARK RESULTS AND DECISION CRITERIA

This section presents the revised criteria for determining whether the desk review results show sufficient improvements in savings estimation methods to warrant further site work. Each decision criterion was derived from a benchmark result. If the desk review results are sufficiently close to the benchmark result, they are within the "No Action Range", indicating that the savings methods have not changed sufficiently to warrant further M&V. However, if the desk review results fall outside of the No Action Range, it implies significant enough changes to savings estimation methods to warrant moving to the M&V impact evaluation.

In this revised memo, we recommend also applying a hypothesis test to determine if the benchmark and the PY2011 desk review populations are significantly different. The decision to move forward has two parts: first the PY2011 desk review population value must be statistically different from the benchmark population using a hypothesis testing; secondly, the PY2011 value must fall outside the 'No Action' range. The exact form of the hypothesis testing is not known at this time, but the team plans to test multiple methods.

An attempt was made to develop an analytical model relating the criteria to the realization rate using regression analysis. The model we developed only weakly explained the realization rate. We speculate 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. We assigned the Baseline criterion the highest weighting and then assigned equal weights to the remaining criteria.

Table 1 summarizes the proposed Decision Criteria with proposed weighting factors.

**Table 1. Statewide Decision Criteria Summary**

| Benchmark | Benchmark Value | No Action Range | Weighting Factor |
|---|---|---|---|
| Baseline is appropriate | 74% of the time | 59% - 89% | 40% |
| Savings method was appropriate | 47% of the time | 38% - 57% | 10% |
| Savings fraction | 8.8% | 7.3% - 9.3% | 10% |
| Document inventory | 44% of documents found | 35% - 53% | 10% |
| Evidence of bills in the file | 24% in agreement | 19% - 29% | 10% |
| Savings was reproducible | 54% of the time | 43% - 65% | 10% |
| Quality of the estimate | 67% reasonable quality | 54%-81% | 10% |
| Threshold standard | 20% | | |

The Decision Criteria are presented in the detail below by each of the three largest PAs and statewide, consistent with realization rate reporting protocols.

## Threshold Standard and Weighting

The Decision Criteria must identify the degree of change considered significant enough to warrant proceeding to the on-site work (the "threshold standard"). This memo proposes a ±20% threshold standard. Table 2 presents annual tracking savings and realization rates for the 2009-2011 custom gas programs. Using the threshold standard of 20%, one may compute a range in savings and realization rates which would theoretically fall within the No Action Range and indicate further impact evaluation work of PY2011 was not warranted.

**Table 2. Threshold Standard Proposal of 20%**

| All Program Administrators | 2009 | 2010 | 2011 |
|---|---|---|---|
| Total tracking savings (therms) | 1,978,536 | 4,427,361 | 7,915,793 |
| Total measured savings (therms) | 1,736,322 | 2,858,553 | 5,834,679 |
| Realization rate (no outlier) | 87.76% | 64.57% | 73.71% |
| **Proposed change threshold** | **20%** | Lower bound | Upper bound |
| Realization rate range, no action | | > 59% | < 88% |
| Savings range, no action | | > 4,667,744 | < 7,001,615 |

*2011 RR% is the weighted average of 2009 and 2010

It is possible that the evaluation team would consider proceeding with the M&V scope with the prospect of half that level change. However, it is not clear how the criteria discussed below will translate to increases or decreases in the realization rate. 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% threshold is likely to be large enough to rise above the noise in the results indicating more systematic changes have occurred and yet not so large as to preclude the identification of any improvements.

The weighting factors reflect the relative value of the criteria.

### Baseline is Appropriate

Changes to baselines by the evaluator accounted for about 5% of the 30% discrepancy in realization rate observed in the previous evaluations. A frequent source of the baseline change occurs when the applicant installs a large capital piece of equipment, such as a boiler, where code is the likely baseline.

Table 3 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 is not clear or indicates that their baseline is different from the evaluator are in the Red Zone, while agreement is shown in the Green Zone. An improvement in agreement between the evaluator and applicant will improve the realization rate.

We propose computing the portion of desk review estimates that fall into the Green Zone on an aggregate basis. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V.

**Table 3. Benchmark Result: Baseline Agreement**

| STATEWIDE | Evaluator Assessed | |
|---|---|---|
| Applicant Assessed | Clearly code or equivalent | Clearly pre-existing conditions |
| Clearly code or equivalent | 343,047 | |
| Apparently code or equivalent | 711,221 | |
| Not clear | 423,921 | |
| Apparently pre-existing conditions | 554,288 | 537,781 |
| Clearly pre-existing conditions | 900,235 | 3,850,840 |
| | | |
| Savings in Green Zone: | 5,686,297 | |
| Percent in Green Zone | 74% | |
| No Action Range: | > 59% | < 89% |

| Columbia | Evaluator Assessed | |
|---|---|---|
| Applicant Assessed | Clearly code or equivalent | Clearly pre-existing conditions |
| Clearly code or equivalent | 9,070 | - |
| Apparently code or equivalent | 251,271 | - |
| Not clear | 287,640 | - |
| Apparently pre-existing conditions | 3,763 | 213,860 |
| Clearly pre-existing conditions | 361,292 | 1,452,106 |
| | | |
| Savings in Green Zone: | | |

| | | |
|---|---|---|
| Percent in Green Zone | 73% | |
| No Action Range: | > 59% | < 88% |

| National Grid<br>Applicant Assessed | Evaluator Assessed<br>Clearly code or equivalent | Clearly pre-existing conditions |
|---|---|---|
| Clearly code or equivalent | 248,949 | - |
| Apparently code or equivalent | 164,868 | - |
| Not clear | 277,549 | - |
| Apparently pre-existing conditions | 159,163 | 188,619 |
| Clearly pre-existing conditions | 381,278 | 1,863,649 |
| | | |
| Savings in Green Zone: | | |
| Percent in Green Zone | 74% | |
| No Action Range: | > 59% | < 89% |

| NSTAR<br>Applicant Assessed | Evaluator Assessed<br>Clearly code or equivalent | Clearly pre-existing conditions |
|---|---|---|
| Clearly code or equivalent | 85,028 | - |
| Apparently code or equivalent | 295,082 | - |
| Not clear | 422,516 | 103,114 |
| Apparently pre-existing conditions | 141,096 | 459,460 |
| Clearly pre-existing conditions | - | - |
| | | |
| Savings in Green Zone: | | |
| Percent in Green Zone | 57% | |
| No Action Range: | > 46% | < 69% |

## Quality of the Savings Estimate

We propose a Decision Criterion which compares the PY2011 desk review sites to the benchmark sites in the quality of the savings estimates. We have characterized that quality of the savings in three assessments:

- Quality of the overall estimate
- Appropriateness of the algorithm employed
- Reproducibility of the savings

Table 4 tabulates the reviewer's judgment of the quality of the engineering estimate. The engineer assigns the site estimate one of five grades, as indicated in Table 4. The values in the table are in therms and represent the portion of savings at the quality level indicated.

For criterion of this type, we have assigned responses as falling in the Green Zone – the quality is acceptable, or in the Red Zone - where it is not. The benchmarked sites show that 65% of the

savings estimates were of at least reasonable quality and included some element of site based information and a defensible algorithm.

We propose computing the portion of desk review estimates that fall into the Green Zone on an aggregate basis. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V. Likewise, if the aggregate desk review result falls outside of the No Action Range, it is indicative that the process has changed and site M&V may be warranted.

**Table 4. Benchmark Results: Overall Quality of the Estimate**

| Quality of Estimate | Columbia | NGRID | NSTAR | Statewide |
|---|---|---|---|---|
| Native files, reasonable, some field measurements, clear documentation | 336,700 | 536,638 | 0 | 873,338 |
| Evidence of good estimation, but no native files to verify | 364,697 | 326,553 | 299,619 | 990,869 |
| Algorithm with some site based information, but poor assumptions | 1,389,695 | 1,349,480 | 242,181 | 3,061,071 |
| Use a fixed savings fraction with no site based data | 219,623 | 334,401 | 267,702 | 927,313 |
| No calculations apparent | 229,952 | 634,070 | 556,554 | 1,468,742 |
| | | | | |
| Savings in the Green Zone | 82% | 70% | 40% | 67% |
| Desk review result is GREATER than | > 66% | > 56% | > 32% | > 54% |
| Desk review result is LESS than | < 99% | < 83% | < 48% | < 81% |

The next table, Table 5, summarizes methods used to estimate savings and whether the method employed was appropriate for that measure. We propose computing the portion of desk review estimates utilize the most appropriate method. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V.

**Table 5. Benchmark Results: Applicability of the Estimation Method**

| Method Used is Most Applicable | Columbia | NGRID | NSTAR | Grand Total |
|---|---|---|---|---|
| Building simulation | 238,290 | 296,209 | 77,835 | 612,333 |
| Proprietary method | 243,400 | 0 | 0 | 243,400 |
| 8760 or bin spreadsheet | 771,858 | 771,699 | 184,550 | 1,728,106 |
| Factor driven, one-line calcs | 165,704 | 660,213 | 43,365 | 892,682 |
| No calculations in the file | 229,952 | 634,070 | 556,554 | 1,468,742 |
| Total savings | 2,540,667 | 3,181,141 | 1,366,056 | 7,321,332 |
| | | | | |
| Sites in the Green Zone | 56% | 54% | 22% | 47% |
| Desk review result is GREATER than | > 45% | > 43% | > 18% | > 38% |
| Desk review result is LESS than | < 67% | < 65% | < 27% | < 57% |

Table 6 summarizes how often the engineer could reproduce the applicant's energy savings results. We propose computing the portion of desk review estimates that fall into the Green Zone

on an aggregate basis. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V.

**Table 6. Reproducibility of the Applicant Savings**

| Savings is Reproducible | Columbia | NGRID | NSTAR | Statewide |
|---|---|---|---|---|
| Yes | 1,505,426 | 1,729,077 | 624,931 | 3,965,021 |
| Partial | | | | |
| No | 1,035,241 | 1,451,344 | 741,125 | 3,355,591 |
| | | | | |
| Sites in the Green Zone | 59% | 54% | 46% | 54% |
| Desk review result is GREATER than | > 47% | > 43% | > 37% | > 43% |
| Desk review result is LESS than | < 71% | < 65% | < 55% | < 65% |

## Savings Fraction of the Billed Usage

The savings fraction is the ratio of the savings to the pre-installed weather normalized bills and provides another method for comparing evaluated and tracked savings. For some measures, the savings fraction is expected to be similar site to site and can therefore provide a more consistent benchmark for comparison of estimates from year to year. The tracking and evaluated savings fractions have an identical denominator, which is the pre-installed and weather normalized billed usage or the best estimate of billed usage available.

Table 7 compares the savings fractions by PA. The tracking savings fraction is larger than the evaluated fraction, which is to be expected given the realization rate which indicates a downward direction in savings. Sites were excluded if they did not have reasonable billing data. We propose computing the desk review aggregate savings fraction. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V.

**Table 7. Benchmark Results: Savings Fraction by PA**

| Savings Fraction | COLUMBIA | NGRID | NSTAR | Statewide |
|---|---|---|---|---|
| Number of measures | 45 | 43 | 27 | 123 |
| Tracking Savings | 2,540,667 | 3,181,141 | 1,366,056 | 7,321,332 |
| Evaluated Savings | 2,036,712 | 2,338,379 | 688,222 | 5,146,045 |
| Pre-installed Billing | 27,041,283 | 43,443,292 | 11,197,126 | 83,286,838 |
| Tracking Savings Fraction | 9.4% | 7.3% | 12.2% | 8.8% |
| Evaluator Savings Fraction | 7.5% | 5.4% | 6.1% | 6.2% |
| of the Difference | 1.9% | 1.9% | 6.1% | 2.6% |
| Savings Fraction NO ACTION | | | | |
| Desk review savings fraction GREATER than | > 9.0% | > 6.9% | > 11.0% | > 8.3% |
| Desk review savings fraction LESS than | < 9.8% | < 7.7% | < 13.4% | < 9.3% |

## Billed Usage Factored in Savings Estimates

It is relatively easy to account for the few end-uses on the gas bill and to weather normalize weather dependent bills. For these reasons, most estimates should include an examination of the gas bills to at least sanity check the estimates.

This benchmark identifies how often gas billing appears to be factored into the applicant savings estimates, as shown in Table 8. We propose computing the portion of desk review estimates that fall into the Green Zone on an aggregate basis. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V.

**Table 8. Benchmark Result: Evidence of Bills**

| Project files with billing data | COLUMBIA | NGRID | NSTAR | Statewide |
|---|---|---|---|---|
| Referenced billing usage is reasonable | 677,694 | 443,898 | 1,028 | 1,235,968 |
| Applicant did not account for other end uses | 161,400 | - | - | 241,115 |
| Appear to be missing accounts or account mismatch | - | 159,163 | 110,678 | 286,408 |
| No reference billed use | 1,701,573 | 2,578,081 | 1,164,234 | 5,444,325 |
| | | | | |
| Savings with billing usage | 33% | 19% | 8% | 24% |
| Desk review savings fraction GREATER than | > 26% | > 15% | > 7% | > 19% |
| Desk review savings fraction LESS than | < 40% | < 23% | < 10% | < 29% |

## Document Inventory

The applicant file contains copies of administrative documents, such as the application and offer letters and documents that support the savings, such as technical assistance studies. Table 9 tabulates the frequency with which typical documents appear in the project file delivered to the evaluation team.

**Table 9. Benchmark Result: Document Inventory**

| Document | Columbia | NGRID | NSTAR | Statewide |
|---|---|---|---|---|
| Application | 17 | 36 | 6 | 59 |
| Technical assistance study | 17 | 15 | 7 | 40 |
| Customer offer letter | 42 | 37 | 24 | 103 |
| Cut sheets | 14 | 7 | 20 | 41 |
| Invoice | 25 | 30 | 15 | 70 |
| Post inspection | 2 | 13 | 0 | 15 |
| TOTAL | 117 | 138 | 72 | 327 |

| Savings Fraction NO ACTION | | | | |
|---|---|---|---|---|
| Benchmark fraction | 43% | 53% | 53% | 44% |
| Document fraction GREATER than | > 35% | > 43% | > 43% | > 35% |

| | | | | |
|---|---|---|---|---|
| Document fraction LESS than | > 52% | > 64% | > 64% | > 53% |

We propose computing the portion documents provided. If that value falls within the No Action Range, it is indicative that the process has not likely changed enough to warrant site M&V. This is a weaker indicator of savings estimation quality because not all PAs typically include an application nor are technical assistance studies part of the review process.

### Weighting Findings and Proposed Statistical Testing

The results in therms, including tracking savings, evaluated savings, and weather normalized billing (used in the computation of savings fractions) are aggregated as the product of the site value in therms and the case weight of the site. Results in therms (whether by PA, program year, or measure) therefore represent appropriate values for the population presented.

## BENCHMARKING PROCESS

The impact team reviewed the applicant and evaluator documents from the 110 sites from impact evaluations conducted in the last two years. The reviewer examined the applicant method of estimating savings, inventoried documents (applications, offer letters, cut sheets, and the like), and examined gas billing data. The findings were systematically cataloged using standard lists of terms in a standard reporting template, one spreadsheet completed for each site.

Each site's key findings were extracted and compiled in a database. Since standard terms, calculations, and grades were used in the review, the results could be tabulated across all the sites.

### Source of Benchmark Sites

The 110 sites reviewed for the benchmarking were the sites selected for M&V activities in four previous evaluations. The site distribution by PA is summarized in Table 1.

**Table 1. Distribution of Benchmark Sites by PA**

| Distribution of Sites | Berkshire | Columbia | National Grid | New England Gas | NSTAR | Unitil | TOTAL |
|---|---|---|---|---|---|---|---|
| 2009-2010 Tracking Savings | 99,794 | 2,254,482 | 2,790,072 | 23,400 | 1,126,737 | 111,412 | 6,405,897 |
| 2011 Tracking Savings | 98,618 | 1,442,479 | 4,280,343 | 95,831 | 1,983,274 | 62,913 | 7,963,458 |
| M&V Sites from 2009-2010 | 2 | 32 | 32 | 1 | 20 | 2 | 89 |
| Desk Review Sites* | 3 | 23 | 35 | 2 | 24 | 4 | 91 |

### Review Process

For benchmarking purposes, the reviewer typically reviewed the final site report to find many of the key benchmarking characteristics. These characteristics include:

- Results of billing analysis and also pre-installed weather normalized bills
- Analysis of applicant baseline and estimation methodology and how appropriate it was

The factors, once identified, are entered in the benchmark template using therms, percentages, or descriptors selected from a pick-list to maintain consistency site to site. A copy of unpopulated

benchmark template accompanies this memo. Virtually the same template was used for the desk reviews of the PY2011 sites to facilite comparison of the desk reviews with the benchmarks.