On-site Measurement and Verification versus Project File Desk Review

Michael Frischmann, Michaels Energy, La Crosse WI
Ryan Kroll, Michaels Energy, La Crosse WI

Abstract

An important segment of Demand Side Management (DSM) program evaluation is the determination of gross program impacts. Gross impacts are generally evaluated using project file desk reviews, on-site measurement and verification, or a combination of the two. The question that arises during evaluation planning is often how many of each type are to be completed? Do utilizing on-site impact evaluation activities produce results that are significantly different and meaningful as opposed to a file review? Why are on-sites necessary if we have an approved technical reference manual (TRM)?

The answers to these questions lay in the data. Comparing the gross impact evaluation results after a desk review to the results after on-site measurement and verification has taken place yields meaningful quantitative and qualitative differences. On-site data is not only useful for retrospective evaluations, but can also be extremely helpful for forward looking program improvement.

This paper will demonstrate the quantitative and qualitative differences between file reviews and on-site evaluations. Evaluation results data from both custom and prescriptive programs will be used to determine if the different evaluation methods produce different sample wide gross impact results. Additionally, we will discuss the gross impact results of programs that utilize third party vetted technical reference manuals.

Introduction

When considering program evaluations, there are a plethora of evaluation methodologies to consider. Over the last several years there have been discussions regarding how cost effective evaluation activities are with respect to the significance of the data they produce. The issue with evaluation activities is that as the complexity and depth of information collected increases, so does the cost. It is no surprise that performing a desk review of project documentation will cost less than conducting an on-site visit with measurement and verification.

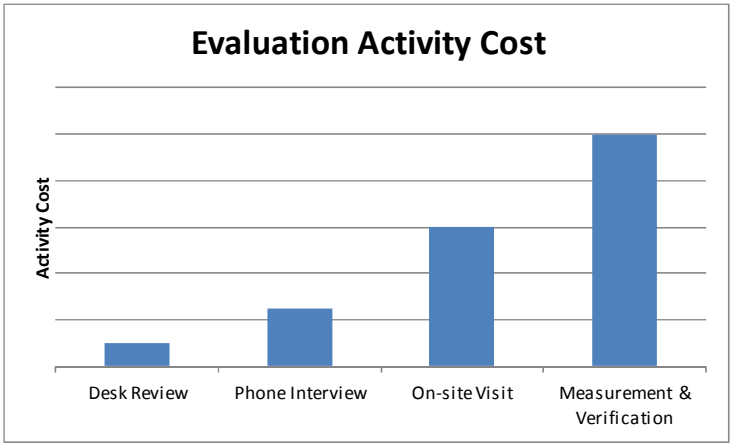

Figure 1. Evaluation Activity Cost

Determining which evaluation method is most appropriate for any given program depends on a large number of factors such as overall budget, program size, total participants, and overall resource savings. There are also a wide range of evaluation activities that can fit any evaluation budget:

- A desk review of project documentation to verify that the information on quantities, or types of equipment are consistent with what was outlined in the utility procedures document (such as a TRM).
- Customer phone interviews to determine if the equipment is installed and operational, as well as some key operation parameters such as hours of operation or temperature set points.
- Field work that focuses on installed measure quantities.
- Field work that includes measurement and/or extended metering of parameters that have a high amount of uncertainty.
- Field work that involves measurement and verification of all project parameters regardless of uncertainty level.
- There are also other forms of evaluation activities such as billing analysis or phone interviews that are not discussed in this paper.

The trouble for utilities is trying to determine which evaluation process is the most cost effective while still meeting the two key function of evaluation (TecMarket Works 2004):

1. To document and measure the effects of a program, and
2. To help understand why those effects occurred and identify ways to improve the program.

This paper attempts to quantify the difference in the results obtained by performing a desk review of project paperwork versus on-site measurement and verification which can include extended metering of key parameters for both custom and prescriptive programs. These are the two most common evaluation options considered and ones where data is readily available.

Data Set

The data used for analysis in this paper is a culmination of custom and prescriptive program evaluations completed over the last three years. These evaluations focused on electricity and demand savings, however, for the purposes of this paper, only the electricity usage (kWh) savings are analyzed. The data sets are all evaluation samples designed to achieve 90-10 precision for their respective programs.

Custom Programs. The custom program data is a combination of evaluations from a total of eight different evaluations that spanned four states and seven utilities. One of the utilities had three separate evaluations, each covering a distinct program year. Two of the evaluations were completed as part of statewide evaluations and included both multiple years and multiple utilities. The remaining evaluations included individual program years for individual utilities. The prescriptive program evaluations entailed desk reviews as well as measurement and verification of a combined 212 projects totaling 145.4 GWh of electricity savings. A summary of the programs evaluated can be seen in Table 1.

Table 1. Summary of Custom Program Data

	Number of Projects	Electricity Savings (kWh)
Utility A PY1	43	11,762,342
Utility A PY2	48	22,948,116
Utility A PY3	46	34,236,454
Utility B	21	2,627,647
Utility C	5	2,938,740
Utility D	5	12,422,190
Utility E	20	35,531,399
Utility F	22	22,930,140
Total	210	145,397,028

All program evaluations were conducted using the evaluation methodology outlined in the California Evaluation Framework (TecMarket Works 2004, 2006), or the International Performance Measurement and Verification Protocol (IPMVP) evaluation options A-D (EVO 2010).

Prescriptive Programs. The prescriptive program data is a combination of evaluations from three different utilities. Two of the utilities' data sets have two full evaluation cycles completed, and the third has only one. The prescriptive program evaluations entailed desk reviews as well as measurement and verification of a combined 643 measures totaling 94.16 kWh of electricity savings. The ex-ante savings for all of the prescriptive measures examined in the data sets were originally determined based on the values presented in either a state wide TRM, or a program specific procedures manual. The deemed savings documents used in these programs all provide a well defined and researched set of assumptions, have varying savings levels dependant on the end user's facility type, and were vetted by the programs evaluators. A summary of the programs evaluated can be seen in Table 2.

Table 2. Summary of Prescriptive Program Data

	Number of Measures	Ex-Ante Savings (kWh)
Utility A PY1	141	36,043,206
Utility A PY2	104	10,745,390
Utility B PY1	17	6,036,718
Utility B PY2	209	26,601,977
Utility C PY2	172	14,728,950
Totals	**643**	**94,156,240**

Similar to the custom evaluations, all program evaluations were conducted using the evaluation methodology outlined in the California Evaluation Framework, or the International Performance Measurement and Verification Protocol (IPMVP) evaluation options A-D.

Data Analysis

The evaluation results for both custom and prescriptive programs were used to quantify the differences between the desk review and on-site visits. The data sets are examined as a whole population. However, the individual program performance is also examined to determine the impact of the different review methods on the specific sample realization rate.

Custom Programs

Entire Data Set. The custom programs evaluated consisted of 210 projects from eight different evaluation cycles, and a total of 145.4 GWh of electricity savings. There were a variety of measures including lighting replacements, motors, variable frequency drives, cooling, agricultural, and process improvements.

The first step in determining if the two evaluation methods differ significantly is to examine the realization rates of the population of projects both after a desk review and after on-site measurement and verification have taken place. The realization rate is defined by the California Evaluation Framework (TecMarket Works 2004) as the original savings claimed divided by the verified savings. This is presented in Figure 2.

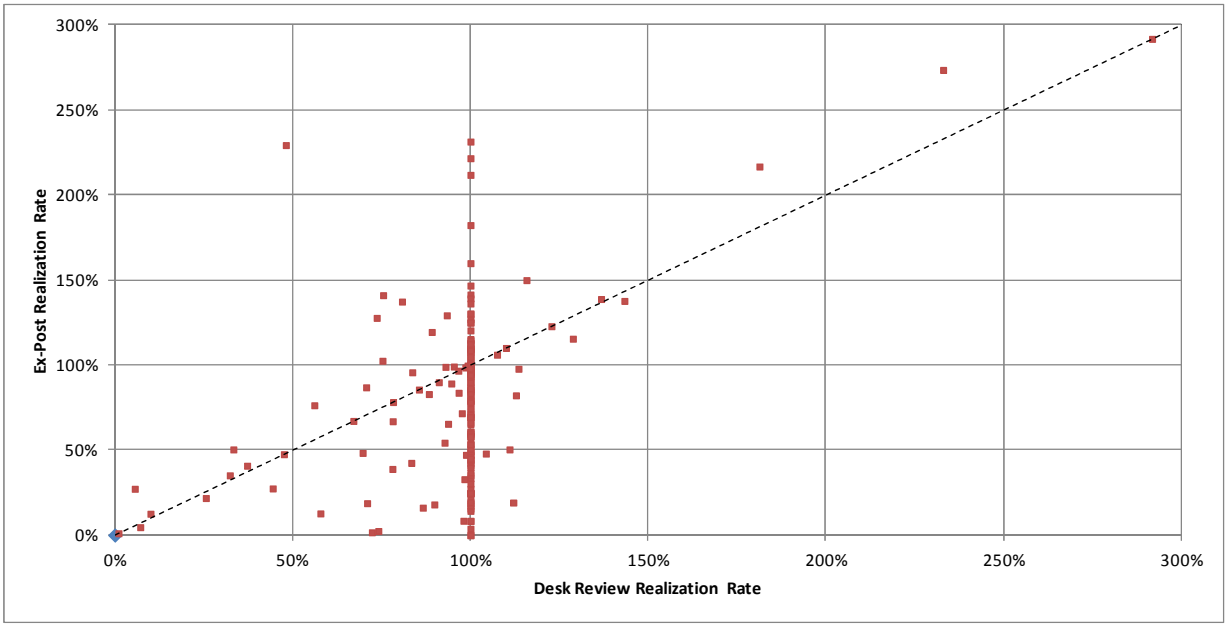

Figure 2. On-site Realization Rate With Respect To Desk Review Realization Rate

The dotted line shown in Figure 1 represents the ideal scenario where the desk review and on-site realization rates would be equivalent. Examination of the data in Figure 2 shows that there is a distinct correlation between the desk review realization rate and the on-site realization rate. This was substantiated by the fact that the statistical correlation of 0.5, showing a moderate correlation.

An interesting characteristic of the data from Figure 2 is the large vertical clump of data at the desk review equals 100%. This is due to the fact that after the desk review, 78% of the measures evaluated had realization rates between 90% and 110%. However, after the on-site measurement and verification, that percentage dropped to 26%. The large clustering of data near 100% after the desk review and the scattering of the data after the on-site measurement and verification can be seen in Figure 3.

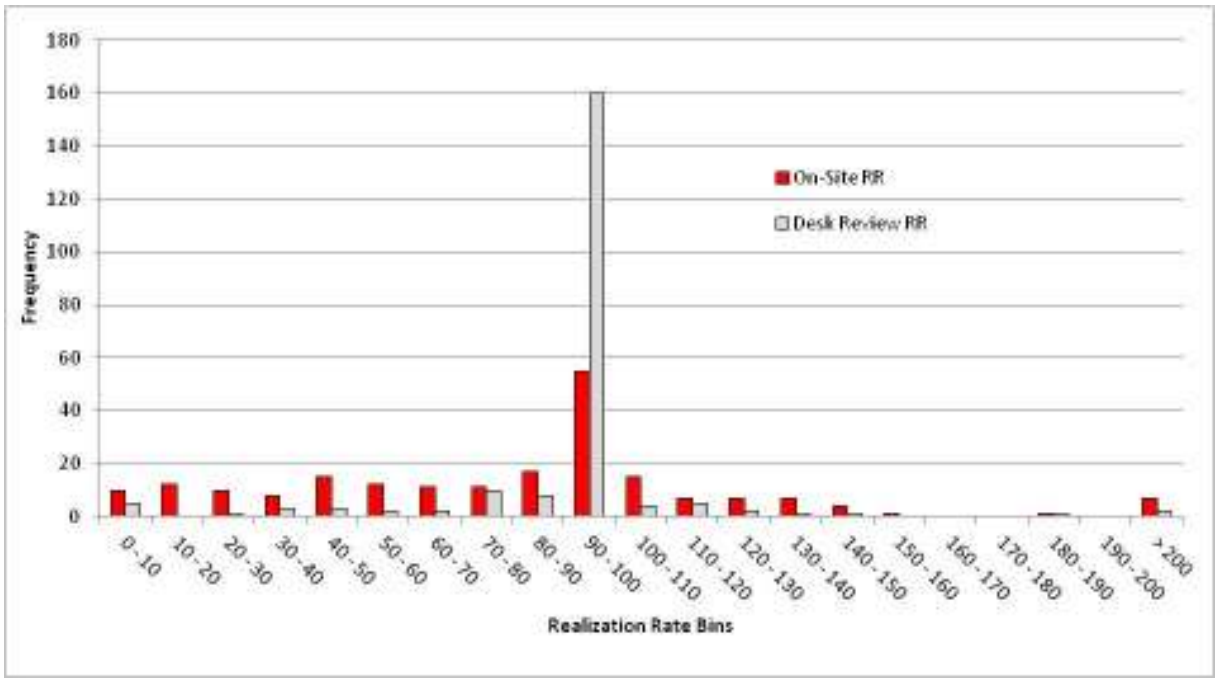

Figure 3. Frequency of Realization Rates After Desk Review and After On-Site M&V

The data from Figure 3 clearly shows that after the on-site visits were completed the realization rates became much less concentrated near 100%. However, the data does still exhibit a somewhat log-normal distribution centered around 100%. After completion of the desk review, the prescriptive measure population examined had a mean realization rate of 95.8%, while after the on-sites were completed the mean realization rate was 82.7%.

Individual Programs. The difference in the evaluation methods can also be seen when examining the programs individually. The custom data set was comprised of eight individual program year evaluation samples. The difference in the sample realization rates after the two evaluation methods can be seen in Table 3.

Table 3. Individual Program Sample Realization Rates

	Desk Review RR	On-site RR	Difference
Utility A PY1	98.9%	94.7%	-4.1%
Utility A PY2	88.4%	105.8%	17.3%
Utility A PY3	97.7%	85.7%	-12.0%
Utility B	99.9%	94.5%	-5.4%
Utility C	73.7%	58.2%	-15.5%
Utility D	99.1%	55.8%	-43.3%
Utility E	89.4%	74.1%	-15.3%
Utility F	99.4%	52.7%	-46.7%

There were two of the programs (Utility A PY1 and Utility B) where the difference between the desk review and on-site review realization rates was not substantial. However, the average magnitude[1] of

1 The average magnitude is calculated as the average of the absolute value of the difference. This provides an estimate of

the difference in realization rate between the two evaluation methods was 20.0%, and the average difference was -15.6%[2].

In order to more fully understand the implications of the effect of the review process on the program ex-post savings, the savings for each sample were extrapolated back to the program population. Population data was available for four programs, and the program realization rates and 90% confidence intervals were calculated using both the desk review and onsite data collection information. The results of this examination are shown in Table 4 and Figure 4.

Table 4. Realization Rate and 90% Confidence Intervals for Select Programs

	Desk Review RR	On-site Review RR
Utility A PY1	98.4% ± 0.7%	94.9% ± 7.4%
Utility A PY2	105.8% ± 10.4%	88.4% ± 8.5%
Utility A PY3	98.9% ± 2.1%	94.9% ± 5.9%
Utility C	77.2% ± 17.2%	57.1% ± 9.5%

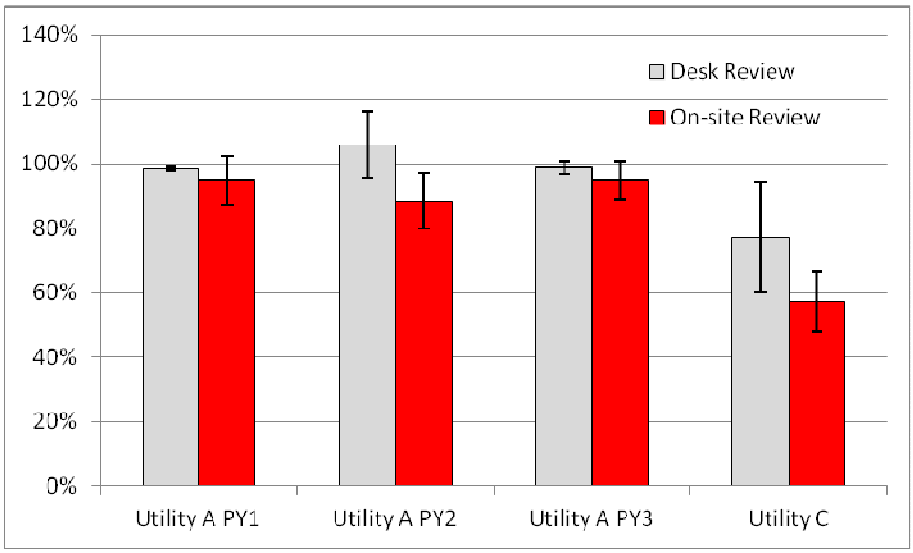

Figure 4. Realization Rate and 90% Confidence Intervals for Select Programs

No clear correlation could be made from this data set to determine the relationship between the expected realization rate for a program with a desk review and the realization rate for the same program with an evaluation that included on-site data collection. However, it is important to note that based on this limited data set, for all four of the programs, the expected realization rate for the program based on the on-sites data collection method is outside the 90% confidence interval for the savings based on the desk review only method.

Prescriptive Programs

Entire Data Set. The prescriptive programs evaluated consisted of 643 measures from five different evaluation cycles, and a total of 94,156,240 kWh of electricity savings. There were a variety of measures including lighting replacements, lighting occupancy sensors, variable frequency drives, high efficiency motors, and heating ventilation and air conditioning equipment.

---

the average difference independent of direction.

2 This is the mean of the numbers found in Table 3.

The first step in determining if the two evaluation methods differ significantly is to examine the realization rates of the population of projects both after a desk review and after on-site measurement and verification have taken place. This is presented in Figure 5.

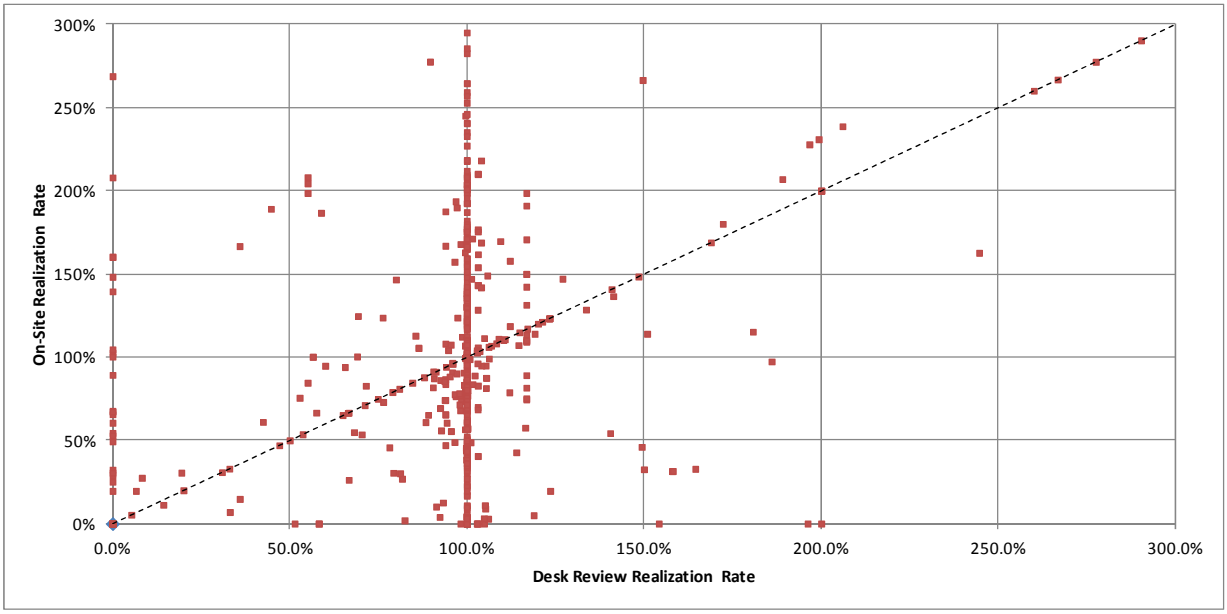

Figure 5. On-site Realization Rate With Respect To Desk Review Realization Rate

The dotted line shown in Figure 5 represents the ideal scenario where the desk review and on-site realization rates would be equivalent. The solid line represents the best linear curve fit to the data[3]. Examination of the data in Figure 5 shows that there is little correlation between the desk review realization rate and the on-site realization rate. This was substantiated by the fact that the statistical correlation of 0.27, showing a weak correlation[4].

An interesting characteristic of the data from Figure 5 is the large vertical clump of data at the desk review equals 100% mark. This is due to the fact that after the desk review, 73% of the measures evaluated had realization rates between 90% and 110%. However, after the on-site measurement and verification, that percentage dropped to 21%. The large clustering of data near 100% after the desk review and the scattering of the data after the on-site measurement and verification can be seen in Figure 6.

3 The linear curve fit is not an accurate representation of the data as the R^2 value is extremely low at 0.0745.
4 The statistical correlation is defined as the covariance of the two data sets divided by the product of their standard deviations.

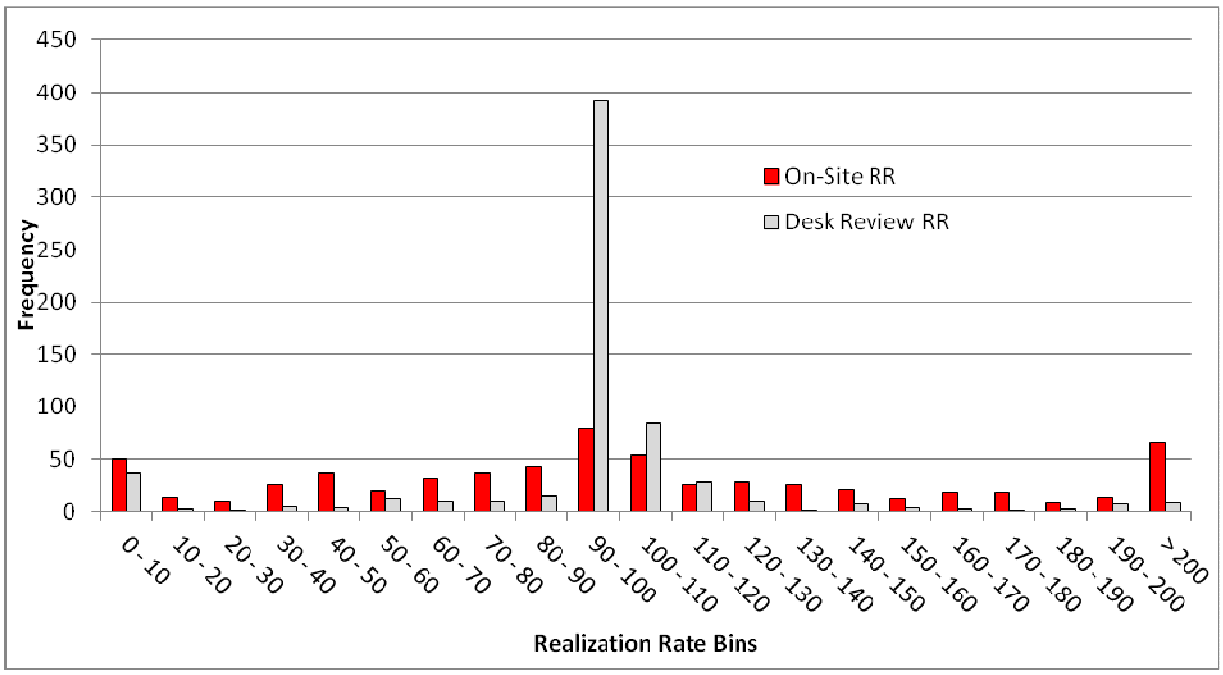

Figure 6. Frequency of Realization Rates After Desk Review and After On-Site M&V

The data from Figure 6 clearly shows that after the on-site visits were completed the realization rates became much less concentrated near 100%. However, the data does still exhibit a somewhat normal distribution centered around 100%. After completion of the desk review, the prescriptive measure population examined had a mean realization rate of 97.0%, while after the on-sites were completed the mean realization rate was 110.2%.

As previously stated, 73% of all measures evaluated had realization rates between 90% and 110%, a 20% range, after the desk review was completed. In order to encompass that same 73% of data points after the onsite visits were completed, the range must be expanded from 30% to 180%, a 150% range.

Individual Programs. The difference in the evaluation methods can also be seen when examining the programs individually. The prescriptive data set was comprised of five individual program year evaluation samples[5]. The difference in the sample realization rates[6] after the two evaluation methods can be seen in Table 5.

Table 5. Individual Program Sample Realization Rates

	Desk Review RR	On-site RR	Difference
Utility A PY2	102.7%	134.2%	31.5%
Utility A PY3	91.0%	69.7%	-21.3%
Utility B PY2	110.3%	127.7%	17.3%
Utility B PY3	82.4%	106.5%	24.1%
Utility C PY2	100.3%	36.5%	-63.8%

5 The five evaluation cycles were two for one utility, two for a second utility, and one for a third utility.
6 The sample realization rate was calculated as the sum of the ex-post savings for all projects divided by the sum of the ex-ante savings for all projects in the sample. No stratification or weighting was applied.

The data available at this time does not allow for the extrapolation of the prescriptive samples back to the respective program populations, as was done with several of the custom programs evaluated. All of the programs had the realization rates differ by greater than 17.0%, and the average magnitude[7] of the difference in sample realization rate was 31.6%, and the average difference was -2.4%.

Programs Based On TRMs. All five of the prescriptive programs examined in this paper are based on a state wide or program specific TRM or similar document. The TRMs all define the deemed savings values or calculation methodologies for all measures, and often for several building types within each measure. Parameters such as lighting hours of operation, effective full load hours, and baseline equipment were all documented, and the calculation algorithms were accurate. The TRMs have also been approved by the program evaluators, and regular updates are being made during the evaluation cycles.

The question surrounding programs that have TRMs in place is what level of evaluation rigor is required to determine how the program is performing. There have been several papers published (Cleff, 2011) stating that the use of TRMs negates the need for traditional measurement and verification evaluation techniques, and that the evaluation can focus on the installation rates of equipment only. There have been numerous discussions regarding this fact with those involved in the programs used as the basis of this paper as well. Three of the programs examined have the installation rate data available[8]. A comparison of the installation rate, desk review realization rate, and on-site review realization can be seen in Table 6.

Table 6. Installation Rate, Desk Review Realization Rate and On-site Realization Rate for Select Programs.

	Installation Rate	Desk Review RR	On-Site Review RR
Utility A PY3	101.0%	91.0%	69.7%
Utility B PY3	110.0%	82.4%	106.5%
Utility C PY3	101.2%	100.3%	36.5%

While the number of programs is not substantial enough to make broad statements about the installation rate argument, the data does show that assuming the installation rate provides evaluation results similar to other evaluation activities may not always be accurate.

Conclusions

Custom Programs. The analysis of the custom programs showed that there is a moderate correlation between the realization rates obtained doing a desk review, and the realization rate obtained after performing on-site measurement and verification for any given measure. This is likely due to several factors. First, the original savings estimates for custom projects are calculated based on information collected about the specific project. Therefore, the calculated savings estimates, at least in theory, should have a high level of correlation to the realized savings. Second, because significant levels of information is collected as part of the original calculation process, more information is available for the evaluator to adjust the savings estimates, compared to prescriptive projects where the information level is often limited to equipment specifications. Finally, because each project is calculated individually,

7 The average magnitude is calculated as the average of the absolute value of the difference. This provides an estimate of the average difference independent of direction.

8 The installation rate was determined during an on-site visit. The quantity of any given measure was visually verified, and its specifications were recorded to ensure it was program qualifying. The installation rate is defined as the visually verified program qualifying quantity obtained during the site visit divided by the claimed ex-ante measure quantity.

as part of the review process, the calculation methodology for the each project is reviewed as well. Therefore, any calculation errors can be identified as part of the desk review process.

Although the realization rates did show some correlation, the program level realization rates were significantly impacted by the evaluation methodology, as all four programs had on-site realization rates that were lower than the desk review realization rates. For all four of the programs reviewed, the expected program realization rate based on the on-site review methodology was outside the 90% confidence interval projected by the desk review only method.

Prescriptive Programs. The analysis of the prescriptive programs showed that there is little correlation between the realization rate obtained doing a desk review, and the realization rate obtained after performing on-site measurement and verification for any given measure. In several cases, the evaluation sample realization rates were also significantly different. The mean difference in evaluation sample realization rates was 31.6%, while the maximum was 63.8%.

The results for prescriptive programs fall in line with what would be expected. The documentation gathered by program implementers for prescriptive programs is generally minimal because the savings calculations are already completed. Often times, the documentation will include only the program application and the installed equipment specifications. Performing a successful desk review is dependant on having enough information to determine how the system is actually functioning. Prescriptive documentation does not meet that requirement, nor is it cost effective to obtain.

The use of on-site measurement and verification allows programs to obtain a better understanding of how customers are actually using equipment. Additionally, results from prescriptive M&V activities can be used for prospective planning, or suggested updates to TRMs, neither of which are possible with only a desk review.

Future Work. The data analyzed for this report focused on a wide range of measures and programs. Future work on this topic will be directed into three main areas. The first area would be how different evaluation methods, especially for prescriptive programs, impact the overall program population realization rates and confidence intervals. Second, examine the differences in realization rates for different technologies. For example, are lighting measure realization rates as dependant on the evaluation methodology as variable frequency drives? Finally, examine if there is a significant difference between other types of evaluation methods such as phone interviews, on-sites with no metering, and metering compliant with forward capacity marketing evaluation requirements.

References

Cleff, P., A. West, H. Haeri, and T. Jayaweera. 2011. "Impact Evaluation Research Design in a Post TRM World." *In Proceedings of the IEPEC 2011 Conference.* Boston, MA.: International Energy Program Evaluation Conference.

Efficiency Valuation Organization (EVO). 2010. International Performance Measurement & Verification Protocol. September, 2010.

TecMarket Works. California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals, prepared for the California Public Utilities Commission. April 2006.

TecMarket Works. The California Evaluation Framework, prepared for the California Public Utilities Commission and the Project Advisory Group. June 2004.