Quality is Job #1

Bill Saxonis, New York State Department of Public Service, Albany NY[1]

Abstract

This paper focuses on the importance of quality evaluation to the future of energy efficiency programs and offers eight lessons learned for improving evaluation quality based on the experience of reviewing hundreds of evaluation reports. The mission of the paper is not to critique, rank or debate the performance of specific Program Administrators, evaluation studies or evaluation firms, but rather to provide a holistic view of our cumulative experiences emphasizing lessons learned for improving evaluation that will find wide applicability to both the public and private sector energy efficiency community.

Introduction

Why is Quality Job #1?

A key to successful evaluation is credibility. Can the results be trusted? Answering this question is more important than ever as evaluation feeds intelligence critical to addressing essential components of a sound energy policy including strategic planning, regulatory requirements, and documenting energy program effectiveness. Over the years, evaluation results were primarily targeted at energy efficiency Program Administrators and state policy makers, but recently we have seen increasing interest from the Federal Government resulting from the Environmental Protection Agency's (EPA) effort to regulate carbon dioxide emissions under section 111(d) of the Clean Air Act. The Department of Energy (DOE), for example, has funded several studies highlighting best evaluation practices and is currently considering a program to certify evaluator competency.[2]

Spending on electric and gas energy efficiency programs by U.S. Program Administrators has increased from $4.6 billion in 2009 to $7.2 billion in 2013, an increase of about 55 percent over just four years (CEE, 2015). Despite this strong financial commitment, the bottom line is that without credible evaluation results the future of energy efficiency programs will be jeopardized. Energy programs must prove themselves over and over again or they will see their support erode among policy makers, consumers and energy utilities. We have already observed that as spending increased, energy programs presented a growing target for critics. Policy makers in several states (e.g., Ohio, Michigan, Illinois, Indiana) have advocated, sometimes successfully, for spending reductions including the elimination of cost effective program portfolios.

In response to the challenge of keeping evaluations credible, the evaluation community has worked tirelessly over four decades to improve the accuracy and quality of not only the quantification of energy savings attributable to specific programs, but to conceptually broader research associated with probing consumer behavior, program design and market characterization. Because most evaluation is

---

[1] This paper reflects the views of the author and does not necessarily, implicitly or explicitly, express the views of the New York State Public Service Commission or the Department of Public Service.

[2] The US Department of Energy is developing a set of protocols for determining savings from energy efficiency measures and programs called the Uniform Methods Project. For more information see: http://energy.gov/eere/about-us/ump-home

conducted at the request of a state regulatory commission or other government entity, evaluation reports are widely available on the internet to both the general public and the research community worldwide. This convenient access to millions of dollars of research helps drive improvements in evaluation quality by sharing results that can help to highlight best evaluation practices, identify program trends and provide benchmarks for comparing results from similar programs. In New York, not only are energy program evaluation reports made public, but in some cases even "raw" non-confidential evaluation data may be available through the "open.ny.gov" initiative introduced by Governor Andrew Cuomo in March 2013 (see: https://data.ny.gov/). Collaborative meetings, where interested parties can freely discuss and share ideas on program and program evaluation issues, have also proven to be a useful tool.

Technological improvements have played a prominent role in the quest for quality as evaluators' typewriters and calculators were replaced with compact, powerful computers and smart phones that are continuing to advance at breakneck speed. Moreover, technological advances in homes and businesses, such as smart meters and Wi-Fi enabled thermostats, are contributing to improved evaluation accuracy and offer potential for future improvements including reductions in data acquisition costs and improved timeliness of results.

Despite growing experience, improving methodologies and technological advances, a basic reality has remained constant-- evaluation is a delicate blend of art and science. Human attributes such as objective analytical judgment, critical thinking, effective communication skills and integrity are still key ingredients in a successful evaluation. Simply stated, there are two overarching quality challenges that must be addressed—design and execution. Design focuses on the tools employed in the research (e.g., phone surveys, analytical software, metering approaches, billing analysis), while execution focuses on how well those tools are utilized.

This paper targets evaluation execution, especially on the more subjective elements associated with the art of evaluation. The foundation for this paper is based on the oversight and reviews of hundreds of evaluation plans and reports (e.g., process, impact, market characteristic) primarily in New York from about 2008 to early 2015. In New York, regulatory staff and their technical consultants play a prominent role in reviewing evaluation plans, key evaluation components (e.g., analytical methodology, sample designs, survey instruments) and final reports. The experiences reflected in this paper generally coincide with New York State's adoption of one of the more ambitious energy efficiency initiatives in the nation, the Energy Efficiency Portfolio Standard (EEPS), with a goal of reducing electricity usage by 15% from the level forecast for 2015 and a comparable efficiency goal for natural gas. This New York based research is supplemented with feedback from interviews with leaders in the evaluation community nationwide representing extensive experience in reviewing and approving evaluation reports (i.e., consultants and regulators).

The mission of this paper is not to critique, rank or debate the performance of specific Program Administrators, evaluation studies or evaluation firms, but rather to provide a holistic view of our cumulative experiences emphasizing lessons learned for improving evaluation that will find wide applicability to both the public and private sector energy efficiency community.[3]

While the quality of the evaluation projects my team and I reviewed has generally been good, it is still not unusual to uncover quality shortfalls with the potential to result in less reliable results.[4] The

[3] In order to adhere to the holistic view, and to offer a candid discussion, we have not always indentified specific studies, evaluators or sources. In cases where we quote from a publically available source, the citation is provided.

[4] Department of Public Service evaluation staff worked with a consultant team headed by TecMarket Works.

majority of the evaluation materials we have reviewed needed to be sent back to the authors to correct a range of defects ranging from minor issues that had little, if any, impact on the results to major issues that significantly influenced the energy savings estimates.

Shortfalls in evaluation quality can take many forms ranging from simple deficiencies, such as typographical errors, to more complex issues associated with, for example, flawed evaluation design, low response rates, non-response bias and estimation errors. We once found the transcript of a telephone order for school supplies imbedded in the text of a report. Fortunately, this type of discovery is rare, but errors such as mislabeled charts and page numbers are fairly common. Of course, simple errors can take on significance beyond their actual magnitude. If a report's page numbers are out of sequence, a skeptic might ask, "if they can't get the page numbers right, how can we trust the results?"

We have uncovered some big problems too, like an error riddled formula producing an inflated realization rate of about 200 percent and spillover estimates based on questionable assumptions that inflated program savings by well over 50 percent. The problems did not always accrue to the program's benefit. For example, one impact evaluation estimated net energy savings by utilizing a control group and also surveyed program participants to identify the level of free ridership in an effort to add more depth to the research. Mistakenly, the evaluator subtracted the free ridership savings from the control group results. While both research components were well executed, the methodological error resulted in a double deduction of savings attributable to free riders. Fortunately, through our review process, we were able to fix all, or at least most, of these types of problems before the release of the final reports.

The quality concerns identified in New York are certainly not unique. Discussions with the members of California's Master Evaluation Contractor Team indicate similar experiences associated with California studies. Likewise, discussions with some of the evaluation managers overseeing the evaluations of national energy programs identified similar concerns.

There is no simple formula for creating the perfect evaluation and there will be errors as long as human beings are involved in the evaluation process. While perfection may never be achievable, we have developed eight specific lessons learned for maintaining and improving evaluation quality that have wide applicability to a diverse array of interested parties including Program Administrators, regulators, policy makers and evaluators.

Eight Lessons

Lesson One - Maintain a Strong Review Process

The key to a strong evaluation effort is having an active, independent and transparent review process utilizing people with solid evaluation credentials to serve both as a quality control check and as a catalyst for ideas for future improvement. In New York, the review process begins with examining evaluation plans and key evaluation components (e.g., survey instruments, proposed sample design) to identify problems early and to synchronize evaluation goals with the program's performance objectives. Regulatory staff plays a major role in this effort, but the Program Administrators are also active partners in promoting quality by sharing in the review process.

As one utility evaluation contractor pointed out, while most states do not examine the components of an evaluation with the same level of detail as New York, New York's reviews have notably enhanced the evaluation products, not only by identifying flaws, but also by offering practical ways to enhance the research. The reviews were consistently performed with a high degree of collegiality and with a constant dedication to the goal of quality evaluation. While occasionally there

would be animated, but constructive, discussions of technical details, the overall level of cooperation has been and remains high. Program Administrators frequently expressed appreciation for a fair and unbiased review. A less obvious benefit of the review process is the concept of the "invisible hand." One evaluator noted that it often takes longer to prepare New York evaluation products because of the recognition their work will need to stand up to rigorous scrutiny. In other words, the review process serves as an incentive to improve quality long before the final product is delivered.

New York's Evaluation Guidelines, developed as part of the EEPS effort, have proven invaluable in guiding the review process. While they don't offer a road map for evaluating every possible program type, they do provide the basics and a framework to guide evaluation quality. For example, they include a quality checklist to help focus both the evaluation plan development and the review process (DPS, 2013). The checklist includes over 30 items for review (e.g., baseline setting approach, adequacy of data sources).

A similar review process is successfully used by regulatory staff in several other states recognized as leaders in energy efficiency programs including California and Massachusetts. Many states, however, are relatively new to energy efficiency programs and their Program Administrators may have had insufficient time or resources to fully develop an effective evaluation oversight infrastructure. In some cases regulatory bodies lack the financial resources to provide an adequate level of evaluation oversight, or staff sufficiently experienced in the art and science of evaluation. Not surprisingly, a representative of a company associated with energy audits noted that his visit to New York was refreshing because he found regulatory staff that understood evaluation terminology and the principles of effective evaluation.

The role of an evaluation quality gatekeeper can be performed outside of the regulatory process. There are many possibilities including a Program Administrator hiring an independent reviewer; creating a review committee of interested parties and evaluation experts; hiring an independent evaluation professional to oversee the evaluation efforts of work performed by others or some combination of these approaches. Not surprisingly, a utility executive with a major southern based utility explained to me the importance of utilities having an EM&V expert in house or under contract.

Lesson Two – Good Communication is Essential

One method of improving evaluation quality doesn't necessarily require actually improving the evaluation methods or data quality; it simply requires effectively communicating the parameters of what the evaluation offers and balancing need with reality. In some cases, the evaluation could prove methodologically sound, but the content may not align with the needs and expectations of the target audience. The study needs to offer the right level of rigor, at the right time, and in an understandable and useful format. Failing to meet these criteria, a technically sound evaluation may receive a frosty reception, by answering questions not asked and not adequately answering the questions asked.

A key to effective communication is to begin the outreach long before the release of the final report. A detailed evaluation plan and a project kick off meeting with the evaluation team and other interested parties are essential. It is also critical to maintain an ongoing dialogue because what may seem possible at a kick off meeting may prove to be impossible in actuality. However, these steps do not always prove adequate because some key players (e.g., regulatory commissioners, senior utility managers, public interest groups) are often unable to engage at this early stage of the evaluation process. Moreover, while a detailed evaluation plan and an effective kick off meeting can be expected to adequately address the goals and the methodology of the evaluation, they sometimes fall short in

capturing the potential limitations of the study and setting realistic expectations. Evaluation doesn't always provide a rock solid answer, especially when dealing with counter factual situations and behavioral related questions. Can we ever be 100 percent sure what a customer would have done if the program's service was unavailable? Readers may not appreciate that free rider and spillover estimates will likely be more uncertain and more prone to error than estimating the delta watts of replacing a 60 watt bulb that runs about 800 hours per year with a 12 watt bulb. They may question the report's quality if free ridership is couched with some uncertainty, but some uncertainty is reasonable and to be expected.

Research is an evolving process in almost all fields including education, health, and social science. Energy efficiency is ***not*** an exception. The Washington Post recently reported in some detail about years of research costing millions of dollars still leaving considerable doubt surrounding the appropriate daily intake of salt as part of a healthy diet. Evidence is now showing that salt may not be as bad for the average person as suggested by earlier studies (Washington Post, 2015). Scholars still debate through dueling research papers if genetics or the environment is the greater influence on human behavior. The lesson is that we should not expect energy evaluations to have definitive answers to every question. It is also important to place evaluation results in context and not overreact to findings, especially preliminary findings. This is not to make excuses for evaluators, but to add a dose of reality and further highlight the need for articulating realistic expectations.

Others may be disappointed that evaluation results are typically not real time and tend to reflect more of a historical perspective. It is a common frustration, but it is important to realize that data collection and analysis take time, often a minimum of 12-18 months of data for some types of analysis (e.g., billing analysis). Moreover, for new programs there may not be enough critical mass to conduct a creditable evaluation until long after the program kicks off. The time required to conduct a process evaluation is usually far less than a rigorous impact study, but still can take months because of the many steps involved (e.g., sampling, data collection, data analysis, report preparation). In the future, technology, such as smart meters, may significantly increase the wide scale potential for real time evaluation results, but regardless of the approach, it is essential that the timing realities be effectively communicated.

A strong case can be made that simply making evaluation reports clearer and better targeted to the intended audience would go a long way to improving the value of evaluations. Admittedly, there can be a difficult balance of promoting clarity, but also offering enough detail to fully explain the data and the results. Terms and acronyms commonly used in evaluations are often unfamiliar to many readers. Terms such as NTG Standard Error, NTG Relative Precision, Interaction Adjustment Factor, and UDC (unit demand consumption) can be daunting and, to add to the confusion, some definitions are not always consistent from region to region.

In the fall of 2014, the New York State Energy Research and Development Authority (NYSERDA) introduced an initiative to streamline their evaluation reports. The recommendations began with a simple, but astute conclusion, "a redesign of NYSERDA's reports should begin with a clear understanding of the audiences and their needs." Another key recommendation was to place the more technical elements in appendices and attempt to have the main body of the report "tell a story and

flow in a narrative manner." Recent NYSERDA reports are now considerably more compact and readable.[5]

Lesson Three – Evaluation Budgets Need to be Prioritized

While there is not a perfect correlation, generally better-funded studies, with larger samples and more sophisticated research methods, result in stronger, more reliable evaluation results. A study of energy program evaluations in the fifty states concluded that the quality varied widely. In some cases this was driven by varying objectives for evaluation, but also funding levels and technical resources (Schiller, 2013). Not surprisingly, a common refrain from evaluation contractors and Program Administrators is "we would have liked to have done more, but we ran out of money," or "yes additional analysis makes sense, but it is beyond the scope of our contract."

Under EEPS, Program Administrators were provided an evaluation budget equal to five percent of their total energy portfolio budget, and afforded the flexibility to spend above or below the five percent level for specific programs depending on need. By national standards, the 5 percent funding level is generous, reflecting the New York State Public Service Commission's (Commission) goal of rigorous evaluation for EEPS programs. Nationally, evaluation spending has averaged around 2 percent (CEE, 2015); a level identified by many experienced professionals as being insufficient to fund the level of accuracy likely to be needed under the anticipated 111(d) Clean Air Act requirements.

A key objective is to design a rational budget strategy targeting funds toward programs with the most need and, where practical, combining the resources of several Program Administrators to conduct evaluations on a joint or statewide basis. New York's policy of flexible funding within the 5 percent cap resulted in some evaluations with small budgets, well under 2 percent, and others with budgets greater than 10 percent of program costs. This funding strategy is based on the principle that different programs have different evaluation needs. Programs that contribute a high percent of energy savings to the overall portfolio may need more evaluation than smaller programs and programs featuring simple, common measures may require less rigorous and less frequent evaluation compared to programs offering new and emerging technologies. We potentially saved millions of evaluation dollars by pooling the resources of seven utilities to conduct a statewide residential gas HVAC impact study and securing joint funding for large scope studies (e.g., residential appliance load shapes).

The cost for key evaluation services can vary widely. The price for an on-site audit of a large commercial building in New York City is very different from the cost of implementing a five minute phone survey to a homeowner in Albany. It is impossible to estimate the cost of an evaluation without fully understanding the research objectives. While New York's Evaluation Guidelines do not provide specific budget guidance, they do highlight cases where the highest standards of accuracy and the greatest frequency of evaluation would typically be directed. They include:

- ✓ Providing expensive infrastructure investments
- ✓ Eligible for utility incentive payments or lost revenue recovery
- ✓ Targeted for a significant budget increase
- ✓ Producing results far above or below expectations
- ✓ Implemented as an innovative program on a pilot basis
- ✓ Containing measures with high energy savings variability

[5] NYSERDA evaluation reports can be found here:http://www.nyserda.ny.gov/About/Publications/Program-Planning-Status-and-Evaluation-Reports/Evaluation-Contractor-Reports

- ✓ Based on a limited existing knowledge base
- ✓ Making large contributions to the overall portfolio savings.

Lesson Four - Present Evaluations in Comparable Terms and Compare Results

Sometimes evaluations are viewed in a silo. The results of a specific program evaluation may serve the specific need of the sponsor, but the results may never be carefully compared to similar programs both in and out of state. This limited viewing can potentially retard valuable learning and insights. Based on our experience, it is clear that evaluations, at least to the extent possible, need to be presented in comparable terms and actually compared.[6]

For example, in New York several utilities operated programs designed to help the small business sector reduce their energy consumption (e.g., Small Business Direct Install Program). The results from five utility impact studies, of essentially the same program, showed realization rates ranging from 58 percent to 80 percent. Over 90 percent of the measures were lighting or lighting related. Representatives of the utilities, regulatory staff and NYSERDA formed a team to pinpoint the reasons for the variances. Were the differences related to program administration? The evaluation design? Both?

After reviewing the data in detail and holding several meetings, the team concluded that the differences in annual lighting operating hours and facility type definitions were responsible for the greatest variability. Some utilities were using hours reported on-site, others used default values and still others used both approaches. Also facility type characteristics were not always consistent from utility to utility. The team recommended the use of a standardized data collection protocol for capturing equipment characteristics, lighting operating hours, and building attributes and use patterns. They also recommended a quality control process for both implementers and evaluation contractors to review the reasonableness of the operating hours reported on-site and to follow up when the results do not appear reasonable (DPS, 2015). The results of this effort will also likely result in updates to New York's Technical Resource Manual (DPS, 2014). Without the collaborative process and collective review of the results, the team's recommendation to improve future evaluations and program implementation may not have been fully realized. This sounds like a simple lesson, but with the demands on staff resources, limited evaluation funding and often inconsistent data quality, making it happen is not always easy. For example, in 1993, the New York Power Pool retained independent consultants to develop a standard reporting format for evaluation reports for energy efficiency programs operated by the New York State utilities. The consultants found that "PSC staff, as well as a number of utility staff, expressed frustration in trying to compare the results of similar programs among the utilities. The data was inconsistent in completeness, format, and definition" (Barakat & Chamberlin, 1993).

Lesson Five - Actionable Recommendations are a Must

In New York there has been a strong emphasis on evaluation reports, both impact and process, producing actionable recommendations. Our objective is to have the evaluation data serve as a vital tool in a process of continual program improvement and not simply live in an academic report that collects dust on a bookcase. All the recommendations in the evaluations become public including the Program

[6] In 2014, the Northeast Energy Efficiency Partnerships' EM&V Forum developed standardized forms to document evaluation results and practices across multiple states. See: http://www.neep.org/initiatives/emv-forum/model-emv-methods-standardized-reporting-forms New York's Evaluation Guidelines adopted the Forum's glossary of EM&V terms. See: http://www.neep.org/emv-glossary

Administrator's responses and planned actions (e.g., implement, partially implement, reject). Since EEPS evaluation reports started rolling in around 2010, they have generated approximately 500 recommendations, with a large majority of the recommendations being implemented. It would be impossible to quantify the value of the recommendations in terms of program savings, especially considering the difficulty of determining the number of changes that would have occurred without the evaluation (i.e., free rider recommendations). However, even a saving of just a few percent of the total program budgets could translate into millions of dollars of savings to the ratepayers and potentially more than offset the cost of the evaluation efforts.

The recommendations were not limited to program improvement, but also included evaluation. For example, a NYSERDA impact study of the EmPower program designed to assist low–income consumers had three recommendations for improving the program, and four recommendations for improving future evaluation efforts (NYSERDA, April 2012). Evaluation should not be exempt from on-going improvement. After spending months of intense work on an evaluation project, the evaluation team should be in a good position to uncover ways of improving the evaluation process. Evaluators need to be sensitive to this need.

Lesson Six – Keep Pace with a Changing Environment

New York State is undergoing an extensive reexamination of both how it regulates the electric and gas utilities and the utility business model. This effort is driven by several factors including the increasing competitiveness of renewable energy resources, an aging energy transmission infrastructure, climate change, poor system efficiency, security concerns and rapid changes in technology impacting both the consumer and the distribution network (e.g., smart meters, energy storage).

To meet this challenge, the Commission commenced its "Reforming the Energy Vision" (REV) initiative to update New York State's energy industry and regulatory practices. The Commission determined that the response to these challenges is not to cling to traditional models, but to identify and build regulatory, utility and market models that create new value for consumers and encourage competitive markets including wider deployment of "distributed" energy resources (e.g., micro grids, on-site power supplies, energy storage) and promote greater use of advanced energy management products to enhance demand elasticity and efficiencies (e.g., demand response programs, real time pricing). The Commission's goal is "to embrace the changes that are shaking the traditional system and turn them into New York's economic and environmental advantage" (Case 14-M-0101).

A fundamental question is what do we need from evaluation under this emerging rubric? Will evaluation be able to produce the necessary data at the right time? The most prominent deficiency we experienced under the EPPS evaluation strategy was the timeliness of the studies, a problem exacerbated by several factors including the time required for several Program Administrators to establish an evaluation infrastructure (e.g., hire evaluation staff, secure contractors), and delays in program implementation because of economic conditions and start-up issues. Moreover, evaluators, regulatory staff and Program Administrators needed time to get accustomed to the EEPS policies and procedures and, admittedly, there were some learning pains. This will likely be a factor as the regulatory framework in New York continues to evolve under REV.

Under EEPS the evaluation objectives usually focused on quantifying energy savings, developing actionable recommendations, and providing insight for updating the Technical Resource Manual. A major push under REV is to place evaluation reports on a schedule to synchronize the results with three-year program cycles. Plans will be developed to establish an annual date by which evaluation

studies of programs in previous cycles will be filed in order to inform overall program design and operation and update the Technical Resource Manual (Case 14-M-0101, Appendix C).

The next two lessons touch upon insights that might help to inform and align evaluation with this emerging energy vision.

Lesson Seven - Embrace Best Practices

Back in the eighties and nineties management techniques such as Total Quality Management (TQM), Six Sigma, and Business Process Management (BPM) received a great deal of attention. In 1992, the New York State Energy Office, Division of Energy Services, adopted TQM principles and the Evaluation Team implemented Evaluation Quality Principles (EQP). The principles were the subject of a paper I presented at the 1993 IEPEC conference and 22 years later they still make sense (Saxonis, 1993). The key principles included:

- ✓ Know the target audience—explore their needs, desires and limitations in depth
- ✓ Know thyself: evaluate the evaluators—examine and address the evaluation team's strengths and weaknesses in meeting the needs of the target audiences
- ✓ Know the competition—learn from the best; benchmark
- ✓ Commit to quality evaluation- engage everyone in the organization from the senior executives to the rank and file
- ✓ Set quality metrics for evaluation- factors could include timeliness, report readability, report quality (e.g., minimal redrafting)

Not all of these principles were formally implemented under EEPS, but we did adopt a philosophy of "continual improvement" and, as a result, captured, at least in spirit, TQM principles. Under REV the Commission has requested the development of an Energy Efficiency Best Practices Guide "to ensure shared learning and the evolution of programs across service territories outlining best practices under a REV framework." While not specifically mentioned in the Commission directive, a similar effort should be considered for evaluation practices.

There is also value to creating a "lessons learned bank" where evaluators would share evaluation successes and failures. The topics would typically focus on issues not so major to merit a formal recommendation in an evaluation report, but still useful to the evaluation community. It could be identifying a survey format that flopped (e.g., surveys of over 20 minutes targeting certain sectors failed because of their length) or a better approach for estimating free ridership. This wouldn't necessarily be proprietary information because it would be derived from public reports. The major advantages are improved accessibility and increased transparency. This idea has not been formally implemented in New York.

Lesson Eight - Never Stop Improving

Evaluation Guidelines are a must and must be dynamic. During a period of about two years, New York's Evaluation Guidelines grew with four new sections created largely in response to concerns derived from our reviews of evaluation plans and reports. The new sections included guidance on sampling and uncertainty, spillover rigor levels, net to gross estimation using self-reports and calculating the relative precision of program net saving estimates. Our evaluation review process proved to be an effective tool for highlighting problems, creating opportunities for discussion and providing solutions (DPS, 2013).

We found examples of surveys asking net to gross questions of program participants a year or more after their program-associated action by simply asking if the measure would have been implemented without program assistance. The surveys failed to include questions to capture the importance of the motivations and the sequence of events surrounding the decision to take the energy efficiency related action. Evidence suggests that this approach could introduce bias, not only because of the period of time between the key events, but also because it failed to "set the mind of the respondent into the train of events that led to the installation." We strongly recommended adding "set up" questions.

Another issue revolved around spillover, especially as spillover savings were becoming significant, at least in a few programs. However, we lacked standards or guidance on how spillover was to be measured and what level of rigor would be acceptable and under what circumstances. We saw examples of evaluations with rigorous analysis of savings resulting from program funded measures combined with claims of large savings from spillover based on tenuous assumptions. This imbalance in the rigor level did not make sense. In response we added to our Evaluation Guidelines two basic levels of rigor for spillover measurement—standard and enhanced—and provided detail on selecting the appropriate rigor level (DPS, 2013).

An increasingly prominent problem is the potential for sample bias or response bias in almost any type of survey, whether targeting the American voter (e.g., surveys of the presidential race), or a highly targeted audience (e.g., building code officials in mid-sized communities). The challenge is finding people willing to spend time answering questions in response to a request from a complete stranger. In 1997, survey firms were usually able to connect with an adult in the targeted households 90 percent of the time, but the percentage dropped to 62 percent by 2012. The percent of households that actually yielded an interview dropped from 36 percent to 9 percent during the same period (Pew Research). The CEO of a major evaluation firm explained that their survey contact completion rate for some types of residential programs has declined from 68 percent to 11 percent over the last 25 years.

This issue is more of a problem when the population is small, as is common in energy efficiency programs. If a program only has 100 participants, finding a statistically valid and unbiased sample can be challenging and it is not unusual to see survey efforts fail to meet relatively modest survey response goals. Highly targeted survey efforts have proven even more challenging. We have observed examples where evaluators set out to survey some of the major product distributors, installers, and manufacturers and were forced to abandon the effort because of a poor response rate.

The good news is that a low response rate doesn't necessarily guarantee the sample is biased. The bad news is that a large sample size doesn't guarantee a lack of bias.

There is no simple solution to the non-response dilemma. The combination of the demands on people's time, hassle of unsolicited phone calls, private telephone numbers, and the wide use of call screening technology (e.g., voice mail, caller identification, call blocking) all contribute to the trend of declining response rates. What is clear is that there needs to be a heightened sensitivity to this problem, including increased emphasis on methods for testing for non-response bias.

For example, as part of a baseline study, we dug into survey results designed to be reflective of the New York population statewide, and found that the education level of the respondents appeared high, with over 40 percent claiming advanced degrees (i.e., graduate level). According to US Census data only about 14 percent of the New York's population has an advanced college degree. Considering that research suggests high income and education levels are often associated with higher energy

efficiency, it appeared we had a sample bias problem. A simple solution would be to weight the sample by education levels more in-line with the Census data. Upon further digging, we realized that the key survey question was presented in a format different from the Census. The evaluation survey asked if ***anyone*** in the household had an advanced degree, whereas the Census data estimate is based on a percent of the state's entire population. In other words, a household of six would have potentially had six chances of having someone with an advanced degree. This leaves us with the dilemma of determining the correct value suspecting that it is somewhere between 14 and 40 percent. Clearly, the study would have benefited by more comparable data collection approaches.

As we move into the future, the Commission observed, "as REV recognizes the pace of technology and its ability to redefine our electric system, so too can advances in technology be used to challenge and enhance our traditional approach to EM&V. Current evaluation guidelines… should be reviewed and revised." We agree. The exact changes are difficult to predict, but it is clear evaluators, program implementers and regulators will need to continue to be nimble to keep pace with the changing regulatory policies, technological advances and utility business models.

Conclusions

Evaluation has changed a great deal over the years, mostly for the better. On the other hand, the hot topics from decades ago (e.g., free ridership, quantifying market transformation, sample bias) are still hot topics today and will likely continue to challenge evaluators for the foreseeable future.

Compared to the normally reliable modern electric and gas meters, which trace their history back to the eighteen hundreds, energy program evaluation admittedly provides less certainty in measuring net energy consumption. Of course, the meter tracking the energy usage in a home or business never was expected to identify if the sale of a specific energy efficient light bulb would have ever happened if not for a rebate program or if the savings resulting from a new energy efficient air conditioner encouraged the home owner to keep their home cooler.

The intent of this paper was not to be critical of evaluators, or the evaluation industry, but simply to point out that working together, the utilities, government agencies and evaluators, must maintain a dedication to the goal of keeping quality job #1. As an industry, energy program evaluation is still young and still advancing. Winston Churchill once remarked that "to improve is to change, so to be perfect is to have changed often" We may never reach perfection, but we will change and we will improve. A clear pathway to better evaluation is to have a meaningful review process covering key evaluation components from the initial plan to the final report. These reviews are critical to guard against everything from typos to faulty formulas as well as serving as a catalyst to push evaluation to higher levels of accuracy. An added bonus is reviews, especially early in the process, can help to identify and correct problems before they happen. Regulators and ratepayers will appreciate this bonus. Not losing sight of the importance of good communication, sound budget priorities, using best practices and taking steps to respond to a changing environment are all critical to good evaluation. Our final conclusion is to repeat the first words of this paper, "a key to successful evaluation is credibility." In order to have credibility you must have quality.

References

Barakat & Chamberlin, "*Standard Reporting Format for Annual DSM Program Evaluation Results,*" submitted to the Program Evaluation Task Force, New York Power Pool, Albany NY, June 28, 1993.

Consortium for Energy Efficiency (CEE), "*2014 State of the Efficiency Program Industry,*" May 2015, Boston MA. http://library.cee1.org/content/2014-state-efficiency-program-industry

Messenger Mike, Ranjit Bharvirkar, Bill Golemboski, Charles Goldman, Steven Schiller, "*Review of Evaluation Measurement and Verification Approaches Used to Estimate the Load Impacts and Effectiveness of Energy Efficiency Programs,*" Ernest Orlando Lawrence Berkeley National laboratory, LBNL-3277E, April 2010, Berkeley CA.

New York State Energy Research and Development Authority, "*2007-2008 Empower New York Program Impact Evaluation Report,"* April 2012, Albany NY.

New York State Department of Public Service, "*New York Evaluation Plan Guidance for EEPS Program Administrators,"* August 2013, Albany, NY.

New York State Department of Public Service, "*Final Report of the Small Commercial EM&V Review Subcommittee,"* April 2015, Albany NY.

New York State Department of Public Service, "*New York Standard Approach for Estimating Energy Savings - Residential, Multi-Family and Commercial/Industrial Measures, Technical Manual," Version 2*, December 2014, Albany, NY.

New York State Public Service Commission, Case 07-M-0548, "*Order Establishing Energy Efficiency Portfolio Standard and Approving Programs,"* June 23, 2008, Albany, NY.

New York State Public Service Commission, Case 07-M-0548, "*Proceeding on Motion of the Commission Regarding an Energy Efficiency Portfolio Standard,*" December 21, 2010. Albany, NY.

New York State Public Service Commission, Case 14-M-0101, *"Order Adopting Regulatory Policy Framework and Implementation Plan"* February 26, 2015, Albany NY.

Pew Research Center, *"Assessing the Representativeness of Public Opinion Surveys," Washington DC,* May 2012. http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys/

Saxonis, William *"Evaluation and Regulation: Put to the Test," Proceedings of the 2011 International Energy Program Evaluation Conference,* Boston MA.

Saxonis, William, "A New Evaluation Paradigm for a New Generation of Energy Efficiency Programs" *Proceedings of the 2009 International Energy Program Evaluation Conference,* Portland OR.

Saxonis, William, *"A Tool for Better Evaluation: Total Quality Management," Proceedings of the 1993 International Energy Program Evaluation Conference, Chicago, IL.*

Schiller, Steven, Charles Goldman, *"Developing State and National Evaluation Infrastructure-Guidance for Challenges and Opportunities of EM&V," Proceedings of the 2013 International Energy Program Evaluation Conference, Chicago, IL.*

Washington Post, *More Scientists Doubt Salt is as Bad for You as the Government Says,"* April 6, 2015.