Hard Times for an Honest Logger? Optimizing a Small Business Direct Install Logger Study in an M&V 2.0 Landscape

David Barclay, Michael Strom, Julian Ricardo, NMR Group, Somerville, MA
Joseph Dolengo, National Grid, Waltham, MA

ABSTRACT

Measurement & Verification (M&V) 2.0 aims to move M&V toward instantaneous evaluation. This can be accomplished by making large sets of granular energy usage data available and by relying on it instead of relying on the sampling of customers. As we approach this future and consider how we might shape it, we need a comprehensive understanding of the emerging techniques, including consideration of the relative costs, accuracy, and feasibility of various approaches. To that end, the authors examine how traditional M&V techniques can be used to test, validate, and complement emerging techniques. They note that results presented here are preliminary; finalized results will appear in a forthcoming impact evaluation report to National Grid.

Alignment and variability of savings estimated using traditional photocell lighting loggers are compared against M&V2.0 approaches, against documented hours of use (HOU) and against savings based on a combination of TRM and program data from a small business direct install lighting program. Findings are from 100 retrofit projects in one Northeastern state.

The influence of building end-uses and duration of data collection on savings estimates was analyzed. The overarching goal was to assess opportunities for M&V technologies to improve the reliability of savings estimates derived using M&V2.0 techniques or TRM-based HOU values. The authors also assess whether the lighting loggers meet evaluators' and program administrators' needs for quicker delivery of evaluation findings.

The approach involved calculating sites' realization rates from the various methods, comparing normalized root-mean-squared-error (CV(RMSE) values of savings estimates, and developing documentation and a web app to enable rolling evaluation updates.

Introduction

How should program administrators and evaluators weigh the relative cost, accuracy, and feasibility of emerging Measurement & Verification (M&V) methods against more-established ones? How can traditional M&V play a role in testing, validating, or otherwise improving the reliability of emerging, innovative methods? Within the context of small business lighting programs, the phrasing of these questions can suggest a hard binary exists between M&V 1.0 and 2.0—between billing analysis or photocell loggers and a networked Internet of Things (IoT) or sub-hourly advanced metering.

However, the authors believe there might be opportunities to optimize for cost, statistical rigor, and/or resource allocation of metering equipment by using some combination of traditional and emerging methods. Emphasizing responsiveness to these variables reflects an acknowledgment that evaluators' and program administrators' (PAs) priorities can shift for any number of reasons, no less in a highly fluid world of M&V methods and media for reporting them. The ability to choose from an array of approaches based on well-documented metrics is especially valuable among heterogeneous, harder-to-reach portfolios of small businesses. Furthermore, existing approaches can benefit from tools and methods more often associated with M&V 2.0 and generally larger volumes of data.

The purpose of this paper is to document how we refined existing commercial lighting M&V methods (using lighting loggers) by adopting tools and practices from the M&V 2.0 playbook that improve the reproducibility of the findings. Three goals guided our approach: (1) using a "tidy" data framework where possible, (2) generating rapid feedback reports for quality control during data collection and logger inventory tracking, and (3) packaging the analysis as a web app for rolling evaluation updates.

In addition to documenting the approach, we assess the validity of savings estimates from photocell lighting loggers against savings calculated using 2019 New York Technical Reference Manual (NYTRM) and utility data. We also compare logger-based savings against estimates derived using four M&V methods discussed in Barclay et al. (2018), which were implemented at a subset of the sites studied here. They include account-level billing analysis, whole building metering using two separate intervals (hourly and daily), and subcircuit metering. This expanded study increases the sample size of logger installations by a factor of three and the geographic area covered from county- to state-level.

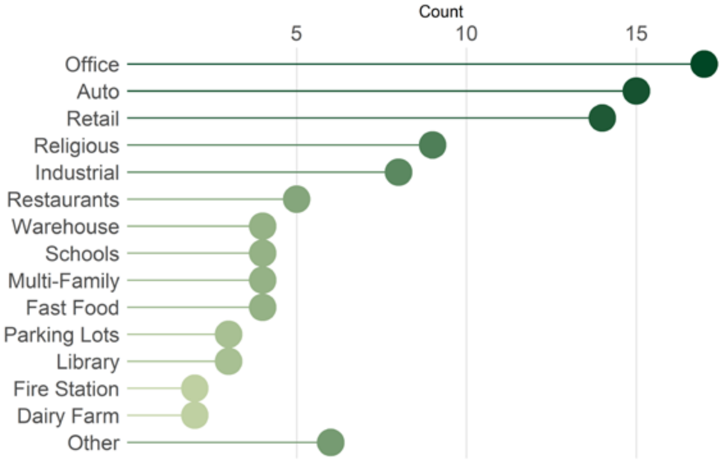
Methods

Sampling and customer recruitment

We developed a representative sample of customers across four strata for the study: year of program participation, program delivery channel, program vendor, and estimated energy savings (based on utility program data[1]).

Our parameter of interest was program energy savings, which we used to determine initial sample sizes following the procedure laid out in the Uniform Methods Project (UMP) Sample Design Cross-Cutting Protocol for deriving a coefficient of variation (CV) from previous evaluation work (NREL 2017). In addition, we used the CV calculated from 32 previously conducted M&V visits (Barclay et al. 2018) and tested all subsequent sampling design for 10% relative precision at the 80% confidence level.[2]

Altogether, the authors contacted 525 customers, spanning three program years and all program territory. From these contacts, we managed visits to audit customers' upgraded lighting and to install lighting loggers covering 100 retrofit projects completed through the program. Offices were the most frequent among sampled business types (n=17), followed closely by auto-related and retail establishments (n=15 and 14, respectively). These three types comprised nearly half (46%) of visited businesses.



[1] Measure descriptions, quantities, and savings estimates in program data, supplemented with wattages from program fixture tables where necessary.

[2] Adheres to ISO-NE requirements for sampling.

Figure 1. Business types visited as part of the study (n=100).

Logging procedure

Figure 2 illustrates the monthly volumes of lighting loggers installed in and retrieved from facilities involved in the study. The authors achieved these volumes using an inventory of 559 loggers, deployed a total of 1115 times.

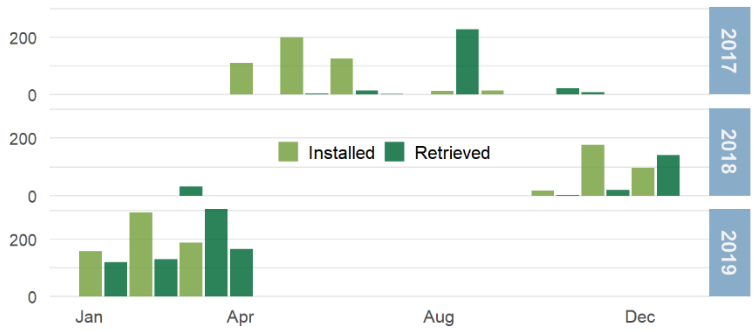
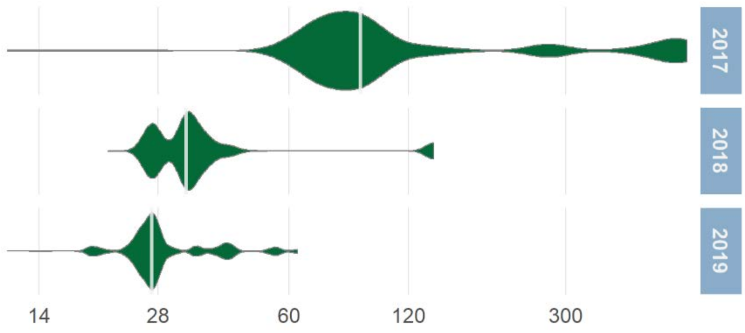


Figure 2. Monthly logger installation and retrieval volume while monitoring small business lighting usage. The figure spans two deployment periods: one involving 34 retrofits from 2017 and the other involving 66 retrofits from 2018 and 2019.

Intermediate analysis of the logger data we collected in 2017 indicated that a four-week metering period would capture most small business' lighting schedule cycles (Ricardo et al. 2018).[3] This threshold exceeded the minimum of two weeks that the UMP recommends for sites with constant schedules and complied with their recommendation for additional metering time with variable or irregular site schedules. Some sites had seasonal usage and/or well-defined inactive periods (e.g., schools); we avoided collecting data during these times and, as needed, accounted for their presence in captured usage data during analysis.

As a result, the overall distribution of metering periods we used for this study ranged from three weeks to over one year, but the distribution among sites visited after 2017 was much narrower. The median metering period for this later set of visits was four weeks. For consistency's sake, we limit comparative analysis of savings across sites to results based on the first five weeks of logged lighting usage, or the maximum metering period achieved if less than five weeks.



[3] More concretely, 75% of the visited sites (n=32) showed less than 2% week-to-week variation in savings estimated after only two weeks.

Figure 3. Distribution of metering periods among business visited in 2017-2019. The white vertical lines indicate annual median metering periods.

In large part, this study follows procedures for installing loggers, analyzing their data, and arriving at estimates of savings from small business lighting retrofits as described in KEMA (2014) and Barclay et al. (2018). We use energy (kWh) savings as the functional unit for analysis, as opposed to hours of use (HOU), and evaluate how data loggers perform in capturing these savings over different metering periods. The final logger savings estimates incorporate HVAC interactive effects multipliers from Appendix D of the 2019 NYTRM.

Excluding fixtures on a timer or photocell sensor, the authors attributed savings to unlogged fixtures using average HOU derived from similar space types at the same site. These scenarios arose primarily at larger sites where the authors could not access and/or install a logger in every room. For unlogged fixtures lacking similar space types, we applied the average HOU for the site. The authors also installed loggers on up to two 24/7 circuits at sites where they comprised part of a facility's lighting system to confirm expected operation, which we otherwise assumed. We developed a procedure for attributing savings to exterior fixtures on a timer, which was confirmed through observation and/or conversations with the customer. The authors calculated the HOU of exterior fixtures on daylight timers and analog timers during desk reviews.

Finally, we selected a slice of logger savings —the estimated savings after week three or four—to calculate savings ratios and other summary statistics to compare logger model performance against TRM and program data, as well as M&V 2.0 data where available.

Reproducible analysis and data management

Other than the proprietary software for accessing lighting logger data, the authors developed tools in R and related open-source software for data cleaning, management, and analysis. In so doing, we aimed to refine a reproducible approach for sampling customers and implementing existing commercial lighting M&V methods. The approach produced a self-contained set of functions and documentation so future similar evaluations could be conducted. This complemented our need to flexibly compare model outputs—logger, TRM, program data, and various M&V 2.0 model results—at both the site and program/domain level of the evaluation. It also allowed us to deliver site visit progress and logger results throughout the evaluation via web app, as opposed to only publishing cumulative results in a single report.

In a future of evaluation where logger deployment occurs primarily as a complementary or backup approach, a reproducible approach to compiling, cleaning, and analyzing logger data will make it easier to integrate logger data into the generally larger, more-automated data flows that characterize M&V 2.0. More abstractly, and beyond the realm of collected data, updating the language of logger analysis so it aligns with M&V 2.0 will make like-to-like comparisons (of advantages, drawbacks, etc.) easier during program design and evaluation.

Over the course of our work, we identified some data handling practices more closely associated with M&V 2.0 and/or Big Data that were useful in improving our existing M&V approaches, including the following:

- replicable computation with complex data sets (Peng 2011)
- "tidy" data framework (Wickham 2014)
- optimizing the computational footprint of web app

Results

M&V methods comparison

Taking the ratio of logger- and TRM-derived savings after a four- and five-week metering periods gave average savings estimate ratios per site of 0.99 and 0.98. The average ratio we calculated after three weeks, 0.82, suggested that we had not yet captured a full lighting schedule cycle at all sites, as NYTRM (2019) recommends. The discrepancy in ratios between the two methods can be attributed to differences in HOU, as we held wattage savings constant across both logger and TRM savings calculations.

Table 1. Summary statistics for logger-TRM kWh savings realization rates (RR) by metering period

Metering period (wk)	N	Average RR (TRM)	Average RR (Program)	Relative Precision (Program, 80% conf.)	CV(RMSE) (TRM)
3	100	0.82	0.76	10.0%	0.66
4	100	0.99	0.93	10.1%	0.62
5	98	0.98	0.92	10.2%	0.58

Figure 4 shows savings estimate ratios disaggregated by business type and weeks of metering data considered. The upper set of axes shows the averages for all business types over time, while the bottom portion highlights RR after logging four weeks of lighting usage for business types where n=3 or greater at that metering period length.

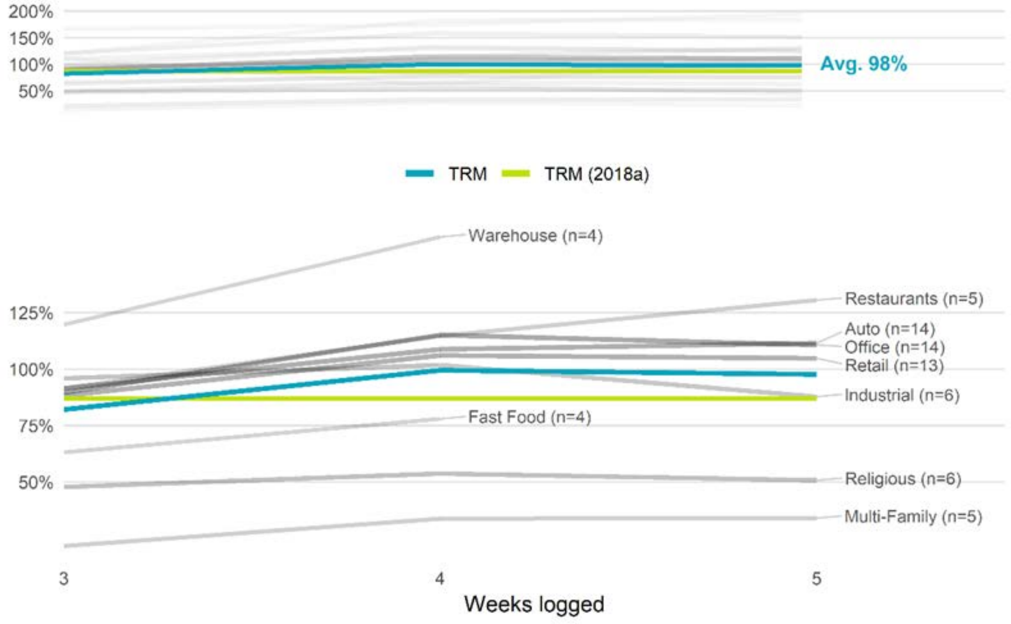


Figure 4. Average site savings estimate ratios of logger-derived kWh savings compared against TRM-derived kWh savings and results from Barclay et al. (2018). The upper set shows the average ratio and the spread of ratios for all sampled business types over time. The lower set shows the same quantities for business types visited with sample size > 3 after a four-week metering period.

Both sets of curves illustrate that most business types (among those with n=3 or greater) produced qualitatively similar realization rate curves over time. Namely, rates that tend to increase between weeks two and three then trend downward between weeks three and four. With the exception of warehouse and restaurant sites, this sequence played out at business types whether their RR were in line with, above, or below TRM expectations.

Figure 5 takes the savings estimates calculated with four weeks of logger data and compares them against the savings attributed to each site using two separate baselines: one with TRM-based savings and another with savings from program data. For comparison, the figure also includes lines indicating the average ratio for the full sample, as well as the average logger-TRM ratio reported in Barclay et al. (2018) for a subset of 32 sites also present in the current dataset.

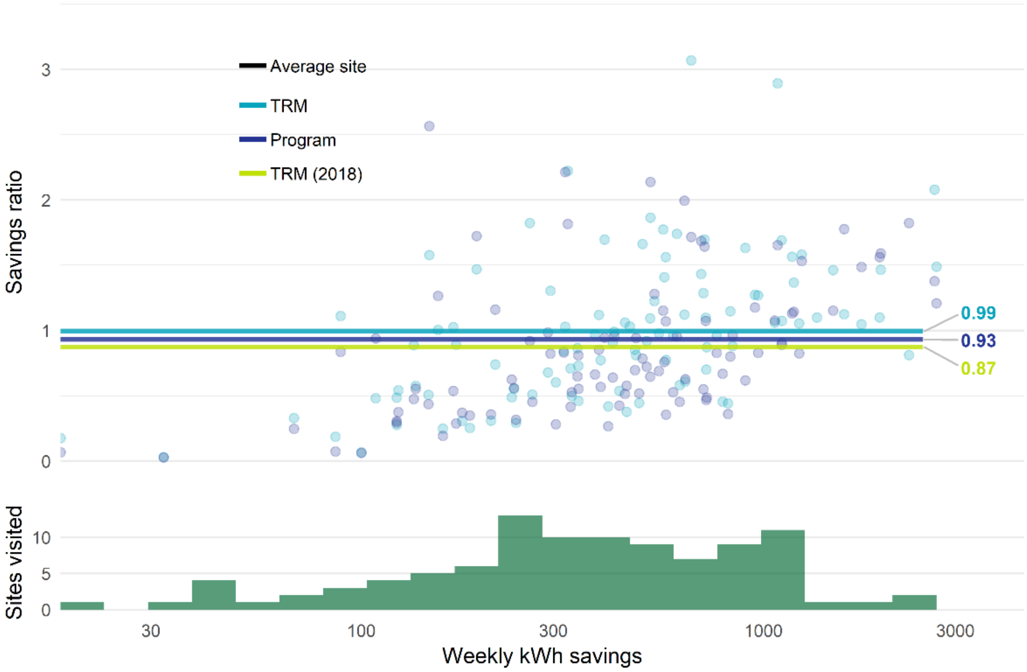


Figure 5. Average savings estimate ratios for logger-derived kWh savings (n=100) after a three-week metering period, calculated with both TRM-derived kWh savings and program savings data. Above, both sets of rates appear as functions of program savings, with the lines indicating the overall average, and points indicating individual site results. Ratios taken between evaluated and program savings at individual sites tended to fall below 1, whereas those taken between evaluated and prescribed TRM savings dominated the higher ratios we observed. Both sets showed a comparable range and variance, however. The average ratio for logger-TRM comparison in Barclay et al. (2018) is also shown. The lower set of axes shows the distribution of program savings among visited sites.

Beyond the average of these site-level RR, we calculate their root-mean-squared-error (RMSE) and its coefficient of variation (CV(RMSE)). With these quantities, we can compare normalized differences in model performance with those we derived in other studies. Also, where sample sizes are sufficient, we can provide an empirical, model-agnostic basis for determining where discrepancies in savings estimates exist at different business types. From there, knowing what parameters the different models do and do not share, we can assess why the discrepancies might exist.

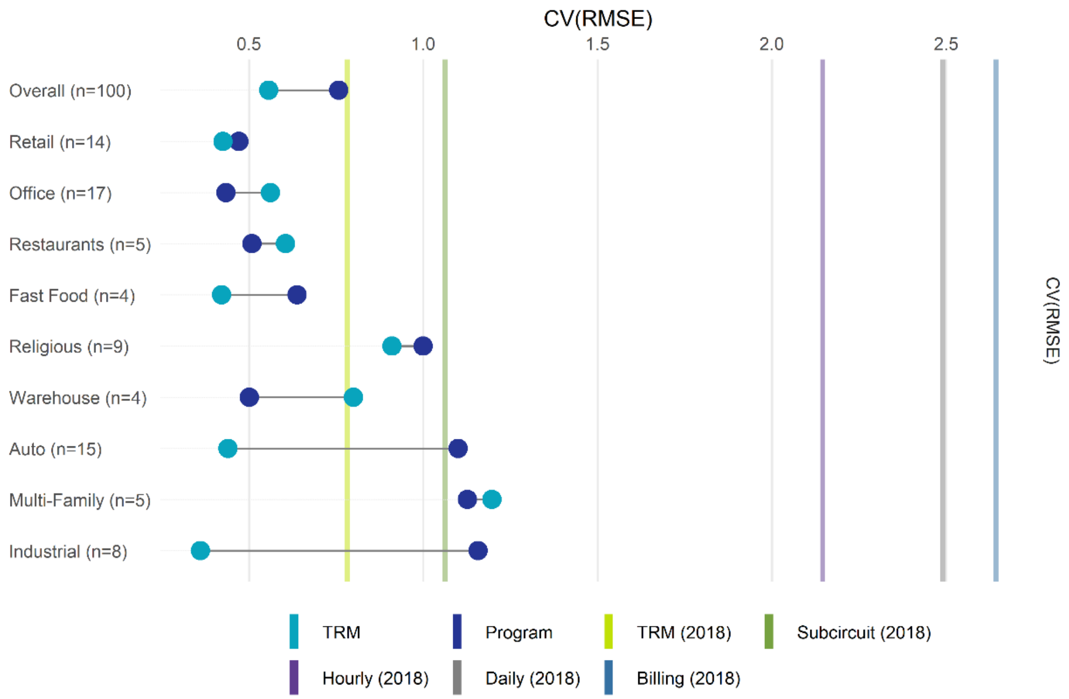


Figure 6. Comparing CV(RMSE) values for RR derived with logger kWh savings (n=100) after a three-week metering period. One set of rates compares logger and TRM-derived kWh savings, while the other uses program savings data. CV(RMSE) reported for logger-TRM comparison and M&V 2.0 methods in Barclay et al. (2018) also shown for reference, though based on smaller sample sizes.

The CV(RMSE) results in Figure 6 corroborate the findings, using average savings estimate ratios: the overall alignment of logger and TRM savings estimates is greater than both the corresponding logger-TRM value in Barclay et al. (2018) and the logger-program rates derived for the first time here. But Figure 6 also demonstrates that a few business types—offices, restaurants, warehouses, and multi-family buildings—showed less error between logger and program savings, which complicates the picture.

Using CV(RMSE) to compare logger, TRM, and program savings estimates with those from M&V 2.0 models, the normalized error from the subcircuit metering approach was the lowest by a factor of two. By this metric, it performed comparably to the logger data at the business types in the bottom half of Figure 6. The other methods described in Barclay et al. (2018)—account-level billing analysis and whole building metering with hourly and daily resolution—did not produce comparable CV(RMSE).[4]

However, the fact that the methods in Figure 6 (marked 2018) reflect a subset (samples ranging from n=17 to 25) of the projects we analyzed for this study (n=100) is a limitation. We would expect some reduction in the CV(RMSE) due to equalizing sample sizes between methods, perhaps on the same order of magnitude (tenths) as the reduction in overall CV(RMSE) using logger data between Barclay et al. (2018) and the present study.

Since logger and TRM savings depend on the same retrofit wattage data, there is some degree of correlation built into their compared values, which is not present to the same extent between logger and program data. The M&V 2.0 methods from Barclay et al. (2018) do not depend on the wattage data used in conjunction with lighting loggers, but the daily and hourly whole-building metering results have the same data source. That is, they vary only in the granularity of data used for the modeling, so there is likely some correlation baked into the comparison of model performance there as well.

[4] We observed the same relationship using average savings estimate ratios.

As such, the logger-TRM CV(RMSE) primarily captures error in the alignment of measured HOU and can serve as a lower bound on the model alignment we might expect from the sample. Similarly, it is likely that the difference in daily-TRM and hourly-TRM CV(RMSE) values in Figure 6 (distance between the blue and grey vertical lines) is largely a reflection of the error that results from adjusting the resolution of energy usage data passed to the models. It is a second-order effect compared to the magnitude of difference between either model's results and those from subcircuit metering.

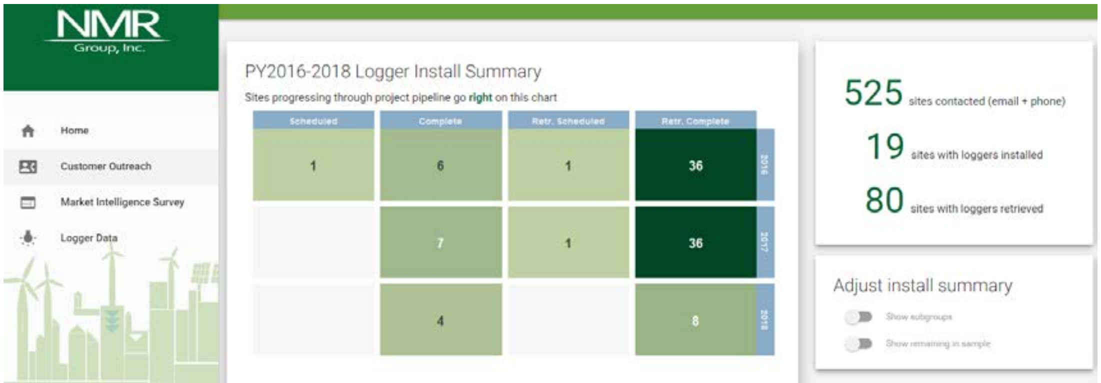
Confining our view only to the TRM and program savings, there is also notable variation in how disparate the results are depending on the choice of baseline. For example, restaurants and warehouses had similar CV(RMSE) values when calculating the logger-program savings ratio, but they differed substantially when looking at TRM-based savings. At the same time, auto-related businesses produced a wider range of CV(RSME) between the two baselines but produced one of the lowest overall CV(RMSE) values when using TRM savings as the baseline.

Logger study web app

The authors provided regular updates on the customer recruitment efforts and logger installations for the study using a dashboard that was built using the Shiny web app framework. In accordance with the prevailing security standards for the project, the dashboard was password-protected, hosted on a local web server with encrypted HTTPS and omitted personally identifiable information (PII).

The two screenshots below capture views of the logger study recruitment and on-site tracker. The first screenshot displays logger installation and retrieval statuses, including toggles for viewing sample strata and the number of visits remaining to reach target completes. The second view is of the contact status summary. We were able to generate these and the figures in the "Logger Data" tab[5] on a roughly weekly basis by selecting tools and approaches at the outset of the evaluation that prioritized reproducibility.

The ability to package functions, documentation and aggregated, anonymized evaluation results made the deployment via web app a substantially easier, more-reliable process. It also reduced barriers to sharing the logger data among stakeholders (i.e., program staff, implementers) for additional analyses.



PY2016-2018 Logger Install Summary

Sites progressing through project pipeline go **right** on this chart

Scheduled	Complete	Retr. Scheduled	Retr. Complete	
1	6	1	36	2016
	7	1	36	2017
	4		8	2018

525 sites contacted (email + phone)

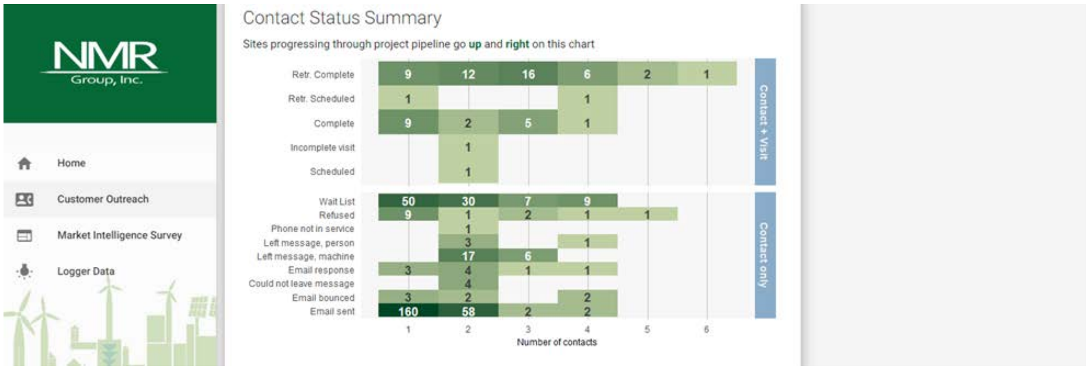
19 sites with loggers installed

80 sites with loggers retrieved

Adjust install summary

Figure 7. Installation and retrieval status tracking dashboard for the logger visits.

[5] Accessible via the menu on the left; already shown as Figures 4, 5, and 6.



Contact Status Summary

Sites progressing through project pipeline go **up** and **right** on this chart

		1	2	3	4	5	6
Contact + Visit	Retr. Complete	9	12	16	6	2	1
	Retr. Scheduled	1			1		
	Complete	9	2	5	1		
	Incomplete visit		1				
	Scheduled		1				
Contact only	Wait List	50	30	7	9		
	Refused	9	1	2	1	1	
	Phone not in service		1				
	Left message, person		3		1		
	Left message, machine		17	6			
	Email response	3	4	1	1		
	Could not leave message		4				
	Email bounced	3	2		2		
	Email sent	160	58	2	2		

Number of contacts

Figure 8. Installation and retrieval status tracking dashboard for the logger visits.

Discussion

Expanding both the sample size and geographic area covered by the logger installations in Barclay et al. (2018), we updated average and CV(RMSE) values for the ratios of lighting logger savings taken against both TRM and program savings data. The results of our study provide evidence that estimating program lighting savings with approximately four weeks of usage data captured via lighting loggers can replicate savings estimates calculated with circuit-based data models. The subcircuit metering model was the only tested M&V 2.0 approach to replicate the average and CV(RMSE) results we calculated with logger-TRM and logger-program savings estimate ratios. Future studies to equalize the sample of sites with both logger and M&V 2.0 data are necessary to fully test this conclusion. A conservative estimate would be that we find a reduction in CV(RMSE) upon additional investigation that is on the same order of magnitude as the reduction in overall CV(RMSE) using logger data between Barclay et al. (2018) and the present study.

Nonetheless, the average and CV(RMSE) savings ratios calculated here serve as a basis for determining where program, implementation, and/or evaluation resources might be supplemented and/or re-allocated to get a more-accurate savings picture (i.e., to reduce error). For example, comparing these values after the initial stage of the study, we determined that a logger metering period of four or five weeks would be sufficient at most businesses we visited for the second study stage.

More generally, the results provide an empirical basis for choosing different business types to pilot M&V 2.0 approaches and compare with existing methods. Stakeholders might decide on a mix of business types to train or test models depending on their program population, as well as the extent of alignment observed between existing methods (e.g., lighting loggers) and TRM or program assumptions.

Among the business types we studied, office, retail, restaurant, and fast food establishments produced relatively low CV(RMSE) values, whether calculated with TRM or program savings. Six of seven industrial businesses also produced results in this range. Given the high degree of confidence in data quality and replication of TRM results, these are business types where logger data could likely serve as the baseline for even piloting M&V 2.0, let alone validating more-mature M&V 2.0 use cases. Warehouses fell in an adjacent group where the CV(RMSE) values were higher depending on the choice of baseline, but aligned with those from our pilot study (Barclay et al. 2018). However, it is worth noting that the sample size for this business type is small compared to the previous category.

Where M&V 2.0 approaches are already in a more-advanced deployment phase, the CV(RMSE) values help determine where loggers would be most effective in validating their results (e.g., performing quality control on already-collected AMI data). By the same token, there could be opportunities for improving existing methods based on the results of M&V 2.0 models. This would primarily be the case at the business types with higher CV(RMSE) values in Figure 6, and generally at larger sites that require additional planning to capture representative usage data with a finite number of loggers to install (e.g., auto, religious, and multi-family buildings).

Model performance aside, there are other considerations to be made in a scenario where program administrators, implementers, and evaluators would be designing an M&V approach from scratch.[6] Stakeholders would weigh pros and cons that include more conventional metrics, such as time and budget constraints on program evaluation. In the context of lighting programs, photocell loggers can be considered a reliable, reusable option, with a known cost profile compared to AMI approaches. They are also still among recommended metering equipment in the UMP for commercial lighting HOU.

Nonetheless, they have limitations: logger attrition, usage changes from stepwise and continuous dimming, natural light pollution, and other disruptions to the signal. There are also labor and time costs to consider, given that temporary metering calls for two site visits, in addition to time for manual data extraction, depending on the logger model. We were able to mitigate some of these costs by working with vendors installing M&V 2.0 metering equipment to also install/retrieve lighting loggers, reducing customer burden in the process too.

For M&V2.0 approaches, advantages include remote data collection, real-time and continuous updates, and the possibility of multiple value streams (e.g., repurposing the usage data for other program or evaluation goals). At the same time, even among the sites in this study, we experienced some obstacles in deploying M&V 2.0 metering: limited network connectivity, constraints due to building characteristics, and customer or other stakeholder preferences regarding permanent installation of metering equipment.

Accounting for these trends and trade-offs, loggers may increasingly become complementary tools, but it would be prudent to expect a continued need for additional inputs to true-up data models based on AMI data, and for alternative methods to deploy where constraints in implementation preclude the use of AMI (budget, time, targeting specific populations, providing network access). Therefore, evaluators should continue looking for ways to extract the greatest value from established approaches using temporary equipment, whether its deployment fulfills a primary or complementary role in data collection. The CV(RMSE) values from lighting logger results presented here provide a measure of variance for making these sorts of decisions.

We established these values using methods that align with and compliment emerging M&V techniques. Our approach generated rapid, reproducible insights from traditional M&V methods and makes them accessible via a web app. We also collaborated with vendors and other stakeholders monitoring M&V 2.0 results to reduce labor costs and customer burden over the course of the evaluation.

Looking forward, the results presented in this study could help determine how best to distribute lighting loggers across a range of business types to train, test, and validate deployed M&V 2.0 models. In so doing, they can help solidify a broader decision-making framework for program administrators and evaluators to use when optimizing for cost, statistical rigor, and/or resource allocation of metering equipment at the outset of an evaluation. This would be a framework that accounts for the maturity of M&V 2.0 deployment in a given program territory and with specific business types. It would also aim to ensure rapid evaluation feedback regardless of the underlying M&V approach(es). Our experience underscores that, though achievable, realizing this vision of the future requires that evaluators continue identifying and extending to existing M&V methods some of the advantages of emerging tools and techniques.

[6] This is as opposed to the present study, which developed out of a comparative analysis of traditional M&V with M&V 2.0 approaches.

References

Barclay, D., J. Dolengo, S. Walker, and J. Ricardo. 2018. “Into the Great Wide Open: A Comparison of M&V 2.0 and Traditional Evaluation Methods for a Small Business Direct Install Program,” *ACEEE Summer Study on Energy Efficiency in Buildings.* Washington, DC: ACEEE.

Dobrowsky, P., I. Wainstein, and P. Hewlett. 2017. “Reduce, Reuse, Recycle, Rethink – Continuously Collecting M&V Data to Redefine the Future of Evaluation,” International Energy Program Evaluation Conference.

KEMA. 2014. *Impact Evaluation of New York Small Business Services Energy Efficiency Program, 2010 and 2011 Program Years*. Albany: New York Public Service Commission.

National Renewable Energy Laboratory. 2017. *The Uniform Methods Project: Sample Design Cross-Cutting Protocol*. https://www.nrel.gov/docs/fy17osti/68567.pdf.

New York State Joint Utilities. 2019. *New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures*. Version 6.

Peng, R. 2011. “Reproducible Research in Computational Science,” *Science*, 334, 1226-1227. New York, NY. doi:10.1126/science.1213847

Powanda, R., H. Lisle, J. Spencer, and J. Rivas. 2015. “Getting Over the Hump: Leveraging Multi-Year Site-Specific Impact Evaluation to Derive C&I Lighting Parameters,” International Energy Program Evaluation Conference.

Ricardo, J., J. Dolengo, D. Barclay, and S. Walker. 2018. “Time to Move On: An Examination of Metering Periods for Small Business Direct Install Participants,” *ACEEE Summer Study on Energy Efficiency in Buildings.* Washington, DC: ACEEE.

U. S. Department of Energy. 2017. *The Uniform Methods Project: Commercial and Industrial Lighting Evaluation Protocol*. https://www.nrel.gov/docs/fy17osti/68558.pdf.

Wickham, H. 2014. “Tidy Data,” *Journal of Statistical Software*, 59, 1-23. Los Angeles, CA: Foundation for Open Access Statistics.