It's About Time: Doing Integrated Real-Time Impact Evaluations

Sue Haselhorst, ERS, and Joe Dolengo, National Grid

ABSTRACT

It has become clichéd that implementers want evaluation results sooner and with actionable recommendations; however, impact results often arrive two to four years after the measures are installed and paid, for and the evaluator recommendations are stale.

National Grid's (NGrid's) New York Commercial and Industrial (C&I) evaluation study manager has launched a bold and innovative approach to evaluations that is focused on quick turnaround measurement and verification (M&V) and simultaneous process-oriented feedback directed at improving program implementation. This thinking is inspired by the New York Reforming the Energy Vision (REV) initiative, which calls for evaluations that are "designed and implemented to yield timely information that [feeds into] the annual iterations of utility programs." This approach incorporates these features:

- A rolling sample to select sites for M&V during the implementation period, permitting reporting of results months after the measures are installed rather than years.
- Leveraging the granular M&V engineering data collection process to provide program implementers with granular feedback on the application process, technology performance, and on-site operation of the measures.

This paper will report on the implementation of this approach, its reception by the implementation staff, the aspects of the program that have worked well, and where adjustments need to be and have already been made.

Introduction

For years, the realization rate (the ratio of evaluated to tracking savings) was the major deliverable of an impact evaluation. In the late 1990s, regulators began requiring verification of the claimed savings to ensure the reliability, cost-effectiveness, and/or appropriateness of the shareholder incentives. The realization rate, with separate factors for energy and demand, encapsulated this gross savings verification in a single number. The paradigm worked well during a period of stable measures, programs, and goals, and delayed evaluation results were acceptable – although recognized as not ideal. Table 1 presents the average time lapse between the mid-point of the evaluated program year (PY) and the publication date for impact evaluation studies published between 2012 and 2015, inclusive.

Table 1. Average time lapse between installation and impact evaluation

Jurisdiction	Number of impact studies	Average lapse (years)[1]
California	14	2.2
Massachusetts	21	1.9
New York	17	3.6
Total	52	2.5

1-Example: lapse between PY 2010-2012 (mid-point of 6/1/2011) and a study published 6/1/2014 is three years.

In recent years, program goals and budgets have increased dramatically, with even more money expected to flow from the investment community. Codes and standards are changing rapidly and striking deep, decreasing the available energy efficiency potential. This confluence of factors has led to more rapid program design changes, thus leading to a need for quicker evaluations. From a program implementer's point of view, program implementation recommendations two and three years after a program's year-end are likely to be stale and out of date since a program's design, quality control procedures, and measure mix most likely changed in that timeframe.

An Opportunity for Change

In 2015, NGrid's New York C&I lighting retrofit program was queued up for a 'routine' impact evaluation, since it had been three years since the last evaluation and this program was a large contributor to the savings portfolio. Rather than commence with the routine, the study manager saw an opportunity to explore an alternate model enabled by changes in the regulatory landscape created by REV. This new model would seek to:

- Compress the time between the program year and the delivery of evaluation results to implementers and regulators
- Return more than a realization rate by adding in meaningful field observations and analysis leading to program improvements
- Provide a nimble platform to address issues as they arise
- Identify subtle program gaps resulting from organization changes and/or program implementation process changes

The study manger engaged an evaluation contractor and together they considered how a new study design might achieve these goals.

An M&V 2.0 whole-building approach was ruled out for multiple reasons. NGrid had been bewildered with the results from a recent billing analysis of the Small Business Direct Install program. On the surface, the SBDI program was a good candidate for utility meter billing analysis because the savings were a large fraction of the billed usage (20%–40% per account for the second and third savings quartiles) and the customers were relatively homogenous. However, the analysis was inconclusive. The prevalence of multi-metered accounts in the population and the subsequent mismatches between the account number on the application and the billing meter actually serving the measure was a significant contributor to the poor results. A billing analysis of the large C&I lighting retrofit program, with its diverse customer base and relatively small savings per site, was not expected to yield reliable results based on this experience.

The team settled on a rolling sample design where sites are selected for M&V from all of the sites that had installed measures in the previous quarter. The fast turnaround results would then be communicated to implementation through quarterly status reports (QSRs).

Fixing the Timing: Rolling Sample

Traditional evaluations wait until the program year of interest is complete and all tracking savings and incentives are reconciled and neatly accounted for in a definitive "data set of record." Even with the most efficient program administrators (PA), definitive data may not be ready until the end of the quarter following the last day of the PY of interest. On-site M&V teams may not set foot inside a facility for 6 to 9 months after the program year is over because

of the time it takes to contract the work, gather population/program data from the PA, design the sample, gather the sampled project documentation, recruit customers, and design the site M&V plans.

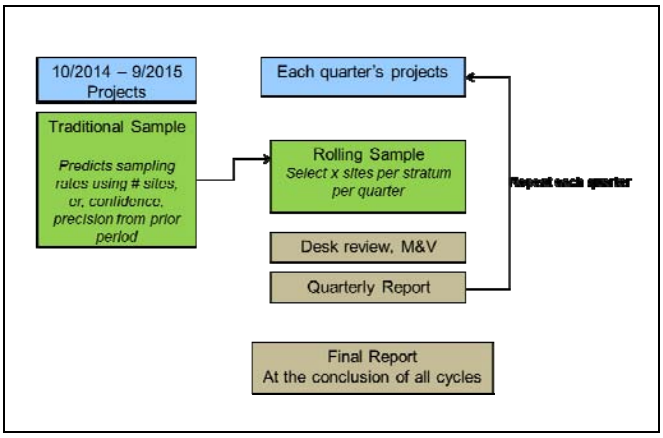
Real-time evaluation cannot wait for a program year to end before the evaluation can begin. A different sampling paradigm is required: this evaluation is using a "rolling sample" strategy in which sites are selected on a regular basis during the program implementation period, with the last sample drawn right after the conclusion of the evaluated program year's closing. With a rolling sample design, program-level results can be reported when the request for proposal for a traditional evaluation might just be making its way through procurement.

A rolling sample is designed to capture the measure performance as the program implementation year proceeds. In conceptualizing the sampling, the team followed these principles:

- Monitor enough sites to produce aggregated realization rates by program track (prescriptive and custom) at ±15% precision by track with a combined precision of ±10% at the 90% confidence level for projects installed between July 2015 and June 2016
- All of the sites within a stratum have an equal probability of selection regardless of quarter.
- The number of sites investigated in a quarter is not fixed but floats as a function of program magnitude in any quarter.
- Since the final population is unknown, the first three quarters are slightly under-sampled, leaving additional budget for deploying more sites in the fourth quarter to balance the final results.

While the concept is simple and intuitive, the mechanics present challenges to ensure that the selected sites are representative and efficiently selected to minimize costs. The team developed the three-part sampling strategy illustrated in Figure 2.

Figure 2. Rolling sample components



Proxy Sample Design

Since the installed population did not exist at the time of the sample design, the evaluation contractor analyzed the population of projects installed between October 2014 and September 2015 to serve as a proxy for the evaluated population. Table 2 presents the results of

the proxy sample design for the prescriptive track. The stratum ranges and sample sizes were derived using the proxy population and a stratified ratio estimation design as outlined in the California framework.

Table 2. Rolling proxy sample design

Stratum	Design sample size	Stratum range(kWh)	Rolling sample size	Probability of selection
0	N/A	0–10,000	N/A	N/A
1	7	10,000–<40,000	6	0.04
2	7	40,000–<120,000	6	0.11
3	7	120,000–<260,000	6	0.21
4	7	260,000+	6	0.67

N/A = Not applicable

Once the proxy sample design was completed, it was adapted to a rolling sample design. This started with a decrease in the stratum quotas size by one, which was designed to result in under-sampling and to accommodate the expected program growth. The probability of selection for each stratum was calculated by dividing the rolling sample size by the population size of the stratum.

Quarterly Sampling

At the conclusion of each quarter in the evaluation period, the tracking data for the previous quarter is collected. The quarter's participants are assigned a random number and binned into the correct stratum using the cut points established in the proxy sample design. Each site with a random number that falls below the random number threshold (the probability of selection in Table 2) is selected for on-site visits. On average, the evaluation contractor expects to conduct desk reviews for approximately twenty sites and on-site M&V for approximately ten sites; however, the actual number varies depending on the level of program activity in the quarter and the random numbers themselves. It is possible no sites will be selected within a stratum in a quarter. The quota of desk reviews in a quarter is defined as twice the number of sites selected for M&V, ranked from least to highest random number, thus ensuring adequate back-ups for primary sites and a larger desk review pool.

Final Adjustment

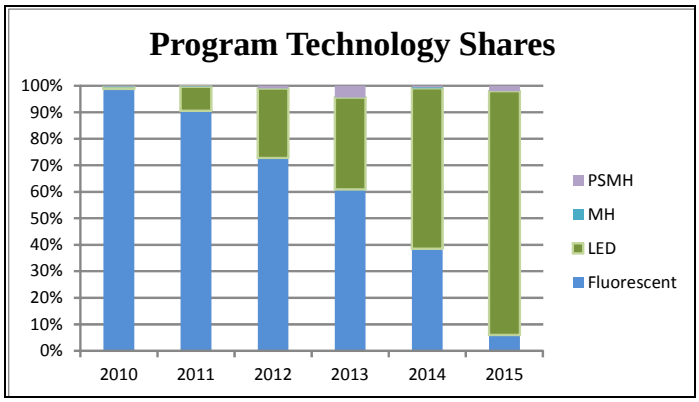
At the conclusion of the last quarter, when a full year's population will be accumulated (July 2016), the entire population will be re-analyzed and the final sample will be selected in a manner that will true-up the final sample sizes to best meet the targeted precision. All of the random numbers assigned to each site in each quarterly sample have been retained. The final stratum sample quotas can be increased to a target quota by increasing the random number threshold thereby sweeping additional sites, across all quarters, into the M&V pool. This is an elegant solution eliminating potential bias quarter to quarter or complicated post hoc stratification.

‘I Want More Than a Realization Rate …’

While the rolling sample solved the timing problem, the next question was: what should the team be looking for in the field and what value might be added to the study beyond the realization rate?

After reviewing recent program data, a clear trend jumped out, as illustrated in Figure 1. The foundation of the lighting retrofit program’s energy efficiency savings had changed from linear fluorescents to light-emitting diodes (LEDs) in the years since the last evaluation (PY2011/2012).

Figure 1. NGRID – NY LED technology ramp-up



This change raised the question, how is this working in the field? A large fraction of the installations were retrofits of existing fixtures designed for fluorescent lamps to a point-source technology. How does this affect the lighting quality? How is it perceived by the customer? This technology transition inspired a research agenda designed to answer the following questions:

- Are the spaces over-lit or under-lit compared to standards for the space and building type?
- What is the quality of the lighting with regards to lumen uniformity, glare, and color rendition?
- Would additional savings have been possible?
- Are occupants who work in the new lighting environment satisfied?

The evaluation team’s initial hypotheses were reviewed by implementation. Implementation requested that data collection include the measurement of lighting power densities (LPD) to help inform the design of a new retrofit performance track. The evaluation team expanded the data collection protocol to include LPD data.

Desk Reviews

The first step in the evaluation is a desk review. In a traditional evaluation, the desk review rarely turns up interesting site-specific findings for lighting. In this evaluation, however, the desk review is as a tool of the process evaluation. The engineer conducts the traditional desk review tasks of confirming the technical aspects of the application and characteristics such as the following:

- Do the project files include the required documents?

- How well does the application accommodate the new technology?
- For custom applications, could the submitted measure qualify as a prescriptive measure?

Lighting Data Collection Protocols

Fixtures designed to optimize the output of a fluorescent linear lamp, with its radial light emissions profile, are being refit with point-source LEDs. Potential problems may include non-uniform distribution of lumens, over-lit spaces (which potentially reduces savings), and other quality issues, like glare. When coupled with the research goal to identify lost opportunities (present if spaces are over-lit), it became imperative to take light-level readings.

It is not a trivial effort to collect systematic light-level measurements in an evaluation context. The methodology must account for ambient lighting and the selection of lighting industry lumen standards that are appropriate for the space and business type. The data collection must be conducted in a manner that is cost effective, not overly intrusive, repeatable, and systematic. A lighting measurement protocol was developed to collect light level, light quality and LPDs within the site sample of spaces that receive light loggers. The lighting protocol includes these elements:

- A single light level measurement is recorded at a key horizontal or vertical (for retail) surface unless the engineer observes uneven light distribution, which will trigger a protocol for taking multiple measurements at regular intervals throughout the space.
- The engineer attempts to determine the artificial light contribution by taking readings that are not affected by sunlight (by selecting an interior office vs. exterior office, for example).
- The engineer identifies how the space is used by the occupant for the purpose of selecting applicable foot-candle benchmarks for determining whether the space is over or under-lit.
- All the lighting in the space in inventoried and the area is measured and recorded for lighting power density calculations
- Multiple occupants within the spaces are surveyed to record their perception of lighting quality, color rendition, glare, level, and distribution.

Reporting Rolling Results

The purpose of M&V innovation is to bring back field observations to the program implementers, allowing them to more rapidly make mid-course corrections. This requires a reimagining of the roles of the program implementers and evaluators and requires quicker, more cohesive communications.

Role Changes – Moving Evaluation to a Participatory Role

For this new paradigm to work, program implementation must change their traditionally distrustful and adversarial views of the evaluator to one in which the evaluator brings timely data to their attention, allowing for mid-stream program adjustments and improvements. The evaluators must also rethink their roles and become “the eyes and ears of the program,” alert to what is happening in the field and prepared to quickly capture observations accurately, completely, and systematically.

As part of the evaluation roll-out, the study manger had discussions with the program implementation counterparts about the research goals. The implementation staff’s response was

polite but skeptical. The evaluation team is hopeful that this attitude will change as evaluation cycles progress. The team also intends to listen to the implementer responses as the cycle results roll out and be responsive to their points of interest. It will take time.

Interestingly, the evaluation M&V site team was also, initially, polite but skeptical that the study manager was sincerely interested in their unscripted observations of the site. However, the study manager has encouraged and listened to individual site observations in every team check-in (weekly). This forum, where the engineers are encouraged to discuss what they see and how the process is working in the field, has led to a more expansive communication of the discrepancies and technology specific observations.

Quarterly Results

The key formal method of communications is the quarterly status report (QSR). The QSR is issued, in theory, within 2 weeks of the close of the quarter and reports both quarterly and cumulative-to-date findings and results. For measures installed in Q1, for example, most desk reviews will be reported in the first QSR and M&V results in the second QSR, or about 6 months after the measures are installed. The content of the report includes both quantitative (metrics) and qualitative (narrative) data.

Metrics

Program implementers intensely monitor the program tracking metrics that measure their progress to the annual goals. Early results in the year provide the program manager with feedback on whether they need to step up or throttle back their efforts. As the year progresses, the tracking data trends provide increased certainty about whether the goals will be met. In this new evaluation paradigm, the evaluation results will unfold in the same way that a program implementation unfolds. The new evaluation paradigm will provide metrics to measuring evaluation progress over the year and feedback for potential mid-course corrections. As the year draws to a close, the final outcome is previewed with increasing certainty.

Table 3 is a compilation of select metrics that were reported in the first QSR (dated April 2016) and captures select desk review findings from the first quarter.

Table 3. C&I project status and select metrics of custom projects as of April 2016

Program period	Q3 2015
Percentage of sites with incomplete documentation	78%
Sampled sites' tracking savings (GWh)	30.1
Technical Resource Manual (TRM) vs. tracking savings realization rate	81%
Desk review (DR) vs. tracking savings realization rate	101%
Tracking estimated full-load hours (EFLH) (average by site, weighted) for custom measures	5,787
TRM vs. tracking EFLH (average by site, weighted)	83%
Percentage of custom savings eligible for prescriptive incentives	69%
Ratio of custom to prescriptive incentive	276%
Precision (90% confidence)	2%

The evaluation will regularly report program and evaluation metrics throughout the course of the program year. The most important of these metrics are key performance indices (KPIs) that are selected to measure progress towards program improvement goals. Examples of these metrics beyond those presented in Table 3 are as follows:

- Gross savings realization rate and precision (KPI)
- Lost opportunities, as a percentage of tracking savings on a program basis (KPI)
- Ineligible measure rate (KPI)
- LPD, as a percentage of the design LPD
- Lighting over-/under-lit, as a percentage of the tracked savings (assuming a linear relationship between the wattages and lumen output)

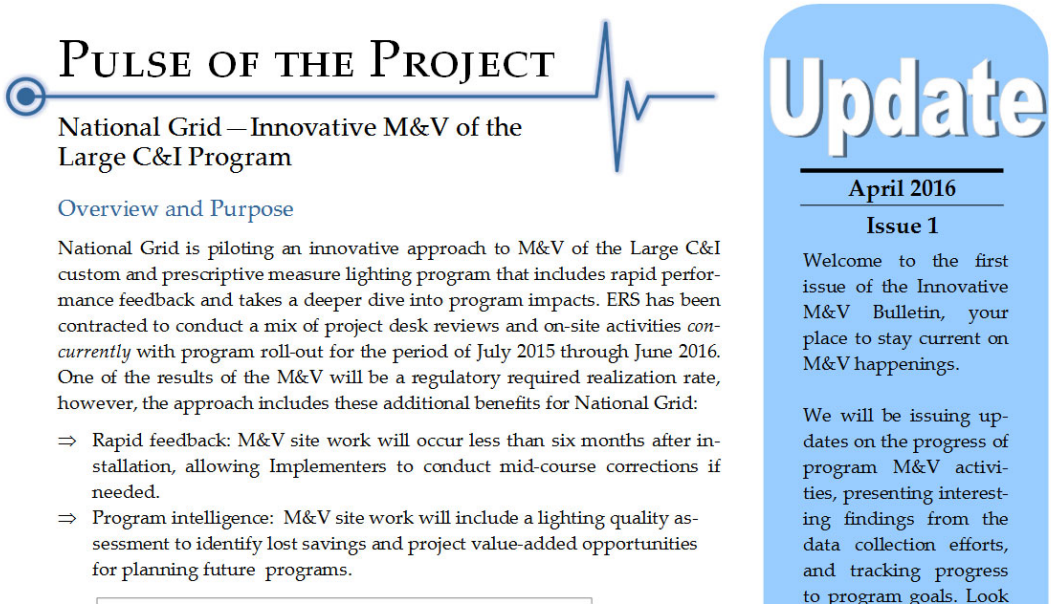
Narrative

The numbers will not tell the whole story. Each QSR will highlight interesting sites where, for example, the savings were significantly different from what had been projected in the application or other issues, as noted in these examples from the first quarter:

- Two hotels selected for on-site M&V had significant discrepancies in the number of fixtures that were actually installed on site. Both of the sites had been installed by the same contractor.
- A customer at a municipal site was very interested in identifying additional opportunities. The lead was characterized and forwarded to the responsible tech rep.

The evaluation team is also experimenting with an alternative reporting format, a two-page quarterly bulletin, which is illustrated in Figure 3. This study manager is gauging the stakeholder responses to both a more formal memo format and the bulletin.

Figure 3. View of the quarterly bulletin



PULSE OF THE PROJECT

National Grid—Innovative M&V of the Large C&I Program

Overview and Purpose

National Grid is piloting an innovative approach to M&V of the Large C&I custom and prescriptive measure lighting program that includes rapid performance feedback and takes a deeper dive into program impacts. ERS has been contracted to conduct a mix of project desk reviews and on-site activities *concurrently* with program roll-out for the period of July 2015 through June 2016. One of the results of the M&V will be a regulatory required realization rate, however, the approach includes these additional benefits for National Grid:

⇒ Rapid feedback: M&V site work will occur less than six months after installation, allowing Implementers to conduct mid-course corrections if needed.

⇒ Program intelligence: M&V site work will include a lighting quality assessment to identify lost savings and project value-added opportunities for planning future programs.

Update

April 2016

Issue 1

Welcome to the first issue of the Innovative M&V Bulletin, your place to stay current on M&V happenings.

We will be issuing updates on the progress of program M&V activities, presenting interesting findings from the data collection efforts, and tracking progress to program goals. Look

Conclusions

NGrid's experiment in rapid and more 'implementation-centric' evaluation looks promising.

The goal for timeliness is on track. This evaluation will be almost finished by the time this paper is published and presented in the ACEEE Summer Study in August 2016. All of the sites should be recruited and three-quarters of the individual M&V site reports completed. The final report is on schedule for December 2016, six months after the close of the program year of study. The PA will have had previews of the evolving realization rate since June, so surprises should be minimal.

The goal to expand data collection and analysis to include program improvements is on track. NGrid has already implemented a study recommendation to add a data field to tracking. The evaluation team is in pursuit of potential recommendations prompted by early observations, including:

- The desk reviews uncovered that the application form is not optimized for LED measures. The evaluation team is cataloging the misalignments and developing potential modifications for implementation's consideration.
- Required documents (like invoices) were not found in all the project files. A more comprehensive look at documentation procedures is underway.
- The desk reviews identified that a number of the custom measures were eligible for prescriptive incentives and were more highly incented than their prescriptive counterparts. The program manager was aware of this trend, which was a result of local market factors, but was surprised by its extent. The implications of this finding are still being explored.

Not everything has gone to plan, however. A lesson learned is that while the organization does seek early program findings, information has implications and stakeholders need to develop appropriate responses to findings that are formative, not final. At the time of this publication in late May, the first quarterly report, drafted in mid-April, is still making its way through the NGrid approval process, one group at a time. The study manager is working with stakeholder groups to develop communication pathways for meaningful, candid, and productive dialogue. The final form of quarterly reporting is likely to evolve and may take the form of multiple sub-group conference calls, a bulletin, or different memos tailored to the needs of different sub-groups.

The field lighting quality protocols have been revised multiple times. Contamination by ambient lighting for taking a valid light level measurement is a problem. This has led to the field engineer scoring the light level measurement to indicate how free it is from a sunlight contribution. The field teams are not able to identify a valid occupant for a lighting quality survey in some spaces (who should be surveyed for spaces like bathrooms or hallways?).

As an affirmation that this approach is heading in the right direction, this evaluation has received positive attention from other jurisdictions. In Massachusetts, both the regulator and program implementation are advocating for including the lighting quality assessment in an upstream lighting program evaluation currently in planning.

References

These following websites were searched for relevant impact evaluation studies for the completion of Table 1.

California Measurement Advisory Council (CALMAC). http://www.calmac.org/search.asp

Massachusetts Energy Efficiency Advisory Council. Commercial and Industrial Studies. http://ma-eeac.org/studies/commercial-and-industrial-studies/

New York State Energy Research Department Advisor (NYSERDA). NYSERDA Evaluation Contractor reports. http://www.nyserda.ny.gov/About/Publications/Program-Planning-Status-and-Evaluation-Reports/Evaluation-Contractor-Reports