NEW YORK STATE OF OPPORTUNITY

Department of Public Service

Public Service Commission
Audrey Zibelman
Chair

Patricia L. Acampora
Gregg C. Sayre
Diane X. Burman
Commissioners

Paul Agresta
General Counsel
Kathleen H. Burgess
Secretary

Three Empire State Plaza, Albany, NY 12223-1350
www.dps.ny.gov

November 1, 2016

Ms. Kathleen Burgess, Secretary
New York State Public Service Commission
3 Empire State Plaza
Albany, NY 12223-1350

RE: Case 14-M-0094 – Proceeding on Motion of the Commission to Consider a Clean Energy Fund.
Case 15-M-0252 – In the Matter of Utility Energy Efficiency Programs

Dear Secretary Burgess:

In its order approving the Clean Energy Fund (CEF Order),[1] the Public Service Commission (Commission) directed Department of Public Service staff (Staff), in consultation with the Clean Energy Advisory Council (CEAC), to conduct a review of "The New York Evaluation Plan Guidance for EEPS Program Administrators" (Evaluation Guidelines) and reissue revised guidelines by November 1, 2016.

In compliance with this requirement, Staff worked with the CEAC, which directed its Metrics, Tracking and Performance Assessment (MTPA) Working Group to conduct the requisite review of the Evaluation Guidelines. Based on this review, the MTPA Working Group filed the Evaluation Guidelines Recommendations Report on October 3, 2016.[2] Staff has considered these recommendations in preparation of the revised evaluation guidelines.

In accordance with the CEF Order, the Evaluation, Measurement & Verification Guidance, enclosed herein, has been posted to the Commission's website and is effective immediately.

Sincerely,
Christina Palmero
Christina Palmero
Deputy Director,
Office of Clean Energy

Enc.

[1] Case 14-M-0094, Proceeding on Motion of the Commission to Consider a Clean Energy Fund, Order Approving the Clean Energy Fund Framework (Issued January 21, 2016).

[2] Matter 16-01008, In the Matter of the CEAC's Metrics Tracking and Performance Assessment Working Group, Evaluation Guidelines Recommendations Report.

OFFICE OF CLEAN ENERGY
CLEAN ENERGY GUIDANCE

Evaluation, Measurement & Verification Guidance

Version History Log:

Version	Date Issued	Approval	Changes
1.0	2016-11-01	Christina Palmero, Deputy Director	N/A

PURPOSE:

This Clean Energy Guidance document provides guidance to the utilities and the New York State Energy Research and Development Authority (NYSERDA), (collectively, the program administrators), and evaluators on the conduct of evaluation, measurement and verification (EM&V) activities associated with ratepayer funded clean energy programs.

All Clean Energy Guidance documents are in effect until revised, rescinded, or superseded.

Background:

The Department of Public Service initially issued Evaluation Guidelines to support implementation and oversight of Energy Efficiency Portfolio Standard (EEPS) programs. These guidelines, the "New York Evaluation Plan Guidance for Energy Efficiency Portfolio Standard Program Administrators" were developed with input from the former Evaluation Advisory Group. In the January 21, 2016 Clean Energy Fund Order[1] the Commission directed Staff, in consultation with the Clean Energy Advisory Council (CEAC), to conduct a review of the Evaluation Guidelines and reissue revised guidelines by November 1, 2016.

Transparency:

To further support transparency of EM&V activities and program administrators' accountability for quality EM&V work, all EM&V plans and final reports shall be filed in Matter 16-02180, In the Matter of Clean Energy EM&V within the Department of Public Service's Document and Matter Management (DMM) System. Use of this designated matter number will not only aid stakeholders and parties interested in tracking EM&V activities but will serve as a tool among program administrators to stay abreast of EM&V plans and results.

[1] Case 14-M-0094, Proceeding on Motion of the Commission to Consider a Clean Energy Fund, Order Authorizing the Clean Energy Fund Framework, issued January 21, 2016.

Contents

I. INTRODUCTION

The New York State Public Service Commission (Commission) has sustained a long-term commitment to energy efficiency as a crucial and cost-effective means of achieving New York State's clean energy goals. In directing Staff to issue Evaluation Guidelines to support transparent and accurate evaluations of the suite of rate-payer funded energy efficiency programs in 2008, the Commission observed that "Evaluation Guidelines are an important step in not only providing the elements of an acceptable evaluation plan but the standard to strengthen the accountability, accuracy, and usefulness of the evaluation results."[2] While the suite of energy efficiency programs and initiatives has and will continue to evolve, the importance of Evaluation, Measurement, and Verification (EM&V) in demonstrating the value and effectiveness of energy efficiency efforts remains.

Reforming the Energy Vision (REV) has set in motion new innovations designed to increase the deployment and utilization of distributed energy resources through utility and NYSERDA investment as well as new investments by energy consumers. This Guidance addresses best practices and common approaches to support consistency across program administrators in the implementation of EM&V activities targeted at energy efficiency initiatives conducted under the REV framework. REV places heightened emphasis on innovation designed to mobilize energy consumer investments that will contribute to energy and environmental goals. Accordingly, this Guidance contemplates the need to continue to adapt EM&V practices to address these innovations, providing for the evolution of best practice standards in response to changing program designs and market conditions. REV's clean energy initiatives are broader than energy efficiency and include the full breadth of market transformation and other Distributed Energy Resources (DER) (e.g. demand reduction, demand response, distributed storage, and distributed generation). This Guidance will need to continue to evolve to fully encompass technology; program and market needs associated with evolving market-based deployment approaches of energy efficiency and other forms of DER.

This Guidance builds upon the knowledge and experience acquired implementing and evaluating more than 100 energy efficiency programs administered by NYSERDA and the utilities under the EEPS portfolio. This experience reveals that it is important for EM&V to produce results that inform program and portfolio design in a timely manner in support of overall policy objectives.

REV is changing the use and role of DER with the goal of ensuring that DER, including energy efficiency, is fully valued and deployed in support of a clean and diverse energy system. Transparent and objective evaluation is important to demonstrating this value and ensuring effective and appropriate use of ratepayer funds. In this context it is important that EM&V activities strike a balance between the level of rigor and accuracy, with the cost, timeliness and usefulness of the information in a rapidly evolving market environment.

This Guidance encourages program administrators to consider advances in EM&V technology and methodologies in support of their EM&V objectives, as tools for improving timeliness and usefulness of the results. This Guidance also requires that EM&V resources be managed strategically, placing a high priority on providing timely information about program performance and the progress toward achievement of key policy objectives and statewide energy goals.

This Guidance recognizes EM&V activities serve two primary functions:

[2] Case 07-M-0548 – Proceeding on Motion of the Commission Regarding an Energy Efficiency Portfolio Standard, Order Establishing Energy Efficiency Portfolio Standard and Approving Programs, Issued June 23, 2008.

- Activities utilities and NYSERDA require to monitor and improve performance on a timely basis, which may place greater emphasis on M&V activities than in the past; and
- Activities policy makers and stakeholders require to document value and benefits in support of assessment of progress toward key policy objectives, assessment of energy demand forecasting and resource planning: and analysis and payment of future utility Earning Adjustment Mechanisms (EAMs).

This Guidance does not represent a rigid doctrine, but offers flexibility to program administrators, allowing quality EM&V work to be completed, using reliable, responsive, and cost effective approaches.

This Guidance is intended to be a 'living document' and therefore is expected to be reassessed and/or updated to remain relevant and responsive as the market and regulatory context evolves under REV.

II. EM&V PLANNING

PLANNING APPROACHES & PRIORITIES

Developing an EM&V plan is a critical component of clearly identifying the goals of the activity, the approach to be taken, the corresponding deliverables/timeline, and associated budget. The most efficient approach to effective EM&V planning is to consider program evaluation needs, as part of the program design process. Developing an initial EM&V plan in preparation for launching a program allows program evaluators to work with program administrators, to identify data collection needs. This can enable evaluation data collection concurrent with program or portfolio implementation, allowing for mechanisms to provide feedback to administrators on an on-going basis, and the establishment of upfront budget estimates and synchronization of EM&V goals with program performance goals. EM&V serves an equally important role as programs mature. Planning in that context should identify specific focus areas to be assessed taking into consideration implementation experience as well as previous EM&V results. EM&V activities are not restricted to comprehensive program (or portfolio) studies, rather meaningful EM&V activities may assess discrete subjects within a program or subjects that transcend individual programs.

Program administrators may also consider pooling resources to conduct EM&V activities on similar programs/subjects on a statewide or regional basis, or to consider collaborating in other ways to achieve EM&V goals while minimizing the cost to ratepayers. For example, often market and baseline assessments are most appropriately conducted at a statewide level through a collaborative approach between multiple entities. To avoid duplicative efforts, these types of initiatives are encouraged.

EM&V activities should be guided by the consistent use of key terms (See Appendix A) and methodologies enabling statewide sharing and analysis of results, and accurate tracking of statewide progress towards established goals and metrics.

In general, EM&V activities should be prioritized to address the largest information gaps within an energy efficiency program portfolio as well as to assess the accuracy of the tools that program administrators rely on to calculate reported energy savings (e.g., the Technical Resource Manual[3],

[3] New York Standard Approach for Estimating Efficiency Programs – Residential, Multi-Family, and Commercial/Industrial Measures.

modeling software, etc.). In particular, program administrators should place a high priority on EM&V activities that target those programs, measures, or technologies that:

- Defer expensive infrastructure investments
- Are eligible for utility EAMs
- Perform far above or below expectations
- Are implemented on a "test and learn" or pilot basis
- Have high energy savings variability
- Are based on a limited existing knowledge base
- Represent large contributions to the program administrator's overall portfolio savings

EM&V FREQUENCY

Early program EM&V efforts should focus on process-related issues to serve as an early warning system, especially for new initiatives. This approach can be used to determine if the program is operating smoothly and is responsive to participants/market needs as well as to identify opportunities to make improvements that can reduce costs and increase program effectiveness. This Guidance does not establish a rigid timetable for process evaluation and impact evaluations. Generally, evaluations focusing on verifying program-level energy savings cannot be completed until a sufficient number of projects have been completed and post-installation operations can be observed. A typical program evaluation timetable includes a process evaluation in program year 1, and an impact evaluation in program year 2 or 3, with an emphasis on obtaining results as soon as reasonably possible.

For programs that have undergone recent full-scale evaluation, repeating such full-scale evaluation may not be necessary every few years, but rather the frequency and scope of the EM&V activity should be based on results of evaluations to date, information gaps, and other areas requiring analysis.

Given the dynamic and transitional nature of current program offerings, shorter and more targeted EM&V activities can be valuable for investigating issues that a typical program evaluation would not cover. Targeted evaluations may be initiated and completed at any time and may serve to better align and serve the planning process for the current suite of clean energy programs.[4]

Concurrent evaluation is also an option that can accelerate M&V data collection, and is a credible technique for assessing decision-making closer to its time of occurrence while still vivid in the respondent's mind, and enables more expeditious feedback regarding program performance.

If an EM&V activity's schedule spans an extended period, (e.g., more than six months) the EM&V plan should include stages for interim feedback to the program administrator(s), in most cases. Presentation of interim results requires clear disclosure regarding the completeness of the data collection and analysis and precision, if calculated. In most cases, interim results are more appropriate for program administrator(s) internal use rather than public consumption.

TIMELY AND TARGETED EM&V

This EM&V Guidance recognizes new methodologies being employed to help inform and improve

[4] The February 26, 2015 Order Adopting Regulatory Policy Framework and Implementation Plan, Case 14-M-0101 Proceeding on Motion of the Commission in Regard to Reforming the Energy Vision, established the elements of a three-year rolling cycle for utility energy efficiency transition programs which, among other things, requires a date by which evaluation studies of programs in previous cycles shall be filed in order to inform overall program design, operations, and an updated TRM.

program offerings that produce more continuous and actionable results, than were possible under traditional "survey-based" EM&V efforts in the past. These methodologies should be considered at the EM&V planning stage to determine the benefits of undertaking such approaches and are potentially applicable across all types of EM&V activities, including impact, process, and market-based studies. They include the following:

- Real-time and rolling surveys that can continually update program related data from customers, implementers, retailers and trade allies. Real-time and rolling surveys embedded into the period being evaluated allow for periodic updates on opinions, reactions and decision making, which may allow for more expeditious feedback and therefore more nimble adjustments to program operations and processes. A major benefit of conducting these types of surveys is that respondents are surveyed shortly after their program-based decisions are made. This ensures that the reasons and motives behind their decision-making processes will be much fresher in their minds, and will lead to less biased results about the program actions they took.
- Smaller more targeted EM&V activities that address specific areas within a program's operations, in lieu of a comprehensive traditional program evaluation study, are encouraged. This approach is enabled by advances in technology that allow for faster data collection and may reduce the time and cost of conducting EM&V activities. This approach can allow program administrators to react to market conditions or shifts in technology, better understanding opportunities for program growth and improvement by taking a deeper dive into specific points of concern or interest.
- Field level assessments (i.e., metering, logging, etc.) of specific measures or equipment can be included into program operations, as opposed to "after the fact" or during the post installation phase of an evaluation cycle. Doing so allows for more proactive planning, and may result in a more integrated installation/EM&V approach, enabling continuous feedback.

Flexibility to commence ad hoc, shorter timeframe assessments, on a specific focus area or when discrete issues arise can yield more actionable and relevant results for planning and implementation purposes.

CHARACTERIZING AND REDUCING UNCERTAINTY IN EM&V RESULTS

In developing study designs, program administrators and evaluators need to make the best use of available resources. The costs of EM&V research are driven primarily by, the selection of primary data collection methods, sample design, and sample size. In making decisions in regard to these aspects of study design, evaluators should take into account the following:

- *Application of study findings in policy and program decision-making.* Not all policy and program design decisions require the same level of certainty. For example, if policy-makers wish to assess with a set level of precision the savings developed from a portfolio of programs that address a given customer segment, rather than the savings of the individual programs in the portfolio, sample sizes at the program level can be reduced. Similarly, if the likelihood of success for a program is the same whether the share of customers who meet eligibility criteria is 40% or 80%, then a sample size required to produce a +/- 10% confidence interval is not necessary. The much smaller sample required to produce a +/- 20% confidence interval will suffice. However, if decisions hinge on precise estimates of the *difference* in populations defined by program participation or other characteristics, required samples and associated costs will be much larger than those needed to characterize the individual populations.

- *Prior experience with the source and magnitude of uncertainty from various sources in regard to the results of interest.* The accuracy of estimates developed from primary data collection can be compromised by a number of factors, most importantly sample bias, measurement error, and sampling error. In many cases, the effects of bias and measurement error can far outweigh the impacts of sampling error on the accuracy of the estimate of savings or other key results. In such cases, expenditure of resources to mitigate potential bias and/or measurement error may contribute more to improving the accuracy of the estimate than similar expenditures on increasing sample size. When developing the detailed scope of work for a study, evaluators should collect and assess studies of similar programs and technical topics to understand the relative importance and likely magnitude of uncertainty linked to potential bias, measurement error, and the underlying variability of the parameters under investigation. The results of this investigation may then be used to identify the research strategies that will develop the most accurate estimates of key program outcomes given the available evaluation budget.

Program administrators and evaluators should strive for a 90/10 confidence/precision level as it represents industry standard. At this level, one can be 90 percent confident that the savings for the program is within +/- 10 percent of the evaluated program savings. However, program administrators must consider the costs associated with achieving this sample precision level in the context of the intended use of the results. As a result, the program administrator may determine to retain the 90/10 program level precision requirement, relax the requirement, or make it more stringent based on the specific information need and use of the study results. In the event the 90/10 confidence/precision level is adjusted, program administrators must document the reason for doing so in the respective EM&V plan.

Additional detail regarding this topic can be found in Appendix B.

CUSTOMER DATA GUIDELINES

Collecting and analyzing customer energy consumption data is often a cost effective approach for documenting energy savings and estimating energy consumption baselines and energy intensity levels by sector. While the availability of customer data may facilitate rigorous and cost-effective EM&V activities, priority must be given to protecting the customer's privacy and data. Appendix C of this Guidance provides guidelines for access to customer data, securing customer consent and maintaining confidentiality of customer data to be followed by program administrators and evaluators in conducting EM&V activities.

ETHICAL AND PERFORMANCE STANDARDS

Program administrators must take all necessary steps to eliminate the opportunities for bias in conducting EM&V activities. This is a critical issue considering that the organization responsible for program administration is also responsible for program EM&V. To protect the integrity of EM&V, program administrators should, to the greatest extent possible, create an organizational separation between the program implementation and program EM&V functions. Program administrators should seek to work with independent third-party consultants who have exhibited a high degree of evaluation ethics. While Staff will not provide direct review and approval of the various elements of EM&V activities, Staff will retain a general regulatory oversight and monitoring role. Adherence to this Guidance, including quality of EM&V work product and incorporation of results resides with each individual program administrator.

In the event deficiencies are found in the conduct and results of EM&V activities, Staff will inform program administrators and determine appropriate corrective actions.

COMPONENTS OF AN EM&V PLAN

An EM&V plan documents and demonstrates a commitment to transparent and credible EM&V activities and results. EM&V plans should include the components outlined in Appendix D. The details of EM&V plans will necessarily vary depending on the size, scope, and type of subject matter being evaluated. However, all EM&V plans are expected to clearly explain how the resulting EM&V activities will be consistent with the core goals of providing reliable, timely, cost conscious and transparent results.

EM&V PLAN FILING REQUIREMENTS

This Guidance stresses the importance of transparency of the EM&V activities and results. In support of this, program administrators are required to file all EM&V plans publically through the Commission's Document Matter Management (DMM) System in Matter 16-02180, In the Matter of Clean Energy EM&V.

III. PROCESS EVALUATION

Process evaluation plays an important role in the overall context of a program evaluation. The primary purpose of process evaluation is to develop actionable recommendations for program design or operational changes that can be expected to cost-effectively improve program delivery by addressing the issues, conditions, or problems being investigated. The types of information addressed in a process evaluation typically include:

- Level of customer satisfaction with the program.
- Effectiveness of program delivery mechanisms
- Effectiveness of program marketing
- Barriers to program participation
- Assessment of remaining program potential
- Assessment of why non-participants did not participate
- Review of program data collection and tracking systems
- Identification of lessons learned and specific actionable recommendations for program improvement

In order to meet the current and future needs of New York's energy programs, future process evaluations will likely need to place increased emphasis on balancing the need for objective and detailed analysis with the need for more timely results. In some cases, highly targeted process evaluations may prove more effective than conducting a traditional, full-scale process evaluation. This type of evaluation could undertake a deeper investigation into specific areas of concern or interest within a program's operations allowing program administrators to better understand potential opportunities for program improvement and to react to market conditions or shifts in technology. Appendix E provides additional detail and protocols related to process evaluations.

IV. IMPACT EVALUATION

There are often multiple approaches for estimating the same evaluation variable. For example, operating hours of a CFL in a residential home can be estimated either by a phone survey that simply asks residents how long they run their lights, or by metering actual usage. The latter approach is dramatically more expensive, but much more accurate. Program administrators and evaluators can refer to the 2012 State and Local Energy Efficiency Action Network (SEE Action) *Energy Efficiency Program Impact Evaluation Guide* (Chapters Three, Four and Seven) for various options for selecting approaches to evaluating gross energy and demand savings.[5] These approaches are based on widely accepted standards such as the *International Performance Measurement and Verification Protocol* (IPMVP), which is often referenced as a general guide to measurement and verification efforts. A more detailed reference document that provides protocols for specific measures is the Uniform Methods Project.[6] This Guide and Protocols have been produced with input from diverse nationwide teams of respected energy and evaluation program experts, including members from New York State. It is impractical and unnecessary for this Guidance to recreate a similar work product, rather these resources are provided as references to program administrators and evaluators as examples of generally accepted evaluation approaches to be considered in the conduct of their EM&V work. Program administrators should clearly articulate the rationale for the selected approach for an EM&V activity in their EM&V Plan.

In addition to documenting program impacts, another key objective of impact evaluation is to provide feedback that can be used to validate or update tools used to estimate energy savings such as the Technical Resource Manual, which is updated annually to reflect enhancements, additions, and modifications resulting from EM&V activities. In the development and implementation of impact evaluations, especially those that are measure specific, attention should be placed on collecting the data necessary to produce actionable recommendations to inform the Technical Resource Manual revision process, to the extent possible.

While process evaluation is normally the primary source of data to inform recommendations for improving program operation and design, experience has demonstrated that impact evaluation can also be a useful tool for further enhancing energy program operations. For example, in developing estimates of program energy savings, an impact evaluation could include interviews with managers of multifamily buildings, who in addition to providing insights regarding the operation of energy measures within their buildings, might also provide valuable insights for improving program effectiveness. Impact evaluations can provide valuable insights in other critical ways, including enhancing data collection and program tracking, and the reliability of energy savings estimates. As a result, impact evaluations should, to the degree practicable, seek information to serve as the basis for actionable recommendations for program improvement. These objectives should be reflected in the EM&V plans.

DATA COLLECTION APPROACHES AND ANALYSIS

Impact analysis requires data to assess the performance of the program (or in some cases the measure(s)). Typical on-site data utilized to characterize existing baselines and performance assessment

[5] State and Local Energy Efficiency Action Network. 2012. Energy Efficiency Program Impact Evaluation Guide. Prepared by Steven R. Schiller, Schiller Consulting, Inc., www.seeaction.energy.gov

[6] Available at http://energy.gov/eere/about-us/ump-protocols. The 2010 *Regional EM&V Methods and Savings Assumption Guidelines*, developed by the NEEP EM&V Forum is a similar but older and more concise resource, available at hwww.neep.org/regional-emv-methods-and-savings-assumptions-guidelines-2010.

includes:

- Energy use (kWh, therms)
- Demand (independent and coincident peak kW)
- Load factors
- Time of use
- Operating hours and set points, and
- Short or long term persistence.

Commonly used techniques for data collection to estimate program impacts include end use metering and billing analysis. End use metering involves placing a meter directly on a specific end use technology (e.g., lighting, refrigerator, industrial process equipment) to collect data such as energy consumption and operating hours. Billing analysis compares energy consumption data from program participants before the energy measure (s) is installed to consumption data for a comparable period afterward. While the study period can vary, it is often 12 months before and 12 months after the installation of the measure(s). While these techniques have proven to be effective, end use metering can be expensive due to the cost of purchasing and installing required hardware and billing analysis usually requires several months of data to perform an accurate analysis.

ADVANCED M&V

Program administrators and evaluators are encouraged to use advanced M&V techniques when appropriate and cost effective, to collect, aggregate and analyze data.[7] Advanced M&V is generally defined as technologies and practices that include, but are not limited to, automated M&V software, data analytics, advanced metering or sub-metering, building or home energy management systems, load monitoring systems, utilization of data science practices, and other emerging technologies.[8]

At the facility level, sources of data may be a whole-premise meter, a sub-meter, a system monitoring device, building device controls or potentially other systems. At the market level, advanced M&V tools that use sector-specific technologies may provide continuous, granular feedback about program performance, especially in broad market-based programs, where immediate information is necessary to assess if the intervention is actually having an effect in the market. A defining criterion for automated M&V software is that it continuously analyzes data as it becomes available. Data analysis is capable of providing ongoing feedback within a short timeframe on program performance that helps identify opportunities to make process improvements.

Evaluators may leverage the use of data generated by systems installed by other entities including program administrators, contractors, and customers, as long as customer privacy is maintained. Program administrators and evaluators should assess the data and analysis before using it to avoid bias and ensure independence. In instances where advanced M&V tools are providing continuous savings estimates for a particular energy efficiency activity, and the data and analysis has been assessed to determine the reliability of the information, program administrators may be able to extend program EM&V cycles and rely on the advanced M&V tools to provide interim impact results. In instances where advanced M&V tools support program implementation and evaluation, the costs of implementing systems that generate data may be

[7] DNV GL. 2015c. The Changing EM&V Paradigm – A Review of Key Trends and New Industry Developments, and Their Implications on Current and Future EM&V Practices.

[8] DNV GL, pgs. 32-52.

shared between program implementation and evaluation budgets. Of note, some advanced M&V technologies, due to the broader benefits to customers, may also be cost-shared by the end-use customer.

BEHAVIOR BASED ENERGY EFFICIENCY PROGRAMS

Some New York program administrators offer behavior based energy efficiency programs designed to provide customers with information and encouragement to voluntarily alter their behavior to reduce energy usage. Behavioral programs vary significantly in terms of scale, customer groups, and end-uses targeted. Regardless of the program design, the expectation for rigorous EM&V remains a constant.

Many residential behavior-based programs encourage customers to undertake energy efficiency actions by comparing their current consumption to their past consumption, as well as the amount of energy used by neighbors in comparable homes. This program design facilitates the use of random control trials in deployment and evaluation, which greatly simplifies estimation of net savings. Designs that require customers to take affirmative action to participate (that is, to opt in or opt out of the treatment group) require additional EM&V efforts to quantify net savings. Recently, government bodies have developed comprehensive, peer-reviewed guidelines for the evaluation of residential behavior based programs. [9] Evaluators of such programs are encouraged to review these guides and consider the adoption of the recommended practices.

Non-residential behavior based programs take a variety of forms. Many focus on enabling and encouraging commercial and industrial facility staff to adopt rigorous energy management practices, by providing training, tools, technical support, and access to networks of peers. Another large group of programs aims to change occupant behavior through information, feedback, and a broad range of other strategies. The principal challenges in estimating energy savings for such programs are to take into account the effects of other influences on consumption, including changes in occupancy and configuration of the facility in question and savings associated with capital measures supported by the sponsor's other programs. EM&V plans need to address issues of quantifying changes in energy consumption in commercial and industrial facilities over time, while accounting for non-program influences.

APPROACHES TO ASSESSING PROGRAM INFLUENCE

Evaluation of ratepayer funded clean energy programs should not be limited to analysis focused only on program-specific impact evaluation because this approach captures only part of the story. It is also important to examine the broader impacts of programs, including assessing market dynamics (e.g., how the market is evolving), understanding the effects of emerging technologies (e.g., growing use of LED lighting) and monitoring product baselines (e.g., the percentage of homes in New York with high efficiency heating equipment). This type of research can provide numerous benefits including insights capable of informing strategic policy decisions, improving program design and implementation, and encouraging more rigorous evaluation results.

[9] Stewart, James and Annika Todd. 2015. *Residential Behavior Protocol: The Uniform Methods Project.* Golden, CO: National Renewable Energy Laboratory. http://energy.gov/sites/prod/files/2015/02/f19/UMPChapter17-residential-behavior.pdf State and Local Energy Efficiency Action Network. 2012. Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. http://behavioranalytics.lbl.gov.

In the REV Track II Order[10], the Commission describes the complexity and challenges related to utility shareholder incentive mechanisms that depend on a determination of what would have taken place in the absence of the program, in other words proving of a counterfactual. These include administrative challenges, contentious ex-post review processes, and the potential for tremendous administrative expense for uncertain net benefit. While this discussion is specific to the development of future electric utility EAMs, consideration should be given to the value of undertaking net-savings evaluations and the intended use of the output of such work. Since ratepayer funded activities are not designed toward actions that would have occurred anyway, some level of net savings assessment, or examination of program influence, can serve as an effective tool for program design and implementation. However, given the variety of activities occurring in the marketplace, including Commission direction for NYSERDA and utility offerings to become more complementary in nature, it will be increasingly more difficult to parse out the effects of any one specific program action.

If and when net-savings evaluations are deemed appropriate, program administrators and evaluators must determine the methods most appropriate for the goals of the EM&V activity. The following methodological approaches can be found in the energy efficiency program evaluation literature for assessing program influence:

- **Analysis of self-reports of program effects by targeted market actors (Self-reports).** This approach typically involves surveying samples of actual and/or potential program participants to elicit their assessment of the program's influence on their decisions to adopt energy efficiency measures or practices. The questions can be structured to probe the effect of the program on the timing, extent, and features of the projects in question, as well as the relative importance of the program versus other decision factors. The responses can then be processed to develop an attribution score using a transparent algorithm. This survey approach can also be applied to non-participants to estimate spillover and free ridership. In recent years, the self-report method was the approach most prominently used in New York and therefore previous versions of Staff-issued Evaluation Guidelines provided additional technical detail in this area and others related to conventional resource acquisition impact evaluation. Appendix F retains this information.

Given the evolution of clean energy activities under the REV framework and direction from the Commission in exploring other approaches, the following alternative approaches are gaining greater use in New York's clean energy programs and are highlighted below for further consideration and use.

- **Theory Based Evaluation.** Theory-based evaluation starts with a program theory or logic model that lays out a theory of change explaining how a program is expected to produce results. The theory or logic model lays out expected outputs, near-term outcomes, medium-term outcomes, long-term outcomes, and the mechanisms by which those outcomes are expected to be achieved. The evaluator may then track the indicators associated with each of these outcomes to draw causal inferences about the influence of the program. Theory-based evaluation can be used on its own and can also serve to validate findings derived using the other methods discussed here by helping explain how and why the observed changes have occurred.

[10] Case 14-M-0101, Proceeding on Motion of the Commission in Regard to Reforming the Energy Vision, Order Adopting a Ratemaking and Utility Revenue Model Policy Framework issued May 19, 2016.

- **Experimental designs.** Experimental design – or random control trials - provides one of the strongest approaches to assessing program influence. Random assignment directly addresses one of the most serious threats to validity that is inherent in other methods for attributing program influence, namely participant self-selection. Self-selection for participation in voluntary programs generally introduces bias to quasi-experimental analyses because participants often differ systematically from non-participants in factors that affect energy savings that cannot be directly observed and controlled for statistically. Experimental designs have been used extensively to evaluate the effect of customer education and information programs. This is a good application of experimental methods because individual participants can be randomly assigned to receive different messages and information products, and the marginal cost of program delivery is very low.

- **Quasi-experimental designs.** This approach uses well-established quasi-experimental social science research designs to assess and quantify program influence. Common strategies include cross sectional methods that compare the rate of measure adoption in an area or market segment not targeted by the program as a baseline for comparison to rates of adoption in the program area. The difference between the two can be viewed as the program's net effect, after taking into account differences between the program and comparison area that might affect measure adoption. This approach accounts for spillover as well as free ridership.

Pre-post designs that compare the rate of adoption before and after the program or policy intervention have also been applied, as have mixed pre-post/cross-sectional approaches. Statistical modeling is often used to apply retrospectively quasi- experimental approaches to datasets that describe the response of a group of market actors to a given program. For example, analysis of variance and regression approaches implicitly invoke quasi-experimental designs by estimating program effects while controlling statistically for the effects of other participant attributes such as income, education, facility size, and so forth.

- **Price elasticity approaches, including conjoint analysis and revealed preference analysis.** In these two approaches, researchers assess the effect on changes in price on customer's likelihood of purchasing an energy-efficient product or service. The results of these assessments can then be combined with information on the actual effect of the program on the price participants paid for the product or service in question to estimate the effect of a program-related purchase incentive on the pace of sales. In the case of conjoint analysis, customers are asked to rank a structured set of hypothetical products that vary along a number of dimensions, including performance and price. In the revealed preference approach, purchasers are intercepted at the point of sale to gather information on product selection they actually made, its price, and other features.

- **Structured expert judging.** Structured expert judgment studies assemble panels of individuals with close working knowledge of the various causes for changes in the market, technology, infrastructure systems, markets, and political environments addressed by given energy efficiency programs to estimate baseline market share and, in some cases, forecast market share with and without the program in place. Structured expert judgment processes employ a variety of specific techniques to ensure that the participating experts specify and take into account key assumptions about the specific mechanisms by which the programs achieve their effects. The Delphi Method is the most widely known of this family of methods.

- **Historical Tracing: Case Study Method.** This method involves the careful reconstruction of events leading to the outcome of interest, for example, the launch of a product, the passage of legislation, or the completion of a large renewable energy project, to develop a 'weight of evidence' conclusion regarding the specific influence or role of the program in question on the outcome.

Evaluators use information from a wide range of sources to inform historical tracing analyses. These include public and private documents, personal interviews, and surveys conducted either for the study at hand or for other applications.

- **Top-Down Modeling** The goal of top-down modeling is to isolate the effects of program activity from other policy variables and from naturally occurring changes at the sector level (e.g., residential or commercial) in order to estimate the effects of those programs on sector-level energy use. Traditional "bottom-up" impact evaluations attempt to measure the gross and/or net energy savings associated with a specific program or set of programs. In contrast, top-down models attempt to measure net changes in energy consumption over time across an entire sector that are attributable to aggregate programmatic interventions in that sector. The savings associated with a particular program or set of programs cannot be isolated with a top-down approach; rather, these models estimate savings in the residential sector as a whole or the commercial sector as a whole that result from all the energy efficiency programs targeting the entire sector. Therefore, top-down modeling studies do not serve the same function as bottom-up evaluations. However, insofar as multiple programs have cumulative effects on efficiency that are not counted by individual bottom-up evaluations – that is, if the whole is greater than the sum of the parts – then top-down modeling may capture savings not measured by traditional bottom-up evaluations. A top-down estimate of savings can also serve to validate cumulative estimates derived from multiple bottom-up impact evaluations.

Most approaches to "top-down" analysis develop econometric models of annual or more frequent total aggregate consumption of one fuel (electricity or gas) in one region as a function of indicators of program effort, controlling for non-program factors such as weather, energy prices, and changes in overall levels of economic activity. At a minimum, time series data on aggregate consumption and the independent variables going back a number of years are required to obtain the variability required for econometric modeling. Recent experience with pilot analyses suggests that, if annual data is to be used, a minimum of 5 years of consistent data is recommended.

Almost all top-down analyses rely on regression methods, and the precision of regression estimates is improved by the inclusion of larger numbers of observations in the model. This can be affected by compiling data over a longer period of time, a greater number of regions, or a subdivision of regions.

Selection of methods for use in evaluation of a given program will depend on a number of factors. Factors that should be recognized include:

- Availability of consumption data for participants and non-participants;

- Size of expected savings compared to total whole-premise consumption;
- Availability of transaction data needed for price elasticity and other econometric analysis;
- Scale of the program and potential risks to portfolio-level claimed impacts or overall cost effectiveness; and
- Developments in the technologies and markets addressed by the program which might be associated with.

Table 1 below provides a high-level summary on estimation methods, for various types of programs for assessing program influence. For individual EM&V activities, method selection should take into account the factors identified above.

Table 1: Assessing Influence of Different Program Types

Primary Approach	= ◊
Secondary Approach	= ●
Unlikely to be Applicable	= ♦

Program Types	**Market Actor Self-reports**	**Theory Based Evaluation**	**Experimental Designs**	**Quasi-Experiments**	**Price Elasticity/ Econometric**	**Expert Judging**	**Case Studies**	**Top Down Modeling**
COMMERCIAL & INDUSTRIAL		◊						◊
Building Retrofit and Equipment Replacement	◊	◊	♦	●	●	●	●	
New Construction	◊	◊	♦	●	♦	◊	●	
Information and Training Programs	◊	◊	♦	●	♦	●	◊	
Codes & Standards Enhancement/Enforcement	●	◊	♦	♦	♦	◊	●	
RESIDENTIAL		◊						◊
Whole-building Retrofit	◊	◊	♦	◊	♦	♦	♦	
Home Energy Reports	●	◊	◊	◊	♦	♦	♦	
Upstream Products	●	◊	♦	◊	◊	●	♦	
New Construction	◊	◊	♦	●	♦	◊	●	

Program administrators are encouraged to use multiple methods, where appropriate and cost effective, to help triangulate and provide more robust and credible information as to impact and influence.

Market transformation activities may require further evolution of approaches over time, in conjunction with the work of the Clean Energy Advisory Council and on-going regulatory requirements (e.g., REV, Clean Energy Fund, and the Clean Energy Standard). Consideration should be given to how market transformation should be properly evaluated, including market-level studies and top-down approaches.

V. EM&V REPORTS

The output of EM&V activities should be transparent, useful, and actionable. In order to be able to judge the reliability of results, EM&V reports must be reviewable – that is, it must be possible for a reviewer to make an independent assessment of the validity of the reported findings. In order to be reviewable, reports must be clearly written, consistently present key variables and statistics of interest, and be easily accessed. While there is good reason to include detail on study objectives, methods, results and recommendations, EM&V reports should not become inaccessible to broader audiences, or obscure the key findings of the work due to ineffective packaging and presentation of the needed information. The varied audiences that will use EM&V reports should be provided with an appropriate level of information to meet their needs. A common format will aid stakeholders in their review of EM&V reports issued by the various program administrators. EM&V reports must balance accessibility with the need for adequate detail to provide an understanding of the EM&V approach, robustness of the results, and clearly articulate the recommended actions.

COMPONENTS OF AN EM&V REPORT

EM&V reports should include the components outlined in Appendix G. The details of EM&V reports will necessarily vary depending on the size, scope, and type of subject matter being evaluated. However, all EM&V reports will be guided by the core principles of providing reliable, timely, cost conscious and transparent results.

EM&V REPORT FILING REQUIREMENTS

This Guidance stresses the importance of transparency of EM&V activities and timeliness of actionable results. In support of this, program administrators are required to file all EM&V final reports through the Commission’s Document Matter Management (DMM) System in Matter 16-02180, In the Matter of Clean Energy Program Evaluation, Measurement & Verification.

Appendix A: Key Terms & Definitions

It is important for all EM&V activities, and resulting reports, to use common terminology to improve the consistency of the results and promote a common understanding by all stakeholders. A number of glossaries of evaluation terms currently exist and are frequently used within the industry. For purposes of this Guidance document, the following list of key terms and definitions are provided as a quick reference and to ensure comparable treatment across EM&V activities in New York.[11]

Adjusted Gross Savings - The change in energy consumption and/or demand that results directly from program-related actions taken by participants in an efficiency program, regardless of why they participated. It adjusts for such factors as data errors, installation and persistence rates, and hours of use. This may be referred to as Ex Post Savings when the focus of the evaluation is on verification of realized energy savings and not related to calculating program influence.

Baseline - Conditions, including energy consumption and related emissions that would have occurred without implementation of the subject measure or project. Baseline conditions are sometimes referred to as “business-as-usual” conditions and are used to calculate program-related efficiency or emissions savings. Baselines can be defined as either project-specific baselines or performance standard baselines (e.g. building codes).

Cumulative Annual Savings: The reduction in electricity usage (MWh) or in fossil fuel use in thermal unit(s) realized in a given calendar year as a result of all measures installed to date and still in operation. Cumulative annual savings may differ from cumulative first year annual savings due to the lifetime of measures. (DPS)

First Year Annual Savings - The reduction in electricity usage (MWh) or in fossil fuel use in thermal unit(s) from the savings associated with an energy saving measure, project, or program calculated based on a full year’s installation and operation.

Gross Savings - The change in energy consumption and/or demand that results directly from program-related actions taken by participants in an efficiency program, regardless of why they participated and unadjusted by any factors. This is sometimes referred to as program-reported savings, prior to any evaluations, or Ex Ante Savings.

Lifetime Energy Savings - The electric or gas energy savings over the lifetime of an installed measure(s), calculated by multiplying the annual electric or gas usage reduction associated with a measure(s) by the expected lifetime of that measure(s).

Net Savings: The change in energy consumption and/or demand that is attributable to a particular energy efficiency program. This change in energy use and/or demand may include, implicitly or explicitly, consideration of factors such as free ridership, participant, and non-participant spillover, and induced market effects. These factors may be considered in how a baseline is defined (e.g., common practice) and/or in adjustments to gross savings values.

[11] Unless otherwise noted, all definitions are based on the Northeast Energy Efficiency Partnership EM&V Forum, Glossary of Terms, Version 2.1, July 2011. http://www.neep.org/emv-glossary or the State and Local energy Efficiency Action Network. 2012. Energy Efficiency Program Impact Evaluation Guide. Prepared by Steven R. Schiller, Schiller Consulting, Inc.. www.seeaction.energy.gov Some definitions have been expanded upon for clarity.

Realization Rate: Used in several contexts for comparing one savings estimate with another. The primary and most meaningful application is the ratio of evaluated gross savings to claimed gross savings. Basis for the ratio not being 1.0 can include several considerations, such as (1) adjustments for data errors, (2) differences in implemented measure counts as a result of verification activities, and/or (3) other differences revealed through the evaluation process, such as changes in baseline assumptions.

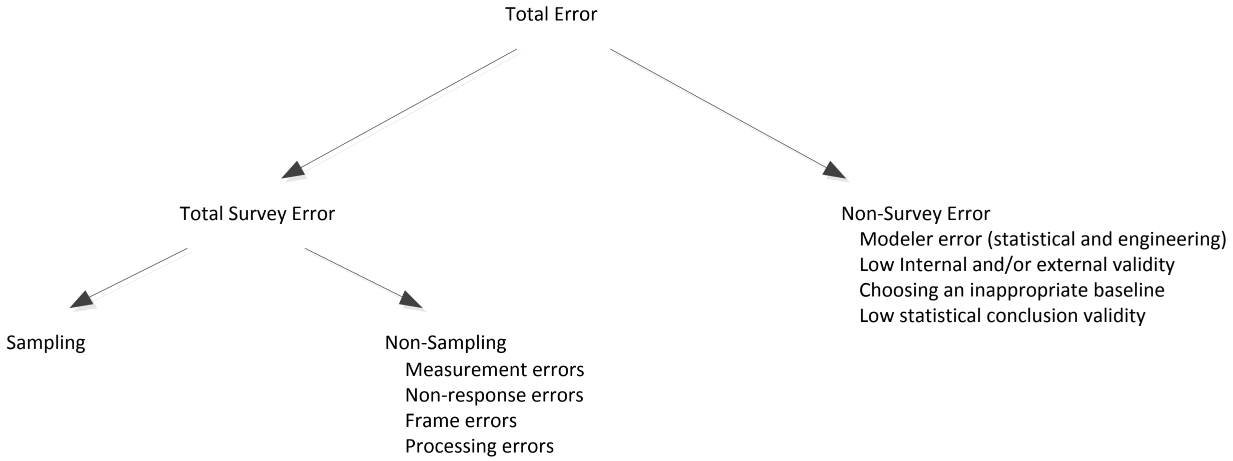
Appendix B: Characterizing and Reducing Uncertainty in Evaluation Results

This Appendix discusses certain methodological principles regarding the reliable estimation of savings, i.e., estimates that are reasonably precise *and* accurate, and reflects clarification or further detail on topics based on program administrators' inquiries and/or Staff's review of previous evaluation plans and completed reports.

Addressing the Multiple Sources of Error

In the design and implementation of any impact evaluation design, evaluators should attempt to cost-effectively mitigate various sources of error in estimating savings. Figure 1 presents a typology of some of the most important sources of error.

Figure 1: Sources of Error in Estimating Energy and Demand Savings


Total Error
Total Survey Error
Non-Survey Error
Modeler error (statistical and engineering)
Low Internal and/or external validity
Choosing an inappropriate baseline
Low statistical conclusion validity
Sampling
Non-Sampling
Measurement errors
Non-response errors
Frame errors
Processing errors

With respect to sampling error, for program-level samples, the minimum standards for confidence and relative precision are set at 90/10 for estimating gross energy savings and for customer surveys at the program level. This Guidance notes that if the planned or achieved confidence and precision could not or did not meet the 90/10 standard, the plan or final report should clearly indicate the reasons it was not practical and provide a detailed justification. Specific approaches to sampling are left up to the evaluator and should consider costs as well as the overall goals of the EM&V activity in determining the appropriate approach. For example, one can choose from a variety of sample procedures recognized in the statistical literature, such as sequential sampling, cluster sampling, stratified random samples, and stratified ratio estimators. Any of these, and others, could be appropriate depending on the circumstances. There are many available books on sampling techniques that can be used as reference, including Cochran (1977), Thompson (2002), TecMarket Works (2004 and 2005), and Sarndal et al. (1992).

However, in any given study, literature has shown the potential for bias could be much more important than sampling error.[12] Some evaluators make the mistake of focusing almost exclusively on reducing sampling error by insisting on large samples while devoting relatively little attention to addressing the many other sources of error. As a result, some studies achieve a high level of confidence and precision around a biased estimate, which compromises the objective of obtaining reliable estimates of energy and demand impacts. As appropriate, evaluators should attempt to mitigate the various sources of error in their evaluations. To do so, the evaluator must have the flexibility to respond to data issues as they arise in order to maximize the reliability of the savings.

Thus, given the multiple sources of error and budget constraints, evaluators are frequently forced to make tradeoffs in the planning and/or implementation of an evaluation resulting, in some cases, in reduced sample sizes and lower confidence and precision or to seek additional funding for the study. Listed below are a few examples of tradeoffs that may occur:

- A program might be so small that expending scarce evaluation dollars to achieve the 90/10 level of confidence and precision might not be cost-effective.
- The expected savings could be so uncertain that more evaluation dollars must be allocated to on-site M&V in order to achieve more accurate estimates of savings.
- The expected or observed non-response rate could be so high that evaluation dollars must be allocated to address potential non-response bias.
- In screening for particular types of customers (e.g., those who have purchased an air conditioner in the last year), the actual incidence could be so low that the planned sample size cannot be achieved.
- In some cases, the evaluator might have underestimated the actual variability in a given parameter in the population (e.g., savings, satisfaction, etc.) making it impossible to achieve the target with the planned sample size.
- After the plan is initiated, the program administrator might decide to increase the level of on-site M&V to improve the accuracy of energy and demand estimates thus forcing the evaluator to reduce the sample size.

In their EM&V plans and final reports, evaluators should clearly explain how the relevant sources of error were addressed and their rationale for doing so.

The major sources of uncertainty in energy program evaluation research along with common steps for mitigating their impact are described below.

Bias: problems in sample design, selection, and realization. The design and execution of data collection from a sample of the population under study must meet certain criteria in order for the averages, totals, and ratios computed from those data to yield estimates that equal the corresponding values for the full population, with calculable probabilities. An unbiased sample is one in which every individual or element in the population has an equal probability of being selected, which implies the application of random sample selection methods. Often, evaluation research will segment or stratify the population to reduce variance or the costs of data collection. Stratified samples can meet the criterion for “unbiased” so long as the appropriate sample weights are used in calculating parameters.

[12] Groves, 2004; Biemer et al., 2004; Lyberg et al., 1997; Biemer and Lyberg, 2003; Lessler and Kalsbeek, 1992; Sonnenblick and Eto, 1995; California Evaluation Framework, 2004; California Protocols, 2005.

In typical evaluation research situations, implementation of unbiased samples may be complicated by a number factors including:

- *Incomplete sample lists.* Often, it is difficult to develop lists of key groups to be contacted for a given study. Such groups include tenants and owners of multifamily buildings, residential building contractors and tradespeople, or designers working on various kinds of commercial energy efficiency projects, among others. Generally, researchers do the best they can working with various sources, such as telephone directories and membership lists for trade and industry associations. These are generally known to be incomplete, but characteristics of individuals or firms not in the lists are not well understood. Practical methods to address this problem include acquisition or purchase of multiple lists and extensive data processing to eliminate duplicates and characterize the population.

- *Non-response bias.* Sizable rates of non-completion raise the concern that the subjects who *have* responded to a survey do not constitute a random sample. Recently, survey response rates have been decreasing due to the proliferation of cell-phone only households and general respondent fatigue. Practical strategies to increase response rates include the following:
    - Use of respondent incentives;
    - Deployment of advance letters to inform sample individuals of the survey's purpose, sponsorship, time requirements, and (where relevant) respondent incentives;
    - Use of multiple media for response, including phone, mail, and e-mail;
    - Use of endorsements for the survey from individuals held in esteem by the target population;
    - Use of social "nudges" such as information on the current completion rate among the respondents' neighbors or commercial peers.

Measurement error. Measurement error occurs when surveys or measurements carried out on individual members of the sample are inaccurate. This may occur if interview subjects misunderstand questions or provide inaccurate answers, or when metering equipment does not function as expected. Measurement error will lead to inaccurate results. For surveys of individuals, practical methods to reduce measurement error include intensive testing of questionnaires with professionals and "live" subjects prior to deployment in the field, careful monitoring of early results to identify evidence of misunderstanding of various items, and use of validated scales for attitudinal and psychographic variables. For surveys of facilities, practical methods to reduce measurement error include professional review of site-level M&V plans, calibration of end use metering, quality control of site data collected, and professional review of site-level savings calculations.

Sampling Error. Sampling error occurs because the estimators of averages, totals, and ratios based on a sample are likely to differ from the *true* value of those quantities that would be obtained by taking surveys or measurements of all facilities or persons in the population (i.e., a full census). The size of sampling error is a function of the type of quantity to be estimated (proportion, mean, total, difference in proportions or means, regression coefficients), the underlying variability of the parameter, the sample design, and the sample size. See Appendix E for a detailed discussion of sampling and uncertainty including addressing multiple sources of error and sample weighting. Appendix H offers guidance for calculating the relative precision for net program-level savings.

Weights

In the design and implementation of any sample, there are various situations when weights must be calculated to correct for differential selection probabilities, to adjust for unit non-response, for post-stratification, or for various combinations of these.[13] The correct calculation and application of weights are critical, therefore EM&V reports must clearly explain:

- Why weighting was necessary,
- The information used to calculate the weights, and
- The formulas used to calculate the weights.

Such detailed information can be included in a technical appendix to the final report.

Detailed Guidance

More detailed guidance is provided below on topics noted during reviews of previous program administrators' impact evaluation plans and reports:

- When sampling supply-side market actors, define the target population appropriately so that its members are reasonably homogeneous in terms of their fundamental role in the market. This is, of course, a matter of degree, and to some extent heterogeneity is exactly what sampling is intended to help manage. For example, one would not want to define the population as something as specific as lighting contractors with 10-25 employees who do 50-75% of their work in the commercial sector. But at the same time, it typically would not be appropriate to define the population as all supply-side actors who have any potential for involvement in the program, because unless the program itself targets a very specific niche, this is likely to include fundamentally different kinds of players, causing summary statistics to have little meaning.

- *Because there are large variations in the size of different market participants within the same category, often it is desirable to oversample larger players in order to enhance sampling efficiency, and then to down-weight these larger players in the analysis stage in order to ensure accurate representation of the population.* There are standard statistical methods for doing this effectively assuming adequate information is available on the size distribution in the market, see TecMarket Works (2004) and Cochran (1977) for discussions of stratified sampling by size.

- Give thought in advance, to what characteristics of the market are being investigated, and shape the weighting schemes accordingly. If the research goal is to represent overall activity and/or transactions in the market, it will generally be desirable to weight by size, reflecting the fact that each large player makes a much larger contribution to overall market activity than does each small player.[14] When the objective is to represent the overall firmographics of the population, then one

[13] Skinner et al., 1989; Groves et al., 2004; Kish, 1965; Cochran, 1977; Lee et al., 2006

[14] Note that weighting by size in order to accurately reflect the disproportionate contributions of large market players to overall market activity and weighting based on size in order to account for over-sampling of large players done for purposes of sampling efficiency are fundamentally distinct issues. The latter is done as part of an overall

should not weight by size. Because the same study often incorporates multiple research objectives, it may be appropriate within a single study to weight by size for some analyses and to not weight by size for others.

- In forecasting likely precision and estimating needed sample sizes, consider the potential need to disaggregate the results for individual sub-sets of the overall population. It is relatively unusual for the analysis of an evaluation dataset to begin and end with the overall population. More often, there are certain researchable questions for which only subsets of the population are of interest, and other questions that require contrasts between different subsets. When this is the case, the expected precision for the sample as a whole is not a good predictor of the reliability of the results, and a sample that is designed solely around precision objectives for the population as a whole is likely to provide results that are more uncertain than may be desired at various levels of disaggregation. Subject to budget constraints, sample designs should therefore take into account what types of sub-population analyses and contrasts are likely to be of interest.

- When there is great uncertainty regarding the overall population size, use the survey itself (to be more specific, typically the screener question(s)) to refine understanding of that issue. The sources available for the development of a sample frame does not always allow the evaluator direct access to the population of interest. Often it is necessary to contact a broader set of respondents, using an up-front screener to identify those who genuinely fall into the target population. This tends to be particularly true of surveys of supply-side actors that use commercial databases such as Dun & Bradstreet. When this occurs, it is critical to use the results of the screener questions to refine the evaluators' understanding of the size and firmographic characteristics of the target population. Such analyses can inform both the current study and future studies of the same market. A corollary is that it is often important to design screener questions in such a manner that, before non-qualifying cases are terminated, enough data are collected from them to use in refining understanding of the target population.

- Comparisons between two or more subsets of the overall sample (e.g., upstate versus downstate New York) that do not take appropriate account of sample definition and weighting issues as discussed above have significant potential to produce false results. If a sample includes cases from what are in reality multiple fundamentally distinct populations, or cases that are not weighted appropriately to reflect differential sampling rates, then comparisons between key subsets of the sample will likely be inappropriate due to the potential for differences in the composition of the subsamples being compared. For example, if a single statewide sample includes both distributors and contractors, and if distributors tend to be disproportionately based upstate, then the results of unadjusted comparisons between upstate and downstate may simply reflect differences between

sampling strategy that includes differential sampling rates for different size categories and is done in order to enhance sampling efficiency. Such weights are referred to as *stratum* or *expansion* weights. The former is done in order to capture an accurate representation of total market activity, may occur regardless of whether or not large players have been over-sampled for purposes of sampling efficiency, and may be applied only to certain analyses. These weights are the same as those used to calculate weighted means. It is possible for both types of size-related weighting issues to arise in the same study, and even for the purpose of the same analysis. When both types of weighting occur in the same study, it is important to maintain conceptual clarity about these differences.

distributors and contractors rather than meaningful differences between regions.

- Beware of the tendency for samples of the general population of supply-side actors to result in a disproportionate number of participants due to differential acceptance rates. This tendency may call for financial or other types of incentives for cooperation with data collection activities, particularly for non-participants.

REFERENCES

1. Biemer, Paul P., Robert M. Groves, Lars E. Lyberg, Nancy A. Mathiowetz, and Symour Sudman. (2004). *Measurement Error in Surveys*. Hoboken, New Jersey: John Wiley & Sons, Inc.
2. Biemer, Paul P. and Lars E. Lyberg. (2003). *Introduction to Survey Quality*. Hoboken, New Jersey: John Wiley & Sons, Inc.
3. Cochran, William G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
4. Enders, Craig K. (2010). *Applied Missing Data Analysis*. New York: Guilford Press.
5. Groves, Robert M., Floyd J. Fowler Jr., Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. (2004). *Survey Methodology*. Hoboken, New Jersey: John Wiley & Sons.
6. Groves, Robert M. (2004). *Survey Errors and Survey Costs*. Hoboken, New Jersey: John Wiley & Sons, Inc.
7. Kish, Leslie. *Survey Sampling*. (1965). New York: John Wiley & Sons.
8. Lee, Eun Sul, Ronald N Forthofer, Ronald J. Lorimor. (2006). *Analyzing Complex Data* (Quantitative Applications in the Social Sciences - #71). Newbury Park, CA: SAGE Publications.
9. Lyberg, Lars, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. (1997). *Survey Measurement and Process Quality*. New York: John Wiley & Sons.
10. McKnight, Patrick E., Katherine M. McKnight, Souraya Sidani, and Aurelio Jose Figueredo. (2007). *Missing Data: A Gentle Introduction*. New York: Guilford Press.
11. Sarndal, Carl-Eric, Bengt Swensson and Jan Wretman. (1992). *Model Assisted Survey Sampling*. New York: Springer-Verlag.
12. Skinner, C. J., D. Holt and T. M. F. Smith. (1989). *Analysis of Complex Surveys*. New York: John Wiley & Sons.
13. Sonnenblick, Richard and Joe Eto. (1995). *Calculating the Uncertainty in Estimates of DSM Program Cost-Effectiveness*. International Energy Program Evaluation Conference: Chicago, IL. pp. 759-768.
14. TecMarket Works. (2004). *The California Evaluation Framework*. Prepared for the Southern California Edison Company.
15. TecMarket Works Team. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*.
16. Thompson, Steven K. (2002). *Sampling*. (2nd Ed.). New York, N.Y.: John Wilson

Appendix C: Customer Data Guidelines[15]

Analyzing utility customer energy consumption data is often a cost effective approach for documenting energy savings from Commission-approved energy efficiency programs. A common technique is to collect customer data before and after energy actions are implemented and use statistical analysis (e.g., adjust for variables such as weather) to estimate energy savings. Customer data can also be used for other evaluation, measurement & verification (EM&V) related research such as estimating energy consumption baselines and energy intensity levels by sector. These approaches can often provide reliable results at a lower cost than other evaluation techniques. While the availability of customer data may facilitate rigorous and cost effective EM&V, priority must be given to protecting the consumer's privacy and data.

Staff has developed guidelines for access to customer data, securing customer consent and maintaining confidentiality of customer data to be followed by program administrators (i.e., utility, NYSERDA) and their evaluation contractors in conducting EM&V and EM&V related activities.

Program Participant Consent Form

For program participants, the priorities are to ensure that they knowingly agree to disclose confidential data and to provide assurances that the availability of their data is limited to the minimum customer data necessary to conduct the EM&V activity consistent with the EM&V Guidance issued by Staff. The program administrator should provide to program participants a consent form authorizing the release of certain enumerated customer data to the program administrator and, if applicable, the evaluation contractor. This data can include consumption data, but may not include payment histories. The terms of the consent form should be prominently displayed and explain that the data will be used only for program EM&V related research; confidentiality will be strictly protected; and results will only be reported in the aggregate. A customer signature or the equivalent (i.e., an electronic signature) is required.

The consent form should be included as part of the program application material, where possible. If this is not possible, consent forms may be obtained from program participants after the point of application.

It is recommended that to further facilitate the EM&V process, the consent form, program application, or other related documents, also include language requiring program participants to agree to cooperate with activities designed to evaluate program effectiveness, such as responding to questionnaires and allowing on-site inspection and measurement of installed program supported measures.

Data Confidentiality Agreement

Evaluation contractors retained by a program administrator must sign a confidentiality agreement with the program administrator providing assurance that they will keep customer information, including energy consumption data, confidential and secure at all times. If in addition to an evaluation contractor, NYSERDA also has possession of this data, NYSERDA will agree to a similar confidentiality agreement with the utility for the customer data in their possession.

The agreement must specify how the data will be used and reported and address the following

[15] Staff originally issued Customer Data Guidelines for EM&V purposes in June 2009, with subsequent revisions in May 2011, December 2012, and with the issuance of this Guidance in November 2016.

principles:

- The evaluation contractor will maintain the confidentiality of all customer data;
- Data transfers and security policies and protocols shall meet or exceed utility and program administrator standards and requirements for cyber security;
- All customer information provided to the evaluation contractor will be used solely to conduct EM&V activities consistent with the terms of the program administrator's EM&V plan;
- Customer information will be safeguarded from unauthorized disclosure with all reasonable care;
- At the conclusion of the EM&V project, or if the program administrator and the evaluation contractor end their business relationship, the evaluation contractor or NYSERDA, will return to the utility all customer information (including any data or analyses generated from the data) and/or provide proof to the utility that the data was destroyed. If there is expected to be a need for additional analysis of the customer data after the release of the final EM&V report, the rationale for the additional research should be detailed in the EM&V plan.
- If the evaluation contractor and/or the program administrator is affiliated with or doing work for any retail energy business interest, then the evaluation contractor must provide specific details on the evaluation contractor's internal security arrangements that will keep the customer data secure from employees involved in unregulated retail energy business related activities in the service territory from which the data was extracted; and
- Each evaluation contractor that receives customer information must agree to indemnify the providing utility from any and all harm that may result from the inappropriate release of customer information by the evaluation contractor or its representatives.

Non-Participant Customer Data

Analysis of program non-participant energy consumption data can play a key role as a control or baseline against which to measure the participant group results, including helping to identify naturally occurring energy efficiency, and identifying and quantifying clean energy potential. Non-participant information may be used to more fully understand a program's strengths and weaknesses and to analyze market conditions and trends.

Staff recognizes that obtaining consent forms from non-participants could be a burden on program administrators, and in some cases is not feasible. To facilitate quality EM&V, and ensure that EM&V activities are implemented in a cost effective manner, the exchange of personally identifiable information for non-participants between a utility and its evaluation contractor (or a utility and NYSERDA, along with NYSERDA's evaluation contractors) will be permissible provided the use of the data is consistent with the objectives and requirements of the program administrator's EM&V plan and Staff has reviewed and approved the use of non-participant data in compliance with this Guidance. Moreover, the exchange shall be in compliance with the terms articulated in the Data Confidentiality Agreement section, described above.

To the extent practical, the information shall be redacted by the customer's utility to remove customer-identifying data and to only provide consumption information identified by generalized category such as service class, customer type (e.g., single family) or location (e.g., Manhattan). In instances when, after the redacting process, a customer might still be identifiable (e.g., the customer is the single large industrial customer in a small service territory), the utility should seek customer consent for inclusion of the information in the EM&V process through a signed customer consent form; exclude the information from the EM&V process; or aggregate the customer with other large customers to shield any individual customers' identity.

Some EM&V research will require personally identifiable customer information. In these cases, the utility may provide such information to (1) its evaluation contractors, and (2) NYSERDA and its evaluation contractors, under certain conditions, discussed below.

The utility evaluation contractor, or NYSERDA, must demonstrate to the utility providing the data that the information is needed to complete the EM&V activity articulated in the program administrators EM&V plan (e.g. conducting a non- participant survey). The request for customer data must be specific, explaining the need for the data and a detailed discussion of how the data will be used in the EM&V research process (e.g., type of survey, sampling approach). The evaluation contractor, or NYSERDA, must demonstrate that the information sought is the least amount necessary, both in terms of number of customers impacted and level of detail.

If a customer, whose personally identifiable information has been provided to an evaluation contractor without prior written consent, indicates that he/she is unwilling to participate in EM&V activities, or otherwise wishes not to be contacted in relation to program EM&V, the evaluation contractor must report this request to the utility and if appropriate NYSERDA, within a reasonable time. The utility will compile and maintain a "do not contact" list and refrain from including any customers on that list in response to future EM&V related data requests.

Information to be provided by a utility without prior customer consent should never include payment history or detailed usage history. Usage history, if provided at all, shall be limited to general categories of usage (e.g., commercial customers using over 100 kW) and shall only be provided when necessary to ensure EM&V activities are sufficiently targeted. Program administrators shall provide in their EM&V plans details of the customer information that will be provided to evaluation contractors (or NYSERDA), including the exact type of information, the type of evaluation survey for which it will be used, sample sizes and sampling techniques.

Appendix D: EM&V Plan Guidance

Components of an EM&V Plan

The outline below is designed to serve as a guide for preparing EM&V plans. Consistency in the development and presentation of EM&V plans will aid stakeholders in their review and understandings of EM&V activities across multiple program administrators. Program administrators should include all relevant components in their EM&V plans, however it is noted that the level of detail may differ depending on the scope and magnitude of the EM&V activity being planned. At the early stage of program development, program administrators may have some difficulty in determining certain aspects of the EM&V design, however, at a minimum, assumed initial strategies should be included.

Section	**Component**
Program/Initiative Background	Evaluation plans should briefly summarize the following: • Program/initiative description including its objectives (e.g. energy savings, market transformation) • Theory of change (include logic model, if available) • Program/initiative's anticipated benefits (e.g. energy savings, services provided, etc.) • Program/initiative schedule and budget • Links to reference documents should be provided for additional detail.
General EM&V Approach	EM&V goals (primary and secondary), including broad and tactical, as applicable. Clear statement on the intended use of the resulting information. • Brief overview of the EM&V approach • Budget for EM&V activity, including major categories (e.g., site-work, survey work, data analysis, report preparation) • EM&V activity schedule and milestones • Program administrator staff and consultant resources and their respective roles, addressing ethical and operational standards.
Detailed EM&V Approach	As applicable, based on the focus of the EM&V activity, include the following: • Process evaluation methodology • Market evaluation methodology • Impact evaluation methodology • Program influence analysis –factors considered and methods used • Sampling strategies and design, including rolling sampling techniques • Targeted level of confidence and precision, where applicable • Steps to identify and mitigate threats to data reliability, where necessary and applicable • Data collection, any advanced M&V techniques and management process*

Status Reporting	• Frequency and format of status reports to be provided to program administrator, if applicable • Final EM&V reports should adhere to the outline provided in Appendix G of this Guidance.

*As detailed in Appendix C: Customer Data Guidelines, Staff must review and approve use of non-participant data to ensure compliance with the EM&V Guidance.

Appendix E: Process Evaluation Protocols

BACKGROUND

In 2008, the New York State Public Service Commission (Commission) established the Energy Efficiency Portfolio Standard (EEPS) and a framework for evaluation of these programs including establishing Evaluation Guidelines and a statewide Evaluation Advisory Group (EAG) to advise the Commission and Department of Public Service staff (Staff) on evaluation related issues.[16]

In February 2012, Staff's original Evaluation Guidelines were supplemented with additional protocols to further guide and enhance the process evaluation of the EEPS programs. The ***New York State Process Evaluation Protocols - a supplement to the New York State Evaluation Guidelines Updated 2013 (Process Evaluation Protocols)*** were developed under the guidance of Staff and the EAG through a contract with the Johnson Consulting Group, managed and funded by the New York State Energy Research and Development Authority. As part of the Commission required review of the Evaluation Guidelines, the Process Evaluation Protocols were found to continue to offer value, but it was recognized they required an update to better align with the Reforming the Energy Vision (REV) framework. In October 2016, Staff retained and modified certain sections of the original guidance document to create this Appendix.[17]

The primary goal of this Appendix is to develop a common integrated approach to plan, direct, conduct and interpret findings from process evaluations conducted at both the program and portfolio levels. The protocols also are designed to ensure that the key stakeholders are able to integrate the findings from process evaluations into actionable recommendations to improve program design and operation.

Traditional programmatic process evaluations remain an effective tool to understand issues associated with newly launched program efforts, programs undergoing major changes, or pilot "test and learn" initiatives after they are deployed into the marketplace. However, in order to meet the current and future needs of New York's energy programs, future process evaluations will likely need to place increased emphasis on balancing the need for objective and detailed analysis with the need for more immediate results In some cases, highly targeted process evaluations may prove more effective and more timely than conducting a traditional, full -scale process evaluation. This type of evaluation could focus on specific areas within a program's operations allowing program managers to react to market conditions or shifts in technology and to better understand potential opportunities for program improvement through a deeper investigation into specific points of concern or interest. Moreover, surveys could be conducted on a "rolling basis" to regularly update program related data to provide an additional stream of program intelligence. The selection of the process evaluation scope and approaches should be driven by the needs of the program, as applicable.

While the focus of these Protocols was originally on full-scale process evaluation, the information contained within can also be adapted to provide useful guidance for other variations of process evaluation (e.g., targeted, rolling).

[16] Case 07-M-0548 Proceeding on Motion of the Commission Regarding an Energy Efficiency Portfolio Standard.

[17] The NYSERDA Performance Management team wishes to acknowledge Dr. Katherine Johnson of the Johnson Consulting Group and Gregg Eisenberg of Iron Mountain Consulting for their development and advancement of the original set of common process evaluation protocols

This document is divided into two sections:

- Section I – Provides an overview of process evaluation terms, methods, and approaches

- Section II – Contains the process evaluation protocols for NYS programs

SECTION I: PROCESS EVALUATION OVERVIEW

The American Evaluation Association defines evaluation as *"assessing the strengths and weaknesses of programs, polices, personnel, products and organizations to improve their effectiveness."*

A process evaluation is defined as: A systematic assessment of an energy efficiency program for the purposes of

1. documenting program operations at the time of the examination, and
2. identifying and recommending improvements that can be made to the program to increase the program's efficiency or effectiveness for acquiring energy resources while maintaining high levels of participant satisfaction (TecMarket Works 2004).

Process evaluations are effective management tools that focus on improving both the design and delivery of energy efficiency programs. They are most commonly used to document program operations for new programs or those in a pilot or test mode. Process evaluations are also effective at diagnosing problems in programs that are under performing or experiencing operational challenges. Since process evaluations most often examine program or portfolio operations, they can identify ways to make program or portfolio enhancements and improvements that reduce costs, expedite delivery, improve satisfaction, and fine-tune objectives. These evaluations can also be used to assess the effectiveness of various incentive structures and examine program operations, they can identify ways to make program enhancements and improvements that reduce overall program costs, expedite program delivery, improve customer satisfaction, and fine-tune program objectives. These evaluations can also be used to assess the effectiveness of various incentive programs and rebated technologies. Process evaluations can also provide feedback on ways to streamline and enhance data collection strategies for program operations (NAPEE 2007).

Process evaluations are driven by the ways in which the end results will be used (NYSERDA 2004). The goal of a process evaluation is to review how program activities and customers interact and to recommend ways to improve program processes to increase effectiveness (NYSERDA 2004; NAPEE 2006).

KEY RESEARCHABLE ISSUES

Process evaluations explore a variety of researchable issues as a way to determine overall effectiveness. The following types of researchable issues can be addressed in process evaluations:

- Determine if the program is meeting its potential to contribute energy supply resources to NY customers
- Determine if the program is filling a key gap in the energy supply structure
- Determine if the program is meeting state supply objectives and meeting energy efficiency supply

opportunities

- Identify future service gaps that can be filled by energy efficiency programs, products and services in NY
- Documenting overall awareness as well as awareness of the program and measures
- Assessing customer satisfaction with the program
- Determining if the program has led to lasting changes in customer behavior regarding energy efficiency actions, searching for energy efficient information, or influencing customer decision-making
- Identifying areas for program improvement
- Program delivery
- Marketing and customer acquisition activities
- The program's role within the portfolio and its overall interactions with other programs
- Reviewing the program's database to determine accuracy and identify areas for program improvement
- Determining the significance of the program's contribution to the New York portfolio

This list is intended to serve as a guideline for the types of information or issues that may be addressed in a process evaluation. However, the specific issues will be determined by the nature of the program, its place in the overall energy efficiency supply policy program portfolio, current operational status, and specific program needs. Specific research issues are explored using a variety of process evaluation tasks and methodologies, which are discussed next.

Process Evaluation Tasks and Methodologies

In order to investigate the researchable issues associated with energy efficiency programs or portfolios, process evaluations involve a wide range of activities. These activities include, but are not limited to, the following: (TecMarket Works 2006)

- Review of Program Materials and Databases
- Review of Program Procedures and Interrelationships (Reynolds et al 2007)
- Staff/Third-Party Program Implementer Interviews
- Key Stakeholder Interviews
- Trade Ally Interviews/Surveys
- Customer Feedback: Surveys/Focus Groups
- Direct Observations/Site Visits

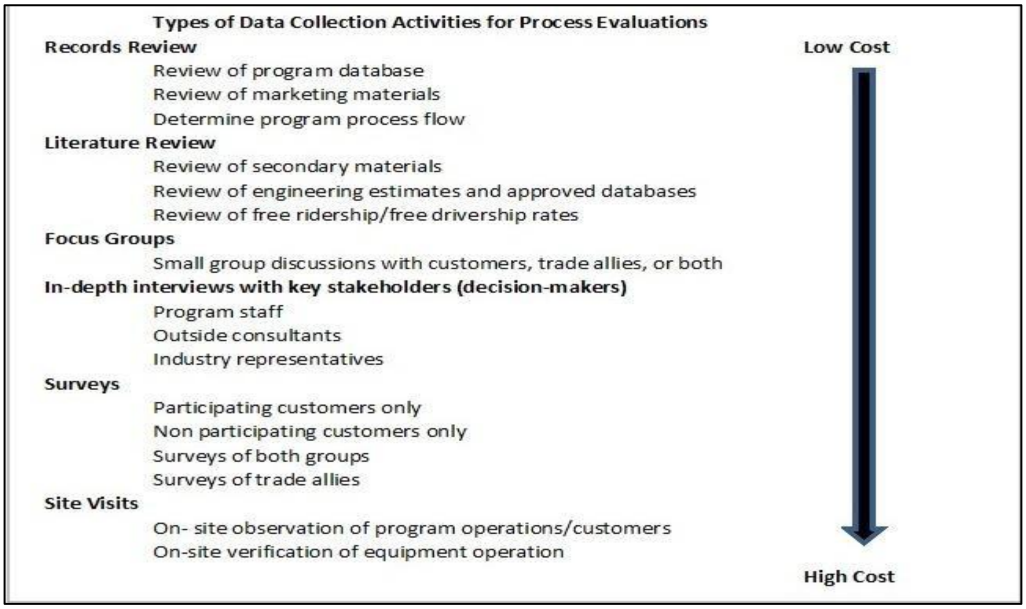
Section II of this document provides guidance for executing these specific process evaluation tactics.

Process Evaluation Budgets

Program evaluation budgets typically range from three to five percent of **overall program budgets**. However, most of these costs are associated with impact evaluations. Smaller utilities face even more intense budget constraints. It is important to remember that while program evaluation budgets may be based on a percentage of the overall energy efficiency spending, the specific cost of conducting a process evaluation will vary depending upon the types of information required and the challenges associated with acquiring this information This will require careful scoping, planning and developing priorities for portfolio and program evaluation needs to ensure that the process evaluations are designed to meet overall state objectives and goals. The first step in conducting effective process evaluations is to prioritize the types of data collection activities that would be conducted, based on the program administrator's specific needs and objectives (Reynolds et al 2007).

Figure 1 summarizes the relationship between cost and level of effort for the various process evaluation activities. For example, it is much less costly to rely on information from secondary sources than it is to gather the data through primary collection methods such as surveys. It is also less expensive to conduct a few in-depth, open-ended surveys with a few respondents, compared to a large-scale survey of many respondents across multiple jurisdictions. The driving factor in determining the costs of surveys is the length of the survey and the number of required surveys based on the desired level of precision and confidence. The current EM&V Guidance recommend 90/10. The most expensive component of data collection is to gather detailed information on-site (Reynolds et al 2007). However, the total costs of moving to a multi- stakeholder planning process should be considered as well

Figure 1: Types of Data Collection Activities for Process Evaluations



Types of Data Collection Activities for Process Evaluations

Records Review
Review of program database
Review of marketing materials
Determine program process flow

Literature Review
Review of secondary materials
Review of engineering estimates and approved databases
Review of free ridership/free drivership rates

Focus Groups
Small group discussions with customers, trade allies, or both

In-depth interviews with key stakeholders (decision-makers)
Program staff
Outside consultants
Industry representatives

Surveys
Participating customers only
Non participating customers only
Surveys of both groups
Surveys of trade allies

Site Visits
On- site observation of program operations/customers
On-site verification of equipment operation

(Source: Reynolds, Johnson & Cullen 2008)

It is important to note that not every process evaluation will require a complete set of data collection activities across all evaluation objectives. Rather, the evaluation plan specifies the data collection strategies that will be used in each phase of the evaluation as well as the anticipated budget expenditures for each data collection activity.

SECTION II: NYS PROCESS EVALUATION PROTOCOLS

The original Process Evaluation Protocols were formulated based on the key findings from the NYS Process Evaluation Literature Review, specifically the key findings and recommendations, and with discussions with the EAG, DPS Staff, and the DPS evaluation advisory team (TecMarket Works). They were designed to provide ongoing feedback for improvements in individual program operations as well provide direction regarding future program design. They are also designed to encourage coordination among the program administrators, where possible, to ensure that process evaluations are conducted in a cost-effective manner. The protocols examine process evaluation activities in two ways:

- Strategic – which focuses on gathering information to provide guidance and direction about future plans including setting or refining policy goals and objectives. Strategic protocols are focused on the overall "big picture" and are designed to ensure that the key findings from the process evaluations are synthesized and analyzed within the context of the broad policy objectives and overall goals of energy program portfolio.
- Tactical-which focuses on the specific process evaluation activities at the program or more granular level. These protocols focus on coordinating research activities and identifying key findings that can lead to program improvements and changes. Ideally, these key findings will be identified as "best practices" that can be further integrated into the overall program portfolio.

In both cases though, the process evaluation findings are designed to provide both feedback and guidance to the individual program administrators, third-party implementers, key decision-makers, and ultimately inform the policy leaders who set the overall goals and objectives for New York State clean energy initiatives.

Program Administrators and their evaluation contractors should review **all the Process Evaluation Protocols** to ensure they are addressing the specific needs for each program.

SECTION II-A: STRATEGIC PROCESS EVALUATION PROTOCOLS

The first set of protocols are defined as "strategic" because they are intended to identify the ways in which a set of process evaluations should work together to provide guidance regarding an entire portfolio of energy efficiency programs. These protocols provide guidance as to how to structure a particular evaluation, whether it is at the portfolio or program level, the decision-making process used to determine if a process evaluation is necessary, and the recommended timing for process evaluations.

PROTOCOL A: Process Evaluation Structure and Timing

Protocol Scope: This protocol provides guidance on how to best structure process evaluations at the state, portfolio, program, service, and market sector level. Process evaluations need to be structured to meet the specific goals and objectives at a particular point in time.

Customer Segments: All

Program Types: All

Approach: The process evaluation decision-maker should determine if a process evaluation is needed, based on any of the criteria described in Protocol A.1 and A.2, which summarize the two major criteria for determining if a process evaluation is necessary. The first criterion is to determine if it is time for a process evaluation; the second criterion is to determine if there is a need for a process evaluation.

Keywords: "timing; portfolio level evaluations; process evaluation structure; diagnostic process evaluations; under-performing programs; programs not meeting targets"

Protocol A.1: Determining Appropriate Timing to Conduct a Process Evaluation
1. New and Innovative Components: If the program has new or innovative components that have not been evaluated previously, then a process evaluation can be useful for assessing their level of success in the current program and their applicability for use in other programs.
2. No Previous Process Evaluation: If the program has not had a comprehensive process evaluation during the previous funding cycle, then the Program Administrator should consider including a process evaluation in the evaluation plan
3. New Vendor or Contractor: If the program is a continuing or ongoing program, but is now being implemented, by a different vendor than in the previous program cycle, then the administrator should consider including a process evaluation in the evaluation plan to determine if the new vendor is effectively implementing the program.
4. State Policy Concerns: Was this program developed as a response to specific policy goals or objectives that now needs to be reviewed or examined?
If any of these criteria are met, it is time to conduct a process evaluation. If none of these criteria are met, then the evaluation decision-maker should proceed to Step 2 in the Process Evaluation Decision Map

Protocol A.2: Determining Appropriate Conditions to Conduct a Process Evaluation
Process evaluations may also be needed to diagnose areas where the program is not performing as expected. These conditions may include the following:
1. Impact Problems: Are program impacts lower or slower than expected?
2. Informational/Educational Objectives: Are the educational or informational goals not meeting program goals?
3. Participation Problems: Are the participation rates lower or slower than expected?

Protocol A.2: Determining Appropriate Conditions to Conduct a Process Evaluation
4. Operational Challenges: Are the program's operational or management structure slow to get up and running or not meeting program administrative needs?
5. Cost-Effectiveness: Is the program's cost-effectiveness less than expected?
6. Negative Feedback: Do participants report problems with the program or low rates of satisfaction?
7. Unusual Practices: Do the program administrators suspect fraud or malfeasance?
8. Market Effects: Is the program not producing the intended market effects?
9. Policy Assessment: Is this program failing to meet an identified service gap or customer segment specially addressed in energy policy objectives?
If any of the criteria is met, a process evaluation is needed to identify ways to address and correct these operational issues. If none of these criteria is met in either Step 1 or Step 2, then a process evaluation is not needed at this time. Re-evaluate the need for a process evaluation at the end of the program year.

(Source: Modified from the CA Evaluators' Protocols TecMarket Works 2006)

PROTOCOL B. Process Evaluation Planning

Protocol Scope: This protocol provides guidance on the key issues that should be addressed in process evaluations. It is especially important to focus on the aspects of program operations to address any deficiencies.

Customer Segments: All

Program Types: All

Approach: The process evaluation plan should use the following outline to identify the key researchable issues that must be addressed in the process evaluation. This outline applies to process evaluations conducted at the program, portfolio, and state level. Process evaluation conducted at a more micro-level may also cover certain elements of this outline.

Keywords: "process evaluation planning; EM&V plan process evaluation timing; portfolio level process evaluations; process evaluation structure; process evaluation components; process evaluation scope"

Protocol B provides more detailed information regarding the key areas for investigation	
Policy Considerations	**Additional Guidance**
• Review of the state policies that led to the development of the energy efficiency portfolio and program	This section provides an opportunity to fully understand the overall program portfolio in the context of overall regulatory policies and goals and therefore can provide guidance for investigating the effectiveness of these policies in the process evaluation.
• Review of state regulatory documents including filings and testimony	
Program Design	**Additional Guidance**
• Program design characteristics and program design process	This area is especially important to address in first- year evaluations and evaluations of pilot programs.
• The program mission, vision and goals and goal setting process	
• Assessment or development of program and market operations theories	
• Use of new or best practices	
Program Administration	**Additional Guidance**
• The program management process	This area should be covered in all process evaluations, bu it is especially important to address in those evaluations where operational or administrative deficiencies exist.
• Program staffing allocation and requirements	
• Management and staff skill and training needs	
• Program tracking information and information support systems	
• Reporting and the relationship between effective tracking and management, including operational and financial management	
Program Implementation and Delivery	**Additional Guidance**

• Description and assessment of the program implementation and delivery process,	This is critical to gathering the information necessary to assess the program's operational flow
• Quality control methods and operational issues	This is an area that should be addressed if the program is not meeting its participation goals or if the program is under-performing.
• Program management and management's operational practices	
• Program delivery systems, components and implementation practices	
• Program targeting, marketing, and outreach efforts	The process evaluator should request copies of all marketing and outreach materials and include an assessment as part of the document review task

(Modified and expanded from the CA Evaluators' Protocols-TecMarket Works 2006)

PROTOCOL C: The Program's Place within the Energy Efficiency Portfolio:

Protocol Scope: This protocol focuses on ensuring that each program is also evaluated within the larger context of the overall energy efficiency program portfolio, which is often overlooked in most process evaluations (Gonzales et al 2003).

Customer Segments: All

Program Types: All

Approach: The process evaluator should review this protocol to ensure that these issues are incorporated in the overall process evaluations planned at either the portfolio or program level. The findings regarding the program's place in the overall portfolio should be reported in the portfolio level evaluation report.

Keywords: "portfolio level evaluations; process evaluation planning; coordination; process evaluation reporting"

Protocol C: The Program's Place within the Energy Efficiency Portfolio
At least once in every funding cycle, the process evaluation results should be compared for all programs within the energy efficiency portfolio to the extent possible.
1. This analysis should include addressing the following researchable issues: a. Does the program fit within New York's energy efficiency state policy goals and objectives fo energy efficiency? Program goals? b. Is the program structured to best contribute to the success of the portfolio within its market segment? c. How is the program meeting its market potential? d. Is it designed to meet some unfilled gap or service need as part of a larger policy goal? e. Are there gaps in program services at the market level (residential / nonresidential)? o Are there overlaps in program offerings that may be causing duplication of efforts or customer confusion?
The key findings and recommendations from this portfolio level analysis should be reported at least once in every funding cycle.

Section II-B: Tactical Protocols

The intent of these tactical protocols is to provide guidance and direction regarding the appropriate use of the most common types of tools in the "process evaluation toolbox." All process evaluations, regardless of the target market or program delivery method will use a mix of these process evaluation methodologies. Where appropriate, the methods best suited to particular types of programs are identified in the process evaluation protocol.

PROTOCOL D: Process Evaluation Work Plan

Protocol Scope: This protocol provides on developing a specific work plan to guide an individual process evaluation. The Process Evaluation Plan, described in Protocol D is intended to be a high-level guidance and planning document. The Process Evaluation Work Plan provides a more detailed description of the key researchable issues, the sampling frame and methodologies, a project schedule and proposed budget and is often the basis of a contractual scope of work between program administrator and evaluation contractor.

Customer Segments: All

Program Types: All

Approach: The process evaluation work plan should be developed after the project kick off meeting to ensure that the research issues are clearly understood and well defined.

Keywords: "process evaluation planning, individual process evaluation"

Protocol D: Recommended Elements of a Process Evaluation Work Plan
1. **Introduction:** The work plan begins with a summary of the program, the process evaluation objectives, and an outline of the specific process evaluation approach that will be used.
2. **Outputs/Outcomes/Indicators Description:** Listing of program outputs/outcomes/indicators, data sources, anc the party responsible for collecting each metric.
3. **Evaluation Scope:** • Synopsis of evaluation including evaluation objectives and main research questions • Evaluation methodology with a detailed sampling plan including proposed sampling methodology and data collection approach for program/third party staff, key stakeholders, trade allies/vendors, and customers, methodologies to be used in the calculation of direct and indirect benefits, and timeframe for long-term data collection. The sampling methodology should be clearly explained with specific targets of completed surveys or interviews clearly described in the work plan. The sampling plans should conform to the current Evaluation Guidelines for confidence and precision.
• **Task Budget, and Schedule:** Identification of key Task, Budget and Schedule milestones including but not limited to a description of the survey instruments that will be developed, timing for data analysis, budget for each key task, timeline and format of deliverables, and staffing and rates per hour.

(Source: Modified and expanded from NYSERDA Draft Work Plans 2011; 2016)

PROTOCOL E: Program Coordination

Protocol Scope: This protocol focuses on identifying ways to effectively plan process evaluations, leverage resources, and report findings in a consistent manner across the entire energy efficiency program portfolio.

Customer Segments: All

Program Types: All

Approach: The process evaluator should review this protocol to ensure that process evaluations activities, such as surveys or onsite visits, are coordinated to avoid respondent fatigue and minimize overall costs. This could be as simple as a single PA coordinating the implementation of multiple process evaluations to minimize the burden on the target audience (e.g., trade allies, program participants). There is also potential for PAs to coordinate research on targeted process related questions (e.g., customer preference for certain type of program delivery mechanics or technologies) or on similar program offerings. This coordinated approach has the potential to reduce costs and offer more rigorous results, but the benefits of a coordinated approach must be balanced against the increased administrative complexities, especially in evaluations involving multiple PAs.

Keywords: "program coordination; portfolio-level evaluations; evaluations across multiple energy providers or program administrators; cost-effective process evaluations: avoiding respondent fatigue; process evaluation budgets"

Protocol E: Program Coordination
1. Where possible, coordinated program or portfolio process evaluations should be considered when: • Multiple programs are targeting similar or identical customer markets • Similar or identical measures are being targeted by multiple programs (e.g., residential lighting; small C&I lighting) • There are a limited number of trade allies serving the same areas (e.g., HVAC installers serve both residential and small C&I markets) or limited populations of other actors targeted for surveys • There are limited evaluation budgets • There are multiple PAs with similar process evaluation related information needs • Multiple PAs are administering similar programs
3. If regional or statewide coordination is not possible, then designing process evaluation survey instruments to capture information as consistently as possible should be considered. This type of collaboration could facilitate expanded analysis and learning across program administrators.

PROTOCOL F: Program Document Review

Protocol Scope: This protocol provides guidance on the types of program information that should be reviewed during a process evaluation. The details of the review process will be influenced by the type of program being evaluated and the study objectives and scope.

Customer Segments: All

Program Types: All

Approach: The process evaluation should include a review of all relevant program materials. At the same time, program administrators should balance the need for review of documents with the desire to be as cost-effective as possible in the use of evaluation funds and deployment of consultant resources.

Keywords: "program records; document review"

Protocol F. Program Document Review	
Scope of Program Document Review	**Additional Guidance**
• Current program records and documents	The program evaluator should request copies of all relevant program materials in a formal data request at the beginning of a process evaluation.
• Relevant Commission orders or filings related to program design and objectives	
• Educational and outreach materials	
• Rebate forms/application materials	
• Program operational manuals and process flow diagrams	
• Website materials	
• Sales data	This information is critical for buy-down programs.
• Retailer store locations	This information is critical for retailer buy-down programs.
• Trade ally contact information	This is critical for equipment replacement programs for residential, multifamily, and C&I programs.
• Program implementer contracts and quality control/quality assurance procedures, and dat tracking requirements	This is critical to determine if the current program practices and procedures are sufficient for documenting program activities.
• Program invoices to the extent possible or payment stream summaries	This is critical for programs using technical assistance, engineering studies, audits, retailer buy-down programs; equipment financing programs.

PROTOCOL G: Assessing Program Flow/Program Inter-Relationships

Protocol Scope: This protocol provides direction on how to document and evaluate the effectiveness of program operations using a program flow chart.

Customer Segments: Residential, Multi-family, Low Income, Institutional, C&I
Program Types: Appliance Recycling, Direct Install, Equipment Replacement, Home Performance, Financing, Rebate, Custom, Technical Assistance/Audit Programs, Pilot Programs

Approach: The process evaluator should review the current program flow materials developed by the program implementer and compare it to the *actual program outcomes*. The resulting flow diagram will be based on findings from the staff and program implementer interviews including a map of the overall database operations and interfaces. The process evaluation should focus on identifying where program "disconnects" or gaps may be. In addition, it should examine the overall flow of program operations to identify possible areas for program improvement or streamlining- especially regarding the integration of information from the application forms.

Keywords: "assessing program operations; program flow; rebate application process; customer interactions"

Protocol G. Assessing Program Flow	
Scope of Program Flow	**Additional Guidance**
Document the ways in which customers interact with the program	These program interactions will vary depending on customer segment and program type; however, it is important to document program flow activities for most programs, except those only dealing with customer outreach and education.
Document the "life of the rebate"	For application driven programs, such as rebate and recycling program, the program evaluator should include questions regarding rebate application processing in interviews with the program implementation staff, trade allies, and program participants. The deliverable should be a flow chart identifying where program gaps may exist.
Identify ways in which the program is promoted to customers	
• Review the program application forms	
• Develop Program Flow Diagram with program "disconnects" highlighted	
• If there appears to be a longer than expected time lag between Technical Assistance Studies and measure installation, then a program flow diagran should be developed to identify where "program disconnects" exist in the decision- making process.	The program flow should be addressed in Custom and Technical Assistance/Audit programs if there seems to b operational lags or difficulties in converting audits to installations.

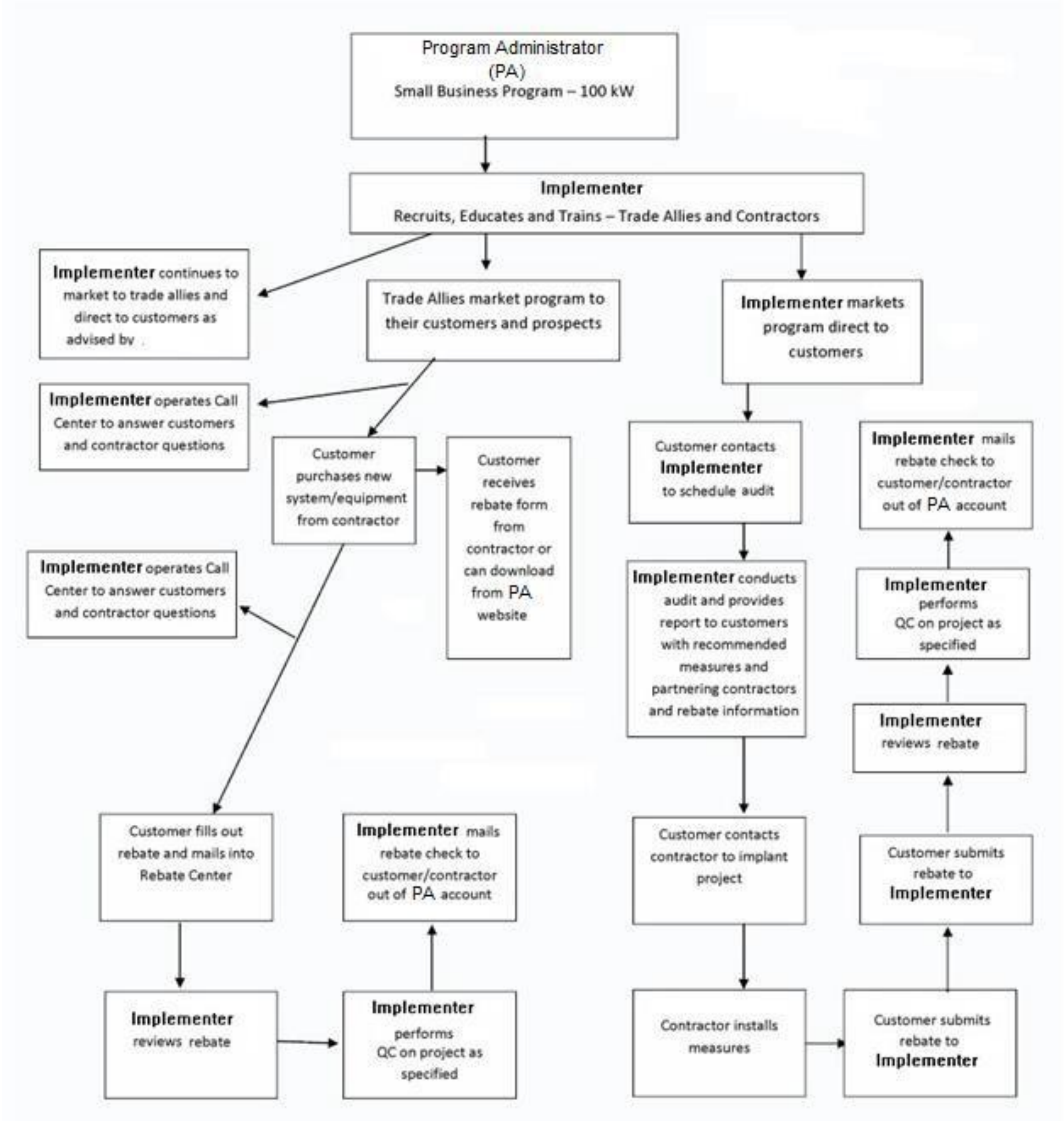

Program Administrator (PA) Small Business Program – 100 kW
Implementer Recruits, Educates and Trains – Trade Allies and Contractors
Implementer continues to market to trade allies and direct to customers as advised by
Trade Allies market program to their customers and prospects
Implementer markets program direct to customers
Implementer operates Call Center to answer customers and contractor questions
Customer purchases new system/equipment from contractor
Customer receives rebate form from contractor or can download from PA website
Customer contacts Implementer to schedule audit
Implementer mails rebate check to customer/contractor out of PA account
Implementer operates Call Center to answer customers and contractor questions
Implementer conducts audit and provides report to customers with recommended measures and partnering contractors and rebate information
Implementer performs QC on project as specified
Implementer reviews rebate
Customer fills out rebate and mails into Rebate Center
Implementer mails rebate check to customer/contractor out of PA account
Customer contacts contractor to implant project
Customer submits rebate to Implementer
Implementer reviews rebate
Implementer performs QC on project as specified
Contractor installs measures
Customer submits rebate to Implementer

Figure 2: Example of a Program Flow Diagram

PROTOCOL H: Staff/Third Party Program Implementation Interviews

Protocol Scope: This protocol provides guidance on the information that should be documented in interviews with program staff or implementers.

Customer Segments: All

All Program Types: All

Approach: Protocol H summarizes the types of information that should be addressed in the in-depth interviews conducted with both the program administrators and program implementers.

Keywords: "assessing program operations; program flow; staff interviews; process evaluation planning"

Protocol H: Staff/Third Party Implementer Interviews	
In-Depth Interview Scope	**Additional Guidance**
In-depth interviews should be conducted either in person or via the telephone with all program staff involved directly in program operations. Types of staff interviewed include: Program manager and implementation lead Key staff involved in program operations, customer database tracking, marketing and outreach, and customer relations	The interviews may also be conducted with program personnel who were involved in the original program design, if this is the first process evaluation conducted for this program.
2. The estimated number of completed interviews should be clearly described based on the Process Evaluation Plan.	Deviations from the estimated number of completed interviews must be explained in the Process Evaluation Report.
3. The interview guide should be designed to be "open ended" to facilitate a discussion.	This is to ensure that the interview guides will be flexible enough to probe on specific issues that may arise during the staff or third party interviews.
4. All respondents must be assured anonymity to the legal extent possible. There will also be no direct attribution of their responses that can easily identify the respondent.	This is to ensure that respondents can speak candidly and honestly about program activities.

<table>
<tr><th colspan="2">Protocol H: Staff/Third Party Implementer Interviews</th></tr>
<tr><th>In-Depth Interview Scope</th><th>Additional Guidance</th></tr>
<tr><td>The topics addressed in these interviews should include the following:
Program Design/Goals & Objectives
Program Results Relative to Goals
Data Tracking/Program Databases
Marketing and Outreach Activities
Participation Process
Barriers to Program Participation
Assessment of Program Operations
Customer Feedback
Trade Ally/Vendor Feedback
Areas for Program Improvement</td><td>The focus of these topic areas will vary according to the type of process evaluation conducted. For those process evaluations that are investigating operational deficiencies, particular attention should be paid to documenting program operations, participation process, feedback from customers and trade allies, and areas for program improvement.</td></tr>
</table>

PROTOCOL I: Trade Ally/Vendor Interviews

Protocol Scope: Protocol H provides guidance on the types of information that should be addressed in interviews conducted with trade allies. The type and number of interviews should be documented in the process evaluation plan.

Customer Segments: Residential, Multi-family, C&I, Institutional

Program Types: Appliance Recycling, Direct Install, Equipment Replacement, Home Performance, Financing, Residential Pilot Program, Rebate Programs, C&I Pilot Programs; Retailer Buy-Down Programs

Approach: Protocol I identifies the types of information that should be addressed in interviews with both participating and non-participating trade allies.

Keywords: "trade allies; installers; vendors; C&I programs; interviews; surveys; process evaluation planning

Protocol I: Trade Ally/Vendor Interviews	
Interview Scope	**Additional Guidance**
Interviews should be conducted either in person or via the telephone with trade allies who sell or install the eligible equipment or services.	Depending upon the scope of the process evaluation, the interviews could be conducted with both participating and non-participating trade allies.
The estimated number of completed interviews should be clearly described based on the Process Evaluation Plan.	Deviations from the estimated number of completed interviews must be explained in the Process Evaluation Report.
The trade ally survey instrument should include a mix of "open" and close-ended questions to facilitate analysis.	The nature of the process evaluation inquiry will determine the appropriate type of survey instrument as well as the length of the interview.
All respondents must be assured anonymity to the legal extent possible. There will also be no direct attribution of their responses that can easily identify the respondent.	This is to ensure that respondents can speak candidly and honestly about program activities.
The topics addressed in these interviews should include the following: Program Awareness Participation Process Trade Ally/Vendor Satisfaction with Program Components Barriers to Program Participation by Customers Barriers to Trade Ally participation Overall Effectiveness of Trade Ally Activities Areas for Program Improvement Trade Ally characteristics, such as type of firm, number of employees, number of jobs completed, etc.	The focus of these topic areas will vary according to the type of process evaluation conducted. Surveys to non-participating contractors will focus primarily on program awareness, barriers to program participation, and trade ally characteristics.

PROTOCOL J.1: Customer Surveys

Protocol Scope: Protocol J summarizes the types of information that should be addressed in the customer surveys that will be targeting mass-market programs.

Customer Segments: Residential, Multi-family, Low Income Customer Segment

Program Types: Appliance Recycling, Direct Install, Equipment Replacement Programs, Home Performance, Financing, Residential Pilot Programs, Education and Outreach Programs, Online Audits, Rebate Programs.

Approach: Protocol J.1 provides guidance on the types of information that should be captured in residential customer surveys.

Keywords: "residential customer surveys; participating customers; non-participating customers; process evaluation planning"

Protocol J.1: Residential Customer Surveys	
Interview Scope	**Additional Guidance**
Interviews should be conducted either in person or via the telephone with customers eligible to participate in the program.	Depending upon the scope of the process evaluation, the interviews could be conducted with both participating and non-participating customers to facilitate comparisons between groups and identify areas for program improvement.
The estimated number of completed interviews should be clearly described based on the Process Evaluation Plan.	Deviations from the estimated number of completed interviews must be explained in the Process Evaluation Report.
The customer survey instrument should include a mix of "open" and close-ended questions to facilitate analysis.	The nature of the process evaluation inquiry will determine the appropriate type of survey instrument as well as the length of the interview.
All respondents must be assured anonymity to the legal extent possible. There will also be no direct attribution of their responses that can easily identify the respondent.	This is to ensure that respondents can speak candidly and honestly about program activities.

Protocol J.1: Residential Customer Surveys	
Interview Scope	**Additional Guidance**
The topics addressed in these interviews should include the following: Program Awareness Participation Process Customer Satisfaction with Program Components (Participants Only) Measure Persistence Spillover if appropriate Barriers to Program Participation Areas for Program Improvement Customer Demographics such as housing type, square footage, number of occupants, income level, educational level and age range.	The focus of these topic areas will vary according to the type of process evaluation conducted. Surveys to non-participating customers will focus primarily on program awareness, barriers to program participation, and customer demographics.

PROTOCOL J.2: Customer Surveys

Protocol J2: Protocol J.2 summarizes the types of information that should be addressed in the customer surveys that will be targeting mass-market programs.

Customer Segments: Small C&I Programs, Institutional Customer Segment

Program Types: Equipment Replacement Programs, Financing, Small C&I Pilot Programs, Education/Outreach Programs, Energy Audits Programs"

Approach: Protocol J.2 provides guidance on the types of information that should be captured in surveys targeting Small C&I customers or institutional segments such as educational facilities, hospitals, or government buildings.

Keywords: "participating customers; non-participating customers; small C&I programs; process evaluation planning"

Protocol J.2: Customer Surveys	
Interview Scope	**Additional Guidance**
Interviews should be conducted either in person or via the telephone with customers eligible to participate in the program.	Depending upon the scope of the process evaluation, the interviews could be conducted with both participating and non-participating customers to facilitate comparisons between groups and identify areas for program improvement.
The estimated number of completed interviews should be clearly described based on the Process Evaluation Plan.	Deviations from the estimated number of completed interviews must be explained in the Process Evaluation Report.

The customer survey instrument should include a mix of "open" and close-ended questions to facilitate analysis.	The nature of the process evaluation inquiry will determine the appropriate type of survey instrument as well as the length of the interview.
All respondents must be assured anonymity to the legal extent possible. There will also be no direct attribution of their responses that can easily identify the respondent. .	This is to ensure that respondents can speak candidly and honestly about program activities.
The topics addressed in these interviews should include the following: Program Awareness Participation Process Customer Satisfaction with Program Components (Participants Only) Measure Persistence Spillover as appropriate Barriers to Program Participation Areas for Program Improvement Customer Demographics such as type of business, number of employees, hours of operation, building square footage, number of years in business.	The focus of these topic areas will vary according to the type of process evaluation conducted. Surveys to non-participating customers will focus primarily on program awareness, barriers to program participation, and customer demographics.

PROTOCOL K: On-Site Visits/Direct Observations

Protocol Scope: Protocol K summarizes the types of information that should be addressed when it is necessary to conduct direct observation of program operations. These types of programs may include those that require substantial on-site interaction with the participants involving measure installations, for example, such as the installation of custom measures.

Customer Segments: Residential, C&I, Multi-family, Institutional, Low Income

Program Types: Retailer Buy-Down Programs, Customer Direct Install Programs, Give- Away Programs, Custom Measure Installations

Protocol Approach: On-site observations may also be included as part of "intercept" studies at retail stores, especially for residential lighting programs. Other times, the evaluator may want to verify the installation rates of self-installed measures, such as through energy efficiency kits, especially when savings estimates are heavily dependent on self-reporting by program participants (California Evaluators' Protocols TecMarket Works 2006; Johnson Consulting 2011).

Keywords: "on-site visits self-reported installations; customer intercept surveys; give- away programs; case studies; process evaluation planning"

Protocol K: On-Site Visits/Field Observations	
Interview Scope	**Additional Guidance**
The Process Evaluation Team should develop an on- site form specifically designed to capture the observations from these visits.	On-site visits/field visits are usually targeted to program participants.
The estimated number of completed on-site visits should be clearly described based on the Process Evaluation Plan.	Deviations from the estimated number of completed interviews must be explained in the Process Evaluation Report.
The on-site survey instrument should be designed to capture actual descriptions of respondent activities.	The nature of the process evaluation inquiry will determine the appropriate type of onsite visits required to complete the process evaluation.
Direct questions to the program participants should be identically worded as in other survey instruments to ensure content validity.	This "best practice" will allow the process evaluator to compare results across groups, thereby increasing the validity and robustness of the overall process evaluation.
All respondents must be assured anonymity to the legal extent possible. There will also be no direct attribution of their responses that can easily identify the respondent.	This is to ensure that respondents can speak candidly and honestly their experiences with the program and its measures.

The topics addressed in on sites should include the following: • Program Awareness • Experience with the Program Participation Process • Customer Satisfaction with Program Components • Measure Persistence • Spillover as appropriate • Barriers to Program Participation • Areas for Program Improvement • Customer Demographics for C&I customers such as: type of business, number of employees, hours of operation, building square footage, number of years in business. • Customer Demographics such as housing type, square footage, number of occupants, income level, educational level and age range.	The focus of these topic areas will vary according to the type of process evaluation conducted.
The Process Evaluator should also document the actual direct observations, physical location or type of measure installed as part of this on site.	This information should be collected to provide further validation of the actual measure installation rate or participant experience.

(Source: Expanded and Modified from California Evaluators' Protocols TecMarket Works 2006; Johnson Consulting Group 2011).

PROTOCOL L.1: Actionable Recommendations

Protocol Scope: Protocol L.1 provides guidance regarding the ways in which recommendations for program improvement should be structured.

Customer Segments: All

Program Types: All

Protocol Approach: The recommendations made from the process evaluations are detailed, actionable and cost-effective. They should identify a clear path for program improvement and specifically tie back to the researchable issues cited in the process evaluation plan.

Keywords: "process evaluation reporting; portfolio-level evaluations; regulatory guidance; policy guidance; findings and recommendations; operational changes"

Protocol L.1: Actionable Recommendations
1. The recommendations from the process evaluations should be realistic, appropriate to the organization's structure, constructive, and achievable using available resources.
2. The recommendations should be linked to specific conclusions.
3. All recommendations need to be adequately supported. Each recommendation should be included in the Executive Summary and then presented in the Findings text along with the analysis conducted and the theoretical basis for making the recommendation. The Findings section should also include a description on how the recommendation is expected to help the program, including the expected effect implementing the change will have on the operations of the program.
4. The recommendations should focus on ways to increase overall program effectiveness and be linked to the researchable issues addressed in the process evaluation such as ways to improve the program design, approach, operations, marketing, or address issues related to program under-performance.
5. To the extent possible, the recommendations will provide specific steps/tasks for implementation.
6. To the extent possible, the Program Administrator will offer specific steps or tasks for implementing the recommendations.
7. The recommendations should be compared across program evaluations to identify areas for portfolio-level improvements.
8. Recommendations should be prioritized by the process evaluator so as to present a reasonable number of recommendations that are truly actionable and important to the efficiency and effectiveness of the program.

(Source: Modified from the CA Evaluators' Protocols, TecMarket Works 2006; Peters 2007; NYSERDA 2011)

PROTOCOL L.2: Tracking/Follow Up for Actionable Recommendations

Protocol Scope: Protocol L.2 provides additional guidance regarding the ways in which recommendations for program improvement should be tracked and monitored over time. Customer

Segments: All

Program Types: All

Protocol Approach: The recommendations also need to be reviewed and monitored in order to ensure they are implemented or documented in subsequent process evaluations why they were not implemented.

Keywords: "process evaluation reporting; portfolio-level evaluations; regulatory guidance; policy guidance; findings and recommendations; operational changes"

Protocol L.2: Tracking/Follow-Up For Actionable Recommendations
1. Each program and portfolio-level recommendation should be documented and tracked over time.
2. Recommendations should also identify "best practices" that may benefit other program operations at the portfolio or statewide level.
3. The status of each prior recommendation should be documented, indicating if they had been implemented, are in the process of being implemented, no longer feasible to be implemented, or cannot be implemented.

(Source: NYSERDA 2011)

SECTION II-C: SPECIAL PROCESS EVALUATION TOPICS

Protocols M and N address special issues that arise in process evaluations.

PROTOCOL M: Pilot Program Evaluations

Protocol Scope: Protocol L describes the recommended procedures for conducting process evaluations of pilot programs.

Customer Segments: All

Program Types: Pilot Programs

Protocol Approach: Pilot programs are trials that need to be closely monitored to understand their effectiveness and ability to scale up, if successful. Process evaluations, or process evaluation approaches, can be helpful in assessing the early performance of a pilot and should be applied where they add value.

Keywords: "process evaluation reporting; portfolio-level evaluations; regulatory guidance; policy guidance; findings and recommendations; operational changes"

Protocol M: Pilot Program Evaluations
1. Pilot programs may include a process evaluation conducted within the first year of program operations to provide early feedback and identify ways to correct operational issues.
2. Pilot programs may also have a process evaluation conducted at the end of the pilot program period to document success or failure and provide a complete history of pilot program activities and outcomes, and support the decision on whether to ramp up the program.
3. Pilot programs that introduce a new technology or approach may also benefit from a process evaluation by demonstrating the applicability of this approach in novel businesses and aiding in the dissemination of this new concept.

PROTOCOL N: Satisfaction

Protocol Scope: Protocol N focuses on some ways to assess overall satisfaction from customers, trade allies, and key stakeholders.

Customer Segments: All

Program Types: Appliance Recycling, Direct Install, Equipment Replacement, Rebate Programs, Financing Programs, Home Performance Programs, Technical Assistance/Engineering Studies/Pilot Programs, Custom Programs and any other programs with customer interactions.

Protocol Approach: Since participant satisfaction is measured in a variety of ways within individual energy organizations, this protocol provides guidance on identifying ways to assess customer satisfaction by investigating not just the direct determinants of satisfaction but also examining the underlying drivers of satisfaction as they relate to the inter- relationships among the participant, the product or service offered and the organization providing the product or service (Hall and Reed 1997).

Keywords: “participating customers; non-participating customers; customer satisfaction; trade ally satisfaction; participant interaction; program experience; under-performing programs”

Protocol N: Satisfaction	
Protocol Scope	**Additional Guidance**
• Satisfaction with the Program is an important element of process evaluation and therefore may be addressed in surveys or interviews with program participants and key stakeholders, including vendors and trade allies.	This topic should be explored through a series of questions to assess satisfaction with the participants’ program experience.
• A consistent satisfaction scale should be used in all customer surveys fielded through the portfolio-level evaluations.	This will ensure that satisfaction ratings can be compared among and between programs in a robust manner.
• Satisfaction should be analyzed in the following ways: • Satisfaction with the various program components (i.e., rebate application process; trade ally interactions; customer service interactions, etc.)	The specific components of customer satisfaction will differ across program evaluations, but the major elements of a customer experience should be addressed in the customer assessment of satisfaction with both the program and the program providers.
• All respondents should report their overall satisfaction ratings with the program and the program provider.	This will allow for comparisons of satisfaction among program participants and non-participants and facilitate analysis over time.
• The customer surveys or interviews should include “open ended” questions that probe for major reasons reported by participants for “dissatisfaction” with the program, for any response under a 7 or 8 on a 10-point scale or under a 5 on a 5-point scale.	This will help to determine the underlying reasons for satisfaction that may not be addressed in a closed-question survey.

(Source: Hall & Reed 1997; Peters 2007).

REFERENCES

1. Gonzales, P., Barata, S, Megdal, L & Pakenas, L, 2003. *"Portfolio Process Evaluation: An Enhanced Perspective with Cost-Efficient Options, "*Proceedings of the 2003 IEPEC Conference, August, Seattle, WA.
2. Hall & Reed, J. 1997. *"Methods for Measuring Customer Satisfaction,"* Proceedings of the 1997 International Energy Program Evaluation Conference, Chicago, IL. pp. 23-33.
3. Johnson Consulting Group, 2011. *"Evaluation, Verification & Measurement Plan for Home Audit Program,"* Columbia Gas of Virginia, Chester, VA. April.
4. *National Action Plan for Energy Efficiency (NAPEE) Action Plan and Resource Guides for Process, Impact Evaluations and Understanding Cost-Effectiveness of Energy Efficiency Programs*, 2007. DOE/EPA, November.
5. New York State Energy Research and Development Authority (NYSERDA), 2004, *New York Energy Smart Program Evaluation and Status Report Executive Summary,* Volume 1. May, 2011, *New York's System Benefits Charge Programs Evaluation and Status Report Year Ending December 31, 2010,* Report to the Public Service Commission- Final Report, March, p. E-21., 2011, *Report Format and Style Guide*, January.
6. Peters, J., 2007.a *White Paper*: *Lessons Learned After 30 Years of Process Evaluation*, Research Into Action.
7. , Baggett, S, Gonzales, P, DeCotis, P, & Bronfman, B. 2007. *How Organizations Implement Evaluation Results*, Proceedings of the 2007 International Energy Program Evaluation Conference, Chicago, IL.
8. Reynolds, D, Johnson, K & Cullen, 2007. G. "*Best practices for developing cost-effective evaluation, measurement, and verification plans: lessons learned from 12 northern California municipal program administrators*," Proceedings from Association of Energy Services Professionals Conference, San Diego, CA.
9. TecMarket Works 2004. *The California Evaluation Framework,*
10. TecMarket Works, 2006. *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals,* Under Contract with and Directed by the CPUC's Energy Division, and with guidance from Joint Staff, April.
11. Zikund, W. 2000, *Business Research Methods-6th Edition*, The Dryden Press, Fort Worth, TX.

Appendix F: Conventional Resource Acquisition Program Impact Evaluation

In the REV Track II Order[18], the Commission describes the complexity and challenges related to utility shareholder incentive mechanisms that depend on a determination of what would have taken place in the absence of the program, in other words proving of a counterfactual. These include administrative challenges, contentious ex-post review processes, and the potential for tremendous administrative expense for uncertain net benefit. While this discussion is specific to the development of future electric utility earning adjustment mechanisms (EAMs), consideration should be given to the value of undertaking net-savings evaluations and the intended use of the output of such work. Since ratepayer funded activities are not geared toward actions that would have occurred anyway, some level of net savings evaluation, or examination of program influence, may inform program design. However, given the number and variety of activities occurring in the marketplace, including Commission direction for NYSERDA and utility offerings to become more complementary, it will be increasingly more difficult to parse out the effects of any one specific program action.

Mindful of the above context, there are still resource acquisition programs operating in New York for which program administrators may seek to address net-to-gross elements in their evaluations. This appendix includes three major sections on resource acquisition program impact evaluation:

1. Program-Level Participant and Nonparticipant Spillover
2. Estimating Net-to-Gross Ratios Using Self-Report Approach
3. Calculating the Relative Precision of Program Net Savings

These sections provide accurate and helpful guidance for program administrators in conducting conventional net-to-gross evaluation of resource acquisition programs.[19] While this guidance provides significant detail on full program evaluations, it does not preclude rolling or targeted evaluation of elements such as freeridership. Program administrators have the flexibility to utilize this approach and guidance as appropriate to meet their particular program evaluation needs.

1. Program-Level Participant and Nonparticipant Spillover

The purpose of this Appendix is to provide certain methodological principles regarding the reliable estimation of spillover savings, i.e., estimates that are reasonably precise *and* accurate.
Spillover is defined as:

> . . . the energy savings associated with energy efficient equipment installed by consumers who were influenced by an energy efficiency program, but without direct financial or technical assistance from the program. Spillover includes additional actions taken by a program participant as well as actions undertaken by non-participants who have been influenced by the program.

[18] Case 14-M-0101, Proceeding on Motion of the Commission in Regard to Reforming the Energy Vision Order Adopting a Ratemaking and Utility Revenue Model Policy Framework (REV Track II Order), issued May 19, 2016.
[19] These sections were previously included as appendices in Staff's former Evaluation Guidelines and have been retained in this Guidance for informational purposes, in contexts where applicable.

This definition is consistent with the somewhat more detailed definition contained in the California Energy Efficiency Policy Manual (2008):

> Reductions in energy consumption and/or demand in a utility's service area caused by the presence of the DSM program, beyond program related gross or net savings of participants. These effects could result from: (a) additional energy efficiency actions that program participants take outside the program as a result of having participated; (b) changes in the array of energy-using equipment that manufacturers, dealers and contractors offer all customers as a result of program availability; and (c) changes in the energy use of non-participants as a result of utility programs, whether direct (e.g., utility program advertising) or indirect (e.g., stocking practices such as (b) above or changes in consumer buying habits)." Participant spillover is described by (a), and nonparticipant spillover, by (b) and (c). Some parties refer to non-participant spillover as "free-drivers." (TecMarket Works Team, 2006)

Some evaluators subdivide participant spillover into "inside" and "outside" spillover. Inside spillover occurs when, due to the project, additional actions are taken to reduce energy use at the same site, but these actions are not included as program savings. Outside spillover occurs when an actor participating in the program initiates additional actions that reduce energy use at other sites that are not participating in the program.[20]

Because causality is inherent in the very definition of spillover, the spillover savings are inherently net.

Free ridership and spillover are captured in the net-to-gross (NTG) ratio to reflect the degree of program-induced actions. Specifically, the gross energy savings estimate, refined by the realization rate, is adjusted to reflect the negative impacts of free ridership and the positive impacts of spillover. Equation 1 illustrates this adjustment.

Equation 1

NTG ratio = (1-Free ridership) + Spillover

Clearly, ignoring spillover results in a downward bias in the NTG ratio.

[20] It is worth noting that one implication of all of these definitions is that how a piece of savings is classified may depend in part on the objectives of the program and what outcomes the program has chosen to track. As a key example, program influence achieved through the provision of technical information (henceforth called information-induced savings for shorthand) is clearly a legitimate source of savings, but, depending on the specifics of the situation, could end up being classified either as in-program savings, participant spillover, or non-participant spillover. If the provision of information is considered sufficiently central to the program objectives for the program to directly track this outcome, then information-induced measures may be classified as in-program savings. If information-induced measures are not tracked but are adopted by participants who also adopted rebated measures, and thus entered the tracking system, then they may end up being classified as participant spillover. If untracked information-induced measures are adopted by end-users who did not also adopt a measure for which they received a rebate, then they may be classified as non-participant spillover. While all of this suggests that the precise meaning of these terms can be somewhat specific to the situation, this document is intended to provide methodological guidance that is resilient in the face of such distinctions.

This Appendix provides general guidelines for estimating both participant and nonparticipant spillover.[21]

KEY DECISIONS FOR EVALUATORS

Before evaluators decide to estimate spillover, they must make a number of critical decisions:

- Will the evaluation address participant spillover, nonparticipant spillover, or both?
- Does the size of the expected savings warrant the expenditure of evaluation funds needed to estimate these savings at an appropriate level of reliability?
- Which of the two levels of methodological rigor discussed in these guidelines, *standard* or *enhanced*, should be used?
- Will spillover be estimated based on data collected from end users, those upstream from end users (e.g., vendors, installers, manufacturers, etc.), or both?
- What is the level of aggregation? Although participant spillover is always estimated at the program level, if an evaluator is attempting to estimate nonparticipant spillover, will the evaluator estimate it at the program level or the market level? One potential reason for estimating nonparticipant spillover at the market level is that, in some circumstances, reliably teasing out the spillover savings attributable to one specific program among many may be nearly impossible due to the difficulty nonparticipants may have in attributing any of their installations to a specific program. In such a case, evaluators can choose to conduct market effects studies which include naturally occurring adoptions, program-rebated adoptions, participant and nonparticipant spillover, other program effects that cannot be reliably attributed to a specific program (e.g., upstream lighting programs and the effects of the portfolio of programs on such things as increases in the allocation of shelving space to efficient measures), and other non-program effects due to such factors as DOE Energy Star, programs funded by the American Recovery and Reinvestment Act (ARRA) and the gradual non-program induced evolution of the market in terms of attitudes, knowledge and behavior regarding energy efficiency. The net savings resulting from market effects studies must be included in the portfolio-level benefits-costs analyses.
- If an evaluator decides to conduct a market effects study, then they must decide whether the study should be focused on the region targeted by a given PA, multiple regions or even the entire state.

Once these questions are answered, evaluators can then use these guidelines in estimating spillover.

PROGRAM-SPECIFIC METHODS

Level of Rigor

Various types of spillover can be estimated using data collected from participating and nonparticipating end users and from participating and/or nonparticipating market actors upstream from the

[21] While the spillover guidance provided in this Appendix focuses entirely on estimating benefits, PAs should not forget that they must also estimate the incremental costs associated with each spillover measure. Both the benefits and costs of spillover measures must be included in the total resource cost (TRC) test and the societal test.

end users (e.g., vendors, retailers, installers, manufacturers). These savings can also be estimated at varying levels of methodological rigor. Program administrators should propose whether a given spillover analysis should receive *standard* or *enhanced* treatment. The primary criterion for whether a given spillover analysis is subject to standard vs. enhanced requirements is the expected magnitude of spillover savings. Factors that the PAs should consider in making a decision regarding treatment include:

- Past results for the same PA program
- Program theory or market operations theory
- National research literature for similar programs.
- Size of the program
- Size and complexity of the market
- Nature of the technology(ies) promoted by the program
- Cost of standard versus enhanced treatment

Table 1 presents the standard and enhanced levels of rigor for estimating both gross spillover savings and program influence for both end users and those upstream from the end users.

Table 1. Level of Methodological Rigor for Estimating Gross Spillover Savings and Program Influence

	Standard Rigor	**Enhanced Rigor**
Overall Methodological Approach	May rely solely on self-reports from end-users and upstream market actors to support estimates of gross savings or program influence.	Basic self-reports from end-users and upstream market actors typically not sufficient as sole method to support estimates of gross savings or program influence
Estimation of average gross savings for spillover measures for end users (participants and/or nonparticipants).	Simplifying assumptions may be made, such as average gross unit savings being the same for spillover measures as for in-program measures.	Average gross unit savings for spillover measures, insofar as practical and cost effective should be documented empirically, based on a combination of self-reports and/or on-site visits.
Estimation of gross savings from upstream actors (participants and/or nonparticipants).	Self-reports generally sufficient.	Researchers must attempt to confirm self-reports using methods such as changes in sales, stocking or shipment data, review of planned or completed project or permits, or on-sites.
Estimation of program influence for end users (participants and/or nonparticipants).	Basic self-reports generally sufficient.	Enhanced self-reports generally sufficient[22].
Estimation of program influence for upstream actors (participants and/or nonparticipants).	Basic self-reports generally sufficient.	Either additional methods such as quasi-experimental design, econometric analysis, or Delphi panels[23] should be deployed or a case should be made that

[22] Basic self-reports typically involve interviewing one participant decision-maker or market actor. Enhanced self-reports on the other hand typically involve more intensive data collection and analysis in the estimation of the net-to-gross ratios. For example, it can include collecting data from more than one participant decision-maker as well as from others such as relevant vendors, retailers, installers, architectural and engineering firms, and manufacturers. It can also include the consideration of past purchases and other qualitative data gleaned from open-ended questions.

[23] Delphi panels can be useful as long as members are provided sufficient market-level empirical data to inform their deliberations. Delphi panels should not be confused with brainstorming.

	Standard Rigor	Enhanced Rigor
		such methods are either not viable or not cost-effective.
Documentation of causal mechanisms	Recommended but not required.	Required, using methods such as self-reports from end-users or market actors regarding the manner in which the program influenced their behavior, and/or theory-driven evaluation practices. [24]

Double Counting

Program administrators should propose methods to avoid double counting both participant spillover and nonparticipant spillover. For example, some participant or nonparticipant spillover measures might have received assistance (information and/or incentives) from some other program administrators' programs. In such cases, measures receiving assistance from other program administrators' programs should be subtracted for the spillover estimates. Or, in other cases, two programs could be targeting the same market for the same measures. In such cases, because it would be challenging to accurately allocate spillover savings attributable to each program, expert judgment may be used. Under no circumstances, when the possibility of double counting exists, should a program administrators claim the sum of the spillover savings separately estimated for each program without making the appropriate adjustments. Determining how the estimated spillover savings should be allocated among different programs within a given program administrators' portfolio and/or across program administrators' portfolios can be based on such factors as the size of the program budgets, program theories and logic models that demonstrate the causal mechanisms that are expected to produce spillover, and the results of theory-driven evaluations (Weiss, 1997; Donaldson, 2007).

Program administrators should consider where top-down and market-level studies can be used in place of program/initiative specific inquiries to help avoid or mitigate the issue of double counting of benefits.

Calculation of the Program-Level Spillover Rate

While PAs are free to calculate spillover rates in a variety of ways, the formulation of a NTG ratio presented in the Guidelines is repeated in Equation 2:

Equation 2

$$\mathit{NTG\ Ratio} = \qquad (1 - \mathit{Free\ Ridership}) + \mathit{spillover}$$

[24] Documentation of causal mechanisms can include verification of the key cause and effect relationships as illustrated in the program logic model and described in the program theory. Weiss (1997, 1998) suggests that a theory-driven evaluation can substitute for classical experimental study using random assignment. She suggests that if predicted steps between an activity and an outcome can be confirmed in implementation, this matching of the theory to observed outcomes will lend a strong argument for causality: "If the evaluation can show a series of micro-steps that lead from inputs to outcomes, then causal attribution for all practical purposes seems to be within reach" (Weiss 1997, 43).

Equation 2 illustrates that the spillover rate is added to 1-Free Ridership to produce the NTGR. Given the additive nature of the spillover rate in Equation 2, the spillover rate must be calculated as in Equation 3:

Equation 3

$$\textit{Spillover Rate} = \frac{\textit{Net PSO+ Net NPSO}}{\textit{Ex Post Gross Program Impacts}}$$

Estimating Spillover at the Market Level

In some cases, it might not be possible to reliably estimate nonparticipant spillover at the program level due to multiple program interventions in the same market involving multiple market actors. In such cases, market effects studies can be performed for specific measures and markets, e.g., commercial lighting combined with trade ally training. This Appendix does not provide any guidelines for conducting such studies, but rather refers evaluators to other sources such as Eto, Prahl, and Schlegel (1996), Sebold et al. (2001) and TecMarket (2005).

Sampling and Uncertainty

Sampling for both program-level and market level spillover studies should be performed in accordance with the *Sampling and Uncertainty Guidelines* in Appendix B.

Levels of Confidence and Precision

As discussed in the main body of the DPS guidelines, the minimum standard for confidence and precision for overall net savings at the program level is 90/10. Here, overall net savings includes both in-program net savings and any reported spillover savings. The achieved level of confidence and precision for overall program net savings must be reported at the 90% level of confidence.

If reported savings results include spillover savings, there is no required level of confidence and precision specifically for the individual components of net savings from in-program measures and net savings from spillover. However, PAs are still accountable for achieving 90/10 for overall program net savings, unless an alternative was proposed and is documented. The standard error of overall program-level net savings can be calculated by combining the achieved levels of confidence and precision for the net savings from in-program measures and for spillover savings using standard propagation of error formulas (Taylor, 2006; TecMarket, 2004).25 While there are no precision requirements for the individual

[25] This is generally true as long as each of the individual components making up the total net savings estimate (e.g., gross savings, free riding, spillover, etc.) has been estimated based on independent random samples and methods that allow for the calculation of standard errors. However, there are legitimate circumstances under which the sample designs and methods for one or more components do not meet these requirements. One example is a market effects study in which total net program impacts are estimated using a preponderance of evidence approach. Another example (some aspects of which are discussed in the next section) is a case in which one or more components are deemed. A third example is a case in which multiple methods are used to estimate net impacts or the net-to-gross ratio, and a Delphi analysis is used to integrate the results. If *none* of the individual components meet these requirements, then clearly the issue of precision does not apply. If some components meet these requirements but others do not, then the program administrator should take clear note of this fact and propose an

components of net savings from in-program measures and the net savings from spillover measures, the precision actually achieved for each of these components should be reported at the 90% level of confidence, in order to help facilitate assessment of the reliability of the results.

Deemed Approaches

Of course, there might be situations in which all key stakeholders are willing to agree that spillover is not zero, but the expense to estimate it reliably is prohibitive. In such cases, a PA may utilize a deemed spillover rate based on a review of the literature and the program theory and logic model, that together describe reasonably well the causal mechanism that is expected to generate spillover.

References

1. California Public Utilities Commission: Energy Division and the Master Evaluation Contractor Team. (2007). *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches*.
2. Donaldson, Stewart I. (2007). *Program Theory-Driven Evaluation Science: Strategies and Applications*. New York: Psychology Press.
3. Eto, Joseph, Ralph Prahl and Jeff Schlegel. (1996). *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs*. Prepared for
4. The California Demand-Side Measurement Advisory Committee: Project 2091T
5. Frederick D. Sebold, Alan Fields, Shel Feldman, Miriam Goldberg, Ken Keating and Jane Peters. (2001). *A Framework for Planning and Assessing Publicly Funded Energy Efficiency:*
6. *Study ID PG&E-SW040.* Prepared for the Pacific Gas & Electric Company.
7. Taylor, John R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito, California: University Science Books.
8. TecMarket Works Team. (2004). *The California Evaluation Framework.* Prepared for the California Public Utilities Commission and the Project Advisory Group, Framework
9. The TecMarket Works Team. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Directed by the CPUC's Energy Division, and with guidance from Joint Staff.
10. Weiss, Carol H. (1997).Theory-Based Evaluation: Past, Present, and Future. In Debra J. Rog and Deborah Fournier (eds.) *Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods*. San Francisco: Jossey-Bass Publishers.
11. Weiss, Carol H. (1998). *Evaluation*. Upper Saddle River, New Jersey: Prentice Hall.

2. Estimating Net-To-Gross Ratios Using the Self-Report Approach

Various methods exist for estimating net energy and demand impacts including true experimental design, quasi-experimental designs and the self-report approach (SRA) among others. The first two approaches estimate net energy and demand impacts directly. The SRA approach is used to estimate a net-to-gross ratio (NTGR), an index of program influence and defined as 1- Freeridership + Spillover. Once the NTGR is estimated, it is then multiplied by the evaluated gross savings to produce the estimate of net energy and demand impacts. The term (1 – freeridership) can be thought of as an estimate of program influence on the measures or practices promoted by the program. Some evaluators estimate this program influence by estimating freeridership (FR) and perform the subtraction. Others estimate this program influence on measures or practices promoted by the program directly. The two approaches differ in terms

approach to ensuring that the components of the study that do meet these requirements are performed in a manner that gives due attention to limiting the effects of sampling error.

of the wording of questions and their interpretation. It makes no difference which approach is used. Once program influence on measure or practices promoted by the program is estimated, it can be adjusted upwards to account for program-induced spillover (SO) measures to produce the final NTGR.

Most evaluation plans and completed reports previously reviewed by the New York Department of Public Service have relied on the SRA method. The SRA is a mixed methods approach that uses, to varying degrees, both quantitative and qualitative data and analysis to assess causality[26]. However, in these reviews, DPS discovered that in both the residential and nonresidential sectors the SRA method is not always designed and implemented according to best practices. Thus, DPS developed these *Guidelines for Estimating Net-to-Gross Ratios Using the Self-Report Approach* (SRA Guidelines) that requires analysts to address certain key issues but does not require analysts to address these issues in a specific way. The primary use of these SRA Guidelines is to assess the influence of the program on measures installed through the program and to make sure that evaluators are adhering, whenever possible, to these best practices. The Guidance Document does not mention all the available methods and leaves it up to the evaluators to select the method that is most appropriate. Finally, the Guidance Document does not preclude the estimation and inclusion of broader program administrator (PA)-induced market effects in place of SO.

It follows that these SRA Guidelines must focus on those methodological issues on which there is general agreement regarding their importance within the social science and engineering communities. The SRA Guidelines will also refer analysts to texts in which more detailed guidance can be found regarding all the issues addressed. Adherence to such guidelines may allow the results to be shaped by the interaction of the situation, the data and the analyst. It is this very interaction and the resulting plethora of legitimate methodological choices that prohibited the creation of a more detailed and prescriptive set of guidelines.

Two key best practice documents were relied upon in developing these guidelines:
Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches
(Ridge et al., 2007[27]) and Framework for Self-Report Net-To-Gross (Attribution) Questions (Winch et al., 2008).

Issues Surrounding the Validity and Reliability of Self-Report Techniques

The SRA deviates from the standard approach to assessing causality, *i.e.*, internal validity (did in fact the treatment make a difference). The standard approach to assessing causality is to conduct an experiment or quasi-experiment[28] in which data are collected from both participants and nonparticipants with the data being subjected to a variety of statistical analyses (Shadish, Cook, and Campbell, 2002). In the early 1970s, many began to realize that such evaluation designs were not always desirable or possible

[26] There is wide agreement on the value of *both* qualitative and quantitative data in the evaluation of many kinds of programs. Moreover, it is inappropriate to cast either approach in an inferior position. The complexity of any decision regarding the purchase of efficient equipment can be daunting, especially in large organizations for which the savings are often among the largest. In such situations, the reliance on only quantitative data can miss some important elements of the decision. The collection and interpretation of qualitative data can be especially useful in broadening our understanding of a program’s role in this decision.

[27] The original document upon which this document was substantially based was prepared by Richard Ridge and Katherine Randazzo for the California Demand Side Management Advisory Committee (CADMAC) in 1998. It was incorporated into the *Protocols and Procedures for the Verification of Costs, Benefits, and Shareholder Earnings from Demand-Side Management Programs* as Appendix J.

[28] In the literature, evaluations of energy efficiency and conservation programs that involve the use of a true experimental design are very rare.

(Weiss, 1972; Weiss and Rein, 1972). As a result, many evaluators began to explore alternatives that would allow them to generate causal conclusions (Guba, 1981, 1990; Cronbach, 1982). Such approaches as the modus operandi method (Scriven, 1976), intensive case studies (Yin 1994), theory-based evaluations (Chen, 1990; Rogers, et al., 2000), and mixed methods (Tashakkori and Teddlie, 1998) have been explored as alternative ways to generate causal conclusions. The SRA fits well with this tradition.

The SRA is useful in a variety of situations. For example, in some cases, the expected magnitude of the savings for a given program might not warrant the investment in an expensive evaluation design that could involve a billing analysis or a discrete choice analysis of both participants and nonparticipants and that would address self-selection bias. Or, key stakeholders might not want to wait for a billing analysis to be completed. Also, if the relationship of the savings to the normal monthly variation in energy use is too small, then a billing analysis should not be attempted owing to a lack of statistical power. Finally, in some cases, it might not be possible to identify a group of customers to serve as a comparison group, since they have been exposed through prior participation or are in some other ways "contaminated." So, for budgetary, timing, statistical, and research design issues, the more traditional designs and analyses must sometimes be replaced with the SRA.

More specifically, the SRA is a mixed method approach that involves asking one or more key participant decision-makers a series of structured and open-ended questions about whether they would have installed the same EE equipment in the absence of the program, as well as questions that attempt to rule out rival explanations for the installation (Weiss, 1972; Scriven, 1976; Shadish, 1991; Wholey et al., 1994; Yin, 1994; Mohr, 1995). In the simplest case (e.g., residential customers), the SRA is based primarily on quantitative data while in more complex cases the SRA is strengthened by the inclusion of additional quantitative and qualitative data which can include, among others, in-depth, open-ended interviews, direct observation, and review of program records 29. Many evaluators believe that additional *qualitative* data regarding the economics of the customer's decision and the decision process itself can be very useful in supporting or modifying *quantitatively*-based results (Britan, 1978; Weiss and Rein, 1972; Patton, 1987; Tashakkori and Teddlie, 1998).

Having presented a very brief history of these alternatives approaches, we move on to discuss a number of special challenges associated with the SRA that merit mentioning.

One of the problems inherent in asking program participants if they would have installed the same equipment or adopted the same energy-saving practices without the program, is that we are asking them to recall what has happened in the past. Worse than that is the fact that what we are really asking them to do is report on a **hypothetical situation**, what they would have done in the absence of the program. In many cases, the respondent may simply not know and/or cannot know what would have happened in the absence of the program. Even if the customer has some idea of what would have happened, there is, of necessity, uncertainty about it.

The situation just described is a circumstance ripe for invalid answers (low construct validity) and answers with low reliability, where reliability is defined as the likelihood that a respondent will give the same answer to the same question whenever or wherever it is asked. It is well known in the interview literature that the more factual and concrete the information the survey requests, the more accurate responses are likely to be. Where we are asking for motivations and processes in hypothetical situations that occurred one or two years ago, there is room for bias. Bias in responses is commonly thought to stem from three origins. First is the fact that some respondents may believe that claiming no impact for the

[29] Of course, in the simplest cases, an evaluator is free to supplement the analysis with additional quantitative and qualitative data.

program is likely to cause the program to cease, thus removing future financial opportunities from the respondent. Closely related to this is the possibility that the respondents may want to give an answer that they think will be pleasing to the interviewer. The direction of the first bias would be to increase the NTG ratio, and the second would have an unclear effect – up or down, depending on what the respondent thinks the interviewer wants to hear.

The second commonly recognized motivation for biased answers is that some people will like to portray themselves in a positive light; *e.g.*, they might like to think that they would have installed energy-efficient equipment without any incentive (the socially desirable response). This type of motivation could result in an artificially low net-to-gross ratio.

The third hypothesized source of bias involves an interaction between the positive perception of taking energy efficiency actions, the often observed difference between stated intentions and actual behaviors, and the fact that the counter-factual outcome cannot be viewed, by the participant or outsiders. Often a series of survey questions are asked of the participant about the actions they would have taken if there had been no program to derive a freeridership estimate. More specifically, this is asking the respondent to state their intentions with respect to purchasing the relevant equipment absent the program. Bias creeps in because people may intend many things that they do not eventually accomplish.

Beyond the fact that the situations of interest have occurred in the past and judgments about them involve hypothetical circumstances, they are often complex. No one set of questions can apply to all decision processes that result in a program-induced course of action. Some installations are simple, one-unit measures, while others involve many units, many different measures, and installations taking place over time. The decision to install may be made by one person or several people in a household, an individual serving as owner/operator of a small business, or, in the case of large commercial, industrial, or agricultural installations by multiple actors at multiple sites. Some measures may have been recommended by the utility for years before the actual installation took place, and others may have been recommended by consultants and/or vendors, making degree of utility influence difficult to establish. Finally, some efficiency projects may involve reconfiguration of systems or practices (such as operations and maintenance) rather than simple installations of energy-efficient equipment.

Another factor that can complicate the SRA is that, in certain situations, the estimated NTGR combines (more often implicitly than explicitly) the probability of a decision/action occurring and whether the *quantity* of the equipment installed would have been the same. This can complicate the interpretation of the responses and the way in which to combine these types of questions in order to estimate the NTGR. This type of complexity and variation across sites requires thoughtful design of survey instruments. Following is a discussion of the essential issues that should be considered by evaluators using SRA, together with some recommendations on reporting the strategies used to address each issue.

These should be regarded as recommendations for minimum acceptable standards for the use of the SRA to estimate net-to-gross ratios. Much of this chapter focuses on self-report methodologies for developing NTGRs for energy efficiency improvements in all sectors regardless of the size of the expected savings and the complexity of the decision-making processes. However, in a given year, energy efficiency programs targeted for industrial facilities are likely to achieve a relatively small number of installations with the potential for extremely large energy savings at each site. Residential programs often have a large number of participants in a given year, but the energy savings at each home, and often for the entire residential sector, are small in comparison to savings at non-residential sites. Moreover, large industrial customers have more complex decision-making processes than residential customers. As a result, evaluators are significantly less likely to conduct interviews with multiple actors at a single residence or to construct detailed case studies for each customer – methods that are discussed in detail in the following sections. *It may not be practical or necessary to employ the more complex techniques (e.g., multiple interviews at the same site, case-specific NTGR development) in all evaluations. Specifically, Sections 2.17 and 2.18 are probably more appropriate for customers with large savings and more complex decision-making processes.* Of course, evaluators are free to apply the guidelines in these sections even to customers with smaller savings and relatively simple decision-making processes.

Within the context of these best practices, there is room for some variation. For example, some programs may be so small, with commensurately small evaluation budgets, that full compliance with these best practices may be entirely impractical. For example, the number of set-up questions (Section 2.3), decision-making questions (Section 2.4), the number of NTGR questions (Section 2.5), or the number of consistency checks (Section 2.8) may be minimal. *Ultimately, each application of the SRA and its level of compliance with these SRA Guidelines must be viewed within a complex set of circumstances. Each PA should describe the level of effort and degree of compliance in light of the specific circumstance surrounding each program evaluation.*

Timing of the Interview

In order to minimize the problem of recall, SRA interviews addressing freeridership should be conducted with the decision maker(s) as soon after the installation of equipment as possible (Stone et al., 2000). It is recognized that interviews or other data collection to assess spillover need to be conducted later to allow enough time for the occurrence of spillover.

Identifying the Correct Respondent

Recruitment procedures for participation in an interview involving self-reported net-to-gross ratios must address the issue of how the correct respondent(s) will be identified. In the residential and small business sectors, this is relatively straightforward. However, in large commercial and industrial facilities, there are complexities that should be addressed such as:

- Different actors have different and complementary pieces of information about the decision to install, *e.g.*, the CEO, CFO, facilities manager, etc.;
- Decisions are made in locations such as regional or national headquarters that may be far away from the installation site;
- Significant capital decision-making power is lodged in commissions, committees, boards, or councils; and
- There is a need for both a technical decision-maker and a financial decision-maker to be interviewed (and in these cases, how the responses are combined will be important).

An evaluation using self-report methods should employ caution to handle all of these situations in a way that assures that the person(s) with the authority and the knowledge to make the installation decision are interviewed.

Set-Up Questions

Regardless of the magnitude of the savings or the complexity of the decision-making process, questions may be asked about the motivations for making the decisions that were made, as well as the sequence of events surrounding the decision. Sequence and timing are important elements in assessing motivation and program influence on it. Unfortunately, sequence and timing will be difficult for many respondents to recall. This makes it essential that the interviewer guide the respondent through a process of establishing benchmarks against which to remember the events of interest (Stone et al., 2000). Failure to do so could well result in, among other things, the respondent "telescoping" some events of interest to him into the period of interest to the evaluator. Set-up questions that set the mind of the respondent into the sequence of events that led to the installation, and that establish benchmarks, can minimize these problems. However, one should be careful to avoid wording the set-up questions in such a way so as to bias the response in the desired direction.

Set-up questions should be used at the beginning of the interview, but they can be useful in later stages as well. Respondents to self-report surveys frequently are individuals who participated in program decisions and, therefore, may tend to provide answers ex post that validate their position in those decisions. Such biased responses are more likely to occur when the information sought in questions is abstract, hypothetical, or based on future projections, and are less likely to occur when the information sought is concrete. To the extent that questions prone to bias can incorporate concrete elements, either by set-up questions or by follow-up probes, the results of the interview will be more persuasive.

An evaluation using self-report methods should employ and document a set of questions that adequately prompt the respondent to the context and sequence of events that led to decision(s) to adopt a DSM measure or practice, including clearly identified benchmarks in the customer's decision-making process.

Such set-up questions could include:

- Confirm or determine whether the project involves new construction, building expansion, replacement of existing equipment, or modification to existing equipment.
- Confirm the type of equipment installed, date, incentive amount, and other information deemed relevant.
- Confirm the evaluator's information regarding key services, incentives, and assistance provided by the program, as well as the type and amount of vendor/implementer involvement.
- Determine when and how the respondent first heard about the services/incentives/assistance available through the program.
- Explore the possibility that new equipment was already installed before hearing about the services/incentives/assistance available from the program.
- Explore any plan(s) to purchase or install equipment before learning about the services/incentives/assistance available through the program.
- Understand the existing plans.
    - Understand the point in the planning process that the respondent/organization: (1) became aware of the program, and (2) began discussing plans with the program representative(s).
    - Understand qualitatively, the impact/changes necessitated by program involvement.

Discuss the working condition of replaced equipment (probe: planned replacement/upgrade, failure, estimated remaining useful life, repair history, etc.)

- Explore what first made the respondent (organization) start thinking about installing/replacing equipment at (home/this facility).
- Identify the age of equipment that was replaced.
- Explore previous program participation.

Decision-Making Process

These questions address key aspects of the customer's decision-making process. In many respects, they are an extension of the context questions and have a similar intent. The intent is to get participants to talk about their project-related decision-making; what factors went into that process, and what decision-makers were involved. The intent of these questions is to elicit how (and the extent to which) the decision-making process was altered as a result of their program participation.

A key purpose of the decision-making questions is to help the respondent recall the particulars of their program-related decision-making and prepare them to answer the direct attribution questions about how the program affected the timing, efficiency level, and quantity of the technology installed. Similar to context questions, answers to these questions provide key indications of program influence (or lack thereof) and should be compared and contrasted with how the respondent answers the direct attribution questions. The idea is to determine whether responses to these questions are consistent with the answers given to the direct attribution questions. For example, with respect to the installed energy-efficient equipment, lack of plans to purchase, lack of awareness, and no prior experience are all indicators of program influence that should be considered, regardless of responses to the direct attribution questions.

Such decision-making process questions could include:

- Organizational policies that specify factors considered when purchasing new (replacing old) equipment/ (probe: payback, return on investment, guidelines on efficiency levels, etc.)
- Major obstacles/barriers faced when seeking approval for project. (probe: budget, time constraints, other priorities, disruption of production, etc.)
- Role of contractor(s)/vendor(s) in project.
- Making respondent aware of the program (or vice versa).
- Decision to participate.
- Recommendation to install certain type/energy efficiency level of equipment.
- Influence of contractor/vendor involvement on decision to install equipment at this time.
- Explore total costs—that is, all financial assistance, plus the costs not covered by financial assistance.
- Budgeting process for new/replacement equipment. (probe: size projects budgeted for, budget planning cycle/length, budgetary approvals required, etc.)
- Who within the organization is responsible for recommending the purchase of new/replacement equipment?
- Who within the organization is responsible for approving the purchase of new/replacement equipment?

It is also important that questionnaires do not launch immediately into a series of questions about extent to which the program influenced the customer's decision. They should first determine *whether* the program influenced their decision in any way.

Use of Multiple Questions

Regardless of the magnitude of the savings, or the complexity of the decision-making process, one should assume that using multiple questionnaire items (both quantitative and qualitative) to measure a construct, such as freeridership, is preferable to using only one questionnaire item, since reliability is increased by the use of multiple items (Blalock, 1970; Crocker & Algina; 1986; Duncan, 1984).

Validity and Reliability

The validity and reliability of *each question* used in estimating the NTGR must be assessed (Lyberg, et al., 1997). In addition, the internal consistency (reliability) of multiple-item NTGR *scales* should not be assumed and should be tested. Testing the reliability of scales includes such techniques as split-half correlations, Kuder-Richardson, and Cronbach's alpha (Netemeyer, Bearden, and Sharma, 2003; Crocker & Algina, 1986; Cronbach, 1951; DeVellis, 1991). An evaluation using self-report methods should employ some or all of these tests, or other suitable tests, to evaluate reliability, including a description of why particular tests were used and others were considered inappropriate.

For those sites with relatively large savings and more complex decision-making processes, both quantitative and qualitative data may be collected from a variety of sources (*e.g.*, telephone interviews with the decision maker, telephone interviews with others at the site familiar with the decision to install the efficient equipment, paper and electronic program files, and on-site surveys). These data must eventually be integrated in order to produce a final NTGR.[30] Of course, it is essential that all such sites be evaluated consistently using the same instrument. However, in a situation involving both quantitative and qualitative data, interpretations of the data may vary from one evaluator to another, which means that, in effect, the measurement result may vary. Thus, the central issue here is one of reliability, which can be defined as obtaining consistent results over repeated measurements of the same items.

To guard against such a threat at those sites with relatively large savings and more complex decision-making processes, the data for each site should be evaluated by more than one member of the evaluation team. Next, the resulting NTGRs for the projects should be compared, with the extent of agreement being a preliminary measure of the inter-rater reliability. Any disagreements should be examined and resolved, and all procedures for identifying and resolving inconsistencies should be thoroughly described (Sax, 1974; Patton, 1987).

Ruling Out Rival Hypotheses

Most significant events in the social world are not mono-causal, but instead are the result of a nexus of causal influences. Both in social science and in everyday life, when we say that Factor A is strongly influential in helping to cause Event B, it is rarely the case that we believe factor A is the sole determinant of Event B. Much more commonly, what we mean to say is that Factor A is among the leading determinants of Event B. Thus, an evaluator should attempt to rule out rival hypotheses regarding the reasons for

[30] For a discussion of the use of qualitative data see Sections 2.15 and 2.17.

installing the efficient equipment (Scriven, 1976). For example, to reduce the possibility of socially desirable responses, one could ask an *open-ended question* (*i.e*., a list of possible reasons is **not** read to the respondent) regarding other possible reasons for installing the efficient equipment. A listing by the interviewer of such reasons such as global warming, energy efficiency programs, the price of electricity, concern for future generations, and the need for the US to reduce oil dependency might elicit socially desirable responses which would have the effect of artificially reducing the NTGR. The answers to such questions about other possible influences can be factored into the estimation of the NTGR.

Consistency Checks

When multiple questionnaire items are used to calculate a freeridership probability there is always the possibility of contradictory answers. Contradictory answers indicate problems of validity and/or reliability (internal consistency). Occasional inconsistencies indicate either that the respondent has misunderstood one or more questions, or is answering according to an unanticipated logic.

Another potential problem with self-report methods is the possibility of answering the questions in a way that conforms to the perceived wishes of the interviewer, or that shows the respondent in a good light (consciously or unconsciously done). One of the ways of mitigating these tendencies is to ask one or more questions specifically to check the consistency and plausibility of the answers given to the core questions. Inconsistencies can highlight efforts to "shade" answers in socially desirable directions. While consistency checking won't overcome a deliberate and well-thought-out effort to deceive, it will often help when the process is subtler or where there is just some misunderstanding of a question.

An evaluation using self-report methods should employ a process for setting up checks for inconsistencies when developing the questionnaire items, and describe the methods chosen as well as the rationales for using or not using the techniques for mitigating inconsistencies. Before interviewing begins, the evaluator should establish a process to handle inconsistent responses. Such process steps should be consistently applied to all respondents.

Based on past experience, one should anticipate which questions are more likely to result in inconsistent responses (*e.g.*, questions of what participants would have done in the absence of the program and reported importance of the program to their taking action could). For such questions, specific checks for inconsistencies along with interviewer instructions could be built into the questionnaire. Any, apparent inconsistencies can then be identified and, whenever possible, resolved before the interview is over. If the evaluator waits until the interview is over to consider these problems, there may be no chance to correct misunderstandings on the part of the respondent or to detect situations where the evaluator brought incomplete understanding to the crafting of questions. In some cases, the savings at stake may be sufficiently large to warrant a follow-up telephone call to resolve the inconsistency.

However, despite the best efforts of the interviewers, some inconsistencies may remain. When this occurs, the evaluator could decide which of the two answers, in their judgment has less error, and discard the other. Or, one could weight the two inconsistent responses in a way that reflects the evaluator's estimate of the error associated with each, *i.e*., a larger weight could be assigned to the response that, in their judgment, contains less error.

However, any inconsistencies are handled, a process for resolving inconsistencies should be established, to the extent feasible, *before* interviewing begins.[31] An evaluation plan using self-report methods should describe the approach to identifying and resolving apparent inconsistencies. The plan should include: 1) the key questions that will be used to check for consistency, 2) whether and how it will be determined that the identified inconsistencies are significant enough to indicate problems of validity and/or reliability (internal consistency), and 3) how the indicated problems will be mitigated. The final report should include: 1) a description of contradictory answers that were identified, 2) whether and how it was determined that the identified inconsistencies were significant enough to indicate problems of validity and/or reliability (internal consistency), and 3) how the indicated problems were mitigated.

However, the process itself has sometimes been found to produce biased results, eliminating these respondents (treating them as missing data) has at times been the selected course of action. Thus, whenever any of these methods are used, one must report the proportion of responses affected. One must also report the mean NTGR with and without these responses in order to assess the potential for bias.

Making the Questions Measure-Specific

It is important for evaluators to tailor the wording of central freeridership questions to the specific technology or measure that is the subject of the question. It is not necessarily essential to incorporate the specific measure into the question, but some distinctions must be made if they would impact the understanding of the question and its potential answers. For instance, when the customer has installed equipment that is efficiency rated so that increments of efficiency are available to the purchaser, asking that respondent to indicate whether he would have installed the same equipment without the program could yield confusing and imprecise answers. The respondent will not necessarily know whether the evaluator means the exact same efficiency, or some other equipment at similar efficiency, or just some other equipment of the same general type. Some other possibilities are:

- Installations that involve removal more than addition or replacement (*e.g.*, delamping or removal of a second refrigerator or freezer in a residence);
- Installations that involve increases in productivity rather than direct energy load impacts;
- Situations where the energy-efficiency aspect of the installation could be confused with a larger installation; and
- Installation of equipment that will result in energy load impacts, but where the equipment itself is not inherently energy-efficient.

An evaluation using self-report methods should include and document an attempt to identify and mitigate problems associated with survey questions that are not measure-specific, and an explanation of whether and how those distinctions are important to the accuracy of the resulting estimate of freeridership.

The challenge of getting the respondent to focus on the measure(s) installed through the program varies by sector. For example, in large nonresidential facilities or with decision-makers across multiple buildings or locations, care must be taken to ensure that the specific pieces of equipment, or group of equipment/facility decisions, are properly identified. The interviewer and respondent need to be referring

[31] One might not always be able to anticipate all possible inconsistencies before interviewing begins. In such cases, rules for resolving such unanticipated inconsistencies should be established before the analysis begins.

to the same things. However, in the residential sector, getting the respondent to focus on the particular measure installed through a program is far simpler.

As part of survey development, an assessment needs to be made of whether there are important subsets within the participant pool that need to be handled differently. For example, any program that contains corporate decision-makers managing building/renovation of dozens of buildings per year requires some type of special treatment. In this case, a standard survey might ask about three randomly selected projects/buildings. Or, a case study type of interview could focus on the factors affecting their decisions in general, for what percentage of their buildings do they take certain actions, and what actions do they take in cases where no incentives are available (if a regional or national decision-making), etc.. Such an approach might offer better information to apply to all the buildings they have in the program. The point is that without special attention and a customized survey instrument, such customers might find the interview too confusing and onerous.

Partial Freeridership

Partial freeridership can occur when, in the absence of the program, the participant would have installed something more efficient than the program-assumed baseline efficiency but not as efficient as the item actually installed as a result of the program. It can also occur when the participant would have installed the same quantity of equipment, fewer, or more at that time without the program. When there is a likelihood that this is occurring, an evaluation using self-report methods should include and document attempts to identify and quantify the effects of such situations on net savings. Partial freeridership should be explored for those customers with large savings and complex decision making processes.

In such a situation, it is essential to develop appropriate and credible information to establish precisely the participant's alternative choice for efficiency and quantity. The likelihood that the participant would really have chosen a higher-than-baseline efficiency option and whether the quantity would have been the same, fewer or more is directly related to their ability to clearly describe these options.

An evaluation using self-report methods should include and document attempts to identify and mitigate problems associated with partial freeridership, when applicable.

Timing of the Purchase

Early replacement is defined as the replacement of equipment before it reaches its Effective Useful Life (EUL) whereas end-of-life or normal replacement refers to the replacement of equipment which has reached or passed the end of its measure-prescribed EUL. An *evaluator* charged with verifying savings associated with early replacements must first verify that a given installation is actually a case of early replacement. This can be accomplished by asking the participant such questions as:

- Approximately how old was the existing equipment?
- How much longer do you think it would have lasted?
- In your opinion, based on the economics of operating this equipment, for how many more years could you have kept this equipment functioning?
- Which of the following statements best describes the performance and operating condition of the equipment you replaced through the PROGRAM?
    - Existing equipment was fully functional
    - Existing equipment was fully functioning, but with significant problems

  - Existing equipment had failed or did not function
  - Existing equipment was obsolete
  - Not applicable, ancillary equipment (VSD, EMS, controls, etc.)
- How much downtime did you experience in the past year, and how did this compare with the previous year(s)?
- Over the last 5 years, have maintenance costs been increasing, decreasing or staying about the same?

If this condition is met, the evaluator should examine the following eight variables and associated documentation and make any necessary adjustments[32]:

- The EUL of the new efficient equipment,
- The RUL of the old equipment,
- The full savings of the equipment (annual energy use of the old equipment in place minus the annual energy of the installed high efficiency equipment supported by the program), and
- The full costs
- The adjusted full cost (full cost multiplied by the full-cost adjustment factor),
- The ratio of incremental savings to full savings,
- The ratio of incremental costs to full costs, and
- Adjusted EUL

Note that, if the claim of early replacement cannot be verified, the evaluator must then determine the more appropriate baseline for estimating gross savings.

However, just because a customer replaced some equipment ahead of schedule doesn't mean that a program should get credit for this early replacement. The evaluator must also determine the extent to which the program caused the customer to replace their equipment ahead of schedule. The approach used should be reasonably robust given the importance of early replacement in PA portfolios. For example, to the extent possible, evaluators should rely on multiple questions in estimating the influence of the program on early replacement. In addition, there are a number of considerations that must be taken into account:

- In conducting NTGR surveys of program participants, *evaluators* will encounter participants for which the *program implementer* claimed early replacement savings. For each individual NTGR survey, the questions bearing on early replacement should *always* be asked by *ex post evaluators* and used in the calculation of the NTGR in order for the program that claimed early replacement to get due credit for accelerating savings to the grid. Consider a situation in which the NTGR is based on questions regarding the influence of the program on *what* and *how many* efficient measures a customer installed. In the case in which the influence on either *what* was installed or the *quantity* that was installed is zero or near zero, the resulting NTGR will be zero or near zero as will the net savings. However, the respondent also indicates that the program had a significant influence on getting them to purchase and install the measures much sooner than they would have

[32] The New York Standard Approach for Estimating Energy Savings from Energy Efficiency Programs (Technical Resource Manual), includes Appendix M: Guidelines for Early Replacement Conditions which provides additional information related to early replacement, including detailed definitions of the variables referenced here.

otherwise. In such cases, the influence of the program on the timing of the purchase should be taken into account in estimating the NTGR.[33]

- Evaluators might also encounter customers for whom a program implementer did not initially claim early replacement for a given measure, but who now claim that the efficiency program caused them to replace the equipment before the end of its useful life. In such cases, the timing question(s), when triggered by a high rate of freeridership, would always be taken into account when calculating the NTGR. This gives the program credit for cases of early replacement that the program did not originally claim.

- Many evaluators have used a participant's forecast of when they would have installed the same equipment absent the program as one indicator of program influence on early replacement. The longer into the future, the greater the influence of the program. When this approach is used, the point at which the length of the deferral is interpreted as meaning no freeridership needs to be explicitly developed in the evaluation plan and should be justified given the length of the measure life (the effective useful life or EUL) and the decision-making process of that type of customer. While additional factors can be taken into account, the evaluator must clearly explain the rationale for their inclusion.

Handling Non-Responses and "Don't Knows"

In some cases, some customers selected for the evaluation sample refuse to be interviewed (unit nonresponse). In other cases, some customers do not complete an attempted interview, complete the interview but refuse to answer all of the questions, or provide a "don't know" response to some questions (item nonresponse). Insoluble contradictions fall into the latter category. Evaluators should explain in advance how they will address each type of problem.

Consider those who choose not to respond to the questionnaire or interview (unit nonresponse). Making no attempt to understand and correct for nonresponse in effect assumes that the non-respondents would have answered the questions at the mean. Thus, their net-to-gross ratios would assume the mean NTGR value. Because this might not always be a reasonable assumption, one should always assess the possibility of non-response bias. To assess the possibility of non-response bias, one should, at a minimum, using information available on the population, describe any differences between those who responded and those who didn't and attempt to explain whether any of these differences are likely to affect one's answers to the NTGR battery of questions. If non-response bias is suspected, one should, whenever possible, explore the possibility of correcting for non-response bias. When not possible, one should explain why not (*e.g.*, timing or budget constraints) and provide one's best estimate of the magnitude of the bias.

[33] As of May 30, 2013, impact evaluations in New York generally have not yet begun encountering cohorts of participants subject to the requirements of Appendix M, and as a result there is little practical experience to date with ex-post evaluation in support of the provisions of Appendix M. Evaluation of early replacement assumptions and of acceleration as a component of free riding is complex. It is therefore possible that, in specific studies, ex-post evaluation in support of the provisions of Appendix M and of the net-to-gross ratio will interact in ways that cannot yet be anticipated. However, the intent here is to anticipate one interaction that seems relatively likely.

When some respondents terminate the interview, complete the interview but refuse to answer all the questions, or who provide a "don't know" response to some questions (item nonresponse), decisions must be made as to whether one should treat such cases as missing data or whether one should employ some type of missing data imputation.

In all cases, one should always make a special effort to avoid "don't know" responses when conducting interviews. However, some survey methods and procedures have been used that do not allow a "don't know" response where that might be the best response a respondent can provide. Forcing a response can distort the respondent's answer and introduce bias. Such a possibility needs to be recognized and avoided to the extent possible.

Scoring Algorithms

A consequence of using multiple questionnaire items to assess the probability of freeridership (or its complement, the NTGR) is that decisions must be made about how to combine them. Do all items have equal weight or are some more important indicators than others? How are probabilities of freeridership assigned to each response category? Answers to these questions can have a profound effect on the final NTGR estimate. These decisions are incorporated into the algorithm used to combine all pieces of information to form a final estimate of the NTGR. All such decisions must be described and justified by evaluators.

In some cases, each of the responses in the series of questions is assigned an ad hoc probability for the expected net savings. These estimates are then combined (additively or multiplicatively) into a participant estimate. The participant estimates are subsequently averaged (or weighted averaged given expected savings) to calculate the overall freeridership estimate. The assignments of the probabilities are critical in the final outcome. At the same time, there is little evidence of what these should be and they are often assigned and justified given a logical argument. With this, however, a multiple number of different probability assignments have been shown to be justified and accepted by various evaluations and regulators. However, we recognize that this can make the comparability and reliability of survey-based estimates problematic.

It is also critical that the NTGR algorithm, which takes responses to multiple questions regarding program influence, not be calculated in a way that produces a biased estimate of the NTGR. A variety of NTGR algorithms have been identified, such as the incorrect use of a multiplicative algorithm, which should be avoided.[34]

Finally, when multiple questions, weights, and complex algorithms are involved in calculating the NTGR, evaluators should also consider conducting a sensitivity analysis (e.g., changing weights, changing the questions used in estimating the NTGR, changing the probabilities assigned to different response categories, etc.) to assess the stability and possible bias of the estimated NTGR. A preponderance of evidence approach is always better than relying solely on a weighted algorithm and sophisticated weighting that is not transparent and logically conclusive should be avoided.

[34] Keating, Ken. (2009). *Freeridership Borscht: Don't Salt the Soup*. Presented at the 2009 International Energy Program Evaluation Conference.

WEIGHTING THE NTGR

The DPS Guidelines require estimates of the NTGR at the program level. Of course, such an NTGR must take into account the size of the impacts at the customer or project level. Consider two large industrial sites with the following characteristics. The first involves a customer whose self-reported NTGR is .9 and whose estimated annual savings are 200,000 kWh. The second involves a customer whose self-reported NTGR is .15 and whose estimated savings are 1,000,000 kWh. One could calculate an unweighted NTGR across both customers of .53. Or, one could calculate a weighted NTGR of .28. In this instance, the latter calculation is the appropriate one.

PRECISION OF THE ESTIMATED NTGR

Most the discussion thus far has been focused on the accuracy of the NTGR estimate and not the precision of the estimate. The calculation of the achieved relative precision of the NTGRs (for program-related measures and practices and non-program measures and practices) is usually straightforward, relying on the standard error and the level of confidence. For example, when estimating NTGRs in the residential sector, one typically interviews one decision maker in each household with the NTGR estimate based on multiple questions. In such a situation, one could report the mean, standard deviation, the standard error, and the relative precision of the NTGR based on the sample at the 90 percent levels of confidence

In the nonresidential sector, things can get much more complicated since the NTGR at a given site can be based on such information as: 1) multiple interviews (end users as well as those upstream from the end user that might have been involved in the decision) that takes into account the propagation of errors, 2) other more qualitative information such as standard purchasing policies that require a specific corporate rate of return or simple payback (e.g., the rate of return for the investment in the energy efficiency measure can be calculated with and without the rebate to obtain another point estimate of the influence of the program), or 3) a vendor's participation in utility training programs. In such a situation, a NTGR will be estimated that uses all of this information.

However, in such situations when the NTGR is based on multiple sources of data, it might be difficult or impossible to track the propagation of errors associated with the estimation of any one measure-level NTGR and incorporate these errors into the relative precision of final program-level NTGR. Thus, the standard errors should be based on the final NTGRs estimated for the sample of measures or projects. For example, consider a large industrial program for which NTGRs have been estimated for each of 70 sites in the sample. The estimation of the NTGR for each site involved interviewing multiple decision makers and associated vendors and the use of other qualitative information. The relative precision for the program is based only on the standard error of the 70 final NTGRs and ignores any error associated with any of the inputs into the final NTGR for any given measure.

Pre-Testing Questionnaire

Of course, as with any survey, a pre-test should be conducted to reveal any problems such as ambiguous wording, faulty skip patterns, leading questions, faulty consistency checks, and incorrect sequencing of questions. Modifications should be made prior to the official launch of the survey.

The Incorporation of Additional Quantitative and Qualitative Data in Estimating the NTGR

When one chooses to complement a mixed method (quantitative and qualitative) analysis of freeridership with additional data, there are a few very basic concerns that one must keep in mind.

Data Collection

Use of Multiple Respondents

In situations with relatively large savings and more complex decision-making processes, one should use, to the extent possible, information from more than one-person familiar with the decision to install the efficient equipment or adopt energy-conserving practices or procedures (Patten, 1987; Yin, 1994).

It is important to inquire about the decision-making process and the roles of those involved for those cases with relatively large savings and with multiple steps or decision-makers. If the customer has a multi-step process where there are "go/no-go" decisions made at each step, then this process should be considered when using the responses to estimate the firm’s NTGR. There have been program evaluations whose estimates have been called into question when these factors were not considered, tested, and found to be important. For example, a municipal program serving cities with financial issues where a department’s facility engineer could say without bias that he definitely intended to install the same measure in the absence of the program and that he had requested that the city manager request the necessary funds from the City Council. However, one might discover that in the past the city manager, due to competing needs, only very rarely include the engineer’s requests in his budget submitted to the City Council. Similarly, there are cases where a facility engineer continues to recommend efficiency improvements, but never manages to get management approval until the efficiency program provides the information in a way that meets the financial decision-makers needs in terms of information or independent verification or leverage by obtaining “free” funds.

These interviews might include interviews with third parties who were involved in the decision to install the energy efficient equipment. Currently, there is no standard method for capturing the influence of third parties on a customer’s decision to purchase energy efficient equipment. Third parties who may have influence in this context include market actors such as store clerks, manufacturers (through promotional literature, demonstrations, and in-person marketing by sales staff), equipment distributors, installers, developers, engineers, energy consultants, and/or architects. Yet, these influences can be important and possibly more so in the continually changing environment with greater attention on global warming and more overlapping interventions. When one chooses to measure the effect of third parties, one should keep the following principles in mind: 1) the method chosen should be balanced. That is, the method should allow for the possibility that the third-party influence can increase or decrease the NTGR that is based on the customer's self report, 2) the rules for deciding which customers will be examined for potential third party influence should be balanced. That is, the pool of customers selected for such examination should not be biased towards ones for whom the evaluator believes the third-party influence will have the effect of influencing the NTGR in only one direction, 3) the plan for capturing third-party influence should be based on a well-conceived causal framework. The onus is on the evaluator to build a compelling case using a variety of quantitative and/or qualitative data for estimating a customer's NTGR.

Other Site and Market-Level Data

Information relevant to the purchase and installation decision can include:

- Program paper files (correspondence between DSM program staff and the customer, evidence of economic feasibility studies conducted by the utility or the customer, correspondence among the customer staff, other competing capital investments planned by the customer).
- Program electronic files (*e.g.*, program tracking system data, past program participation).
- Interviews with other people at the site who are familiar with the program and the choice (*e.g.*, operations staff).
- Open-ended questions on structured interviews with the key decision-maker and other staff who may have been involved with the decision.
- Incremental costs of the equipment.
- Estimates of the equipment's market share.
- The diffusion (saturation) of the equipment in the marketplace

Establishing Rules for Data Integration

In cases where multiple interviews are conducted eliciting both quantitative and qualitative data, and a variety of program documentation has been collected, one will need to integrate all of this information into an internally consistent and coherent story that supports a specific NTGR.

Before the analysis begins, one should establish, to the extent feasible, rules for the integration of the quantitative and qualitative data. These rules should be as specific as possible and adhered to throughout the analysis. Such rules might include instructions regarding when the NTGR based on the quantitative data should be overridden based on qualitative data, how much qualitative data is needed to override the NTGR based on quantitative data, how to handle contradictory information provided by more than one person at a given site, how to handle situations when there is no decision-maker interview, when there is no appropriate decision-maker interview, or when there is critical missing data on the questionnaire, and how to incorporate qualitative information on deferred freeridership.

One must recognize that it is difficult to anticipate all the situations that one may encounter during the analysis. As a result, one may refine existing rules or even develop new ones during the initial phase of the analysis. One must also recognize that it is difficult to develop algorithms that effectively integrate the quantitative and qualitative data. It is therefore necessary to use judgment in deciding how much weight to give to the quantitative versus qualitative data and how to integrate the two. The methodology and estimates, however, must contain methods to support the validity of the integration methods through preponderance of evidence or other rules/procedures as discussed above.

Analysis

A case study is one method of assessing both quantitative and qualitative data in estimating a NTGR. A case study is an organized presentation of all these data available about a particular customer site with respect to all relevant aspects of the decision to install the efficient equipment. When a case study approach is used, the first step is to pull together the data relevant to each case and write a discrete, holistic report on it (the case study). In preparing the case study, redundancies are sorted out, and information is organized topically. *This information should be contained in the final report.*

The next step is to conduct a content analysis of the qualitative data. This involves identifying coherent and important examples, themes, and patterns in the data. The analyst looks for quotations or observations that go together and that are relevant to the *customer's decision to install the efficient equipment*. Guba (1978) calls this process of figuring out what goes together "convergence," i.e., the extent to which the data hold together or dovetail in a meaningful way. Of course, the focus here is on evidence related to the degree of program influence in installing the efficient equipment. Identifying and ruling out rival explanations for the installation of the efficient equipment is a critical part of the analysis (Scriven, 1976).

Sometimes, *all* the quantitative and qualitative data will clearly point in the same direction while, in others, the *preponderance* of the data will point in the same direction. Other cases will be more ambiguous. In all cases, in order to maximize reliability, it is essential that more than one person be involved in analyzing the data. Each person must analyze the data separately and then compare and discuss the results. Important insights can emerge from the different ways in which two analysts look at the same set of data. Ultimately, differences must be resolved and a case made for a particular NTGR.

Finally, it must be recognized that there is no single right way to conduct qualitative data analysis:

> The analysis of qualitative data is a creative process. There are no formulas, as in statistics. It is a process demanding intellectual rigor and a great deal of hard, thoughtful work. Because different people manage their creativity, intellectual endeavors, and hard work in

> different ways, there is no one right way to go about organizing, analyzing, and interpreting qualitative data. (p. 146)

Ultimately, if the data are systematically collected and presented in a well-organized manner, and if the arguments are clearly presented, any independent reviewer can understand and judge the data and the logic underlying any NTGR. Equally important, any independent reviewers will have all the essential data to enable them to replicate the results, and if necessary, to derive their own estimates.

Qualified Interviewers

For the basic SRA in the residential and small commercial sectors, the technologies discussed during the interview are relatively straightforward (e.g., refrigerators, CFLS, T-8 lamps, air conditioners). In such situations, using the trained interviewers working for companies that conduct telephone surveys is adequate. However, in more complicated situations such as industrial process and large commercial HVAC systems, the level of technical complexity is typically beyond the abilities of such interviewers. In such situations, engineers familiar with these more complicated technologies should be trained to collect the data by telephone or in person.

Transparency

The question sequence and analysis process followed for determining program impacts must be transparent. The question sequence, analysis algorithms and the rationale for assigning attribution should be included in the resulting reports. Ideally, this reporting would include a matrix (or flow diagram) showing the combinations of responses given to the attribution questions and the percentage of customers (and percentage of the overall savings) that fall into each category. This allows stakeholders to fully understand how each question (and within each question, the response categories) affects the final result. In addition to the attribution questions, the matrix would include key context, decision-making, and consistency responses. Particular attention should be given to identifying the extent to which answers to these questions are in conflict with responses to the direct attribution questions.

References

1. Blalock, H. (1970), "Estimating Measurement Error Using Multiple Indicators and Several Points in Time," *American Sociological Review*, 35, pp. 101-111.
2. Bogdan, Robert and Steven J. (1975). Taylor. *Introduction to Qualitative Research Methods*. New York: John Wiley & Sons.
3. Britan, G. M. Experimental and Contextual Models of Program Evaluation. (1978). Evaluation and Program Planning 1: 229-234.
4. Chen, Huey-Tsyh. 1990. *Theory-Driven Evaluations*. Newbury Park, CA: SAGE Publications.
5. Cochran, William G. (1977). *Sampling Techniques*. New York: John Wiley & Sons.
6. Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart & Winston.
7. Cronbach, L. J. 1982. *Designing Evaluation and Social Action Programs*. San Francisco:

8. Jossey-Bass.
9. Cronbach L.J. (1951). "Coefficient Alpha and the Internal Structure of Tests." *Psychometrika*, 16, 297-334.
10. DeVellis, R.F. (1991). *Scale Development: Theory and Applications*. Newbury Park, CA: Sage Publications, Inc.
11. Duncan, O.D. (1984). *Notes on Social Measurement: Historical and Critical*. New York: Russell Sage.
12. Guba, E. G. (1978). Toward a Methodology of Naturalistic Inquiry in Educational Evaluation (CSE Monographic Series in Evaluation No. 8). Los Angeles: Center for the Study of Evaluation.
13. Hall, Nick, Johna Roth, Carmen Best, Sharyn Barata, Pete Jacobs, Ken Keating, Ph.D., Steve Kromer, Lori Megdal, Ph.D., Jane Peters, Ph.D., Richard Ridge, Ph.D.,
14. Francis Trottier, and Ed Vine, Ph.D. (2007). *California Energy Efficiency Evaluation: Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Prepared for the California Public Utilities Commission.
15. Keating, Ken. (2009). *Freeridership Borscht: Don't Salt the Soup*. Presented at the 2009 International Energy Program Evaluation Conference.
16. Lyberg, Lars, Paul Biemer, Martin Collins, Edith De Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. (1997). *Survey Measurement and Process Quality*. New York, NY: John Wiley & Sons.
17. Madow, William G., Harold Nisselson, Ingram Olkin. (1983). *Incomplete Data in Sample Surveys*. New York: Academic Press.
18. Maxwell, Joseph A. (2004). Using Qualitative Methods for Causal Explanations. Field Methods, Vol. 16, No. 3, 243-264 (2004)
19. Mohr, Lawrence B. (1995). *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.
20. Netemeyer, Richard G., William O. Bearden, and Subhash Sharma. (2003). *Scaling Procedures: Issues and Applications*. Thousand Oaks, CA: SAGE Publications.
21. New York State Department of Public Service and the Evaluation Advisory Group. (Revised Nov. 2012) *New York Evaluation Plan Guidance for EEPS Program Administrators*
22. Patton, Michael Quinn. (1987). *How to Use Qualitative Methods in Evaluation*. Newbury Park, California: SAGE Publications.
23. Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. (2007). *Guidelines for*
24. *Estimating Net-To-Gross Ratios Using the Self Report Approach*. Prepared for the California Public Utilities Commission.
25. Rogers, Patricia J., Timothy A. Hacsi, Anthony Petrosino, and Tracy A. Huebner (Eds.). (2000). *Program Theory in Evaluation: Challenges and Opportunities*. San Francisco, CA: Jossey-Bass Publishers.
26. Rossi, Peter and Howard E. Freeman. (1989). *Evaluation: A Systematic Approach.* Newbury Park, California: SAGE Publications.
27. Sax, Gilbert. (1974). *Principles of Educational Measurement and Evaluation*. Belomont, CA: Wadsworth Publishing Company, Inc.
28. Schumacker, Randall E. and Richard G. Lomax. (1996). *A Beginner's Guide to Structural Equation Modeling*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
29. Scriven, Michael. (1976). Maximizing the Power of Causal Explanations: The Modus Operandi Method. In G.V. Glass (Ed.), Evaluation Studies Review Annual (Vol. 1, pp.101-118). Bevery Hills, CA: Sage Publications.
30. Shadish, Jr., Thomas D. Cook, and Donald T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.
31. Shadish, Jr., William R. and Thomas D. Cook, and Laura C. Leviton. (1991). *Foundations of Program Evaluation*. Newbury Park, CA: Sage Publications, Inc.

32. Stone, Arthur A., Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. (2000). *The Science of the Self-Report: Implications for Research and Practice*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
33. Tashakkori, Abbas and Charles Teddlie. (1998). *Mixed Methodology: Combining Qualitative and Quantitative Approaches*.Thousand Oaks, CA: SAGE Publications.
34. TecMarket Works, Megdal & Associates, Architectural Energy Corporation, RLW Analytics, Resource Insight, B & B Resources, Ken Keating and Associates, Ed Vine and Associates, American Council for an Energy Efficient Economy, Ralph Prahl and Associates, and Innovologie. (2004). *The California Evaluation Framework.* Prepared for the California Public Utilities Commission and the Project Advisory Group.
35. Weiss, R. S. and M.Rein. (1972). The Evaluation of Broad-Aim Programs: Difficulties in Experimental design and an Alternative. In C. H. Weiss (ed.) *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn and Bacon.
36. Weiss, Carol H. (1998). *Evaluation*. Upper Saddle River, New Jersey: Prentice Hall.
37. Wholey, Joseph S., Harry P. Hatry and Kathryn E. Newcomer. (1994). *Handbook of Practical Program Evaluation*. San Francisco, CA: Jossey-Bass, Inc.
38. Winch, Rick, Tom Talerico, Bobbi Tannenbaum, Pam Rathbun, and Ralph Prahl. (2008). Framework for Self-Report Net-To-Gross (Attribution) Questions. Prepared for the Public Service Commission of Wisconsin.
39. Yin, Robert K. (1994). Case Study Research: Design and Methods. Newbury Park, California: SAGE Publications.

3. Calculating the Relative Precision of Program Net Savings

The EM&V Guidance recommends a 90/10 confidence and relative precision for both net and gross saving at the program level. These requirements apply to each fuel, electric and gas. These guidelines are designed to describe the basic approaches to estimating the relative precision of net savings at the program level at a reasonable level of rigor. In their EM&V plans, program administrators should plan sample sizes so that the 90/10 requirement is likely to be met. However, as the Guidance notes if this level is not realistic, the EM&V plan should clearly indicate the reasons it is not practical and offer a justification and alternative approach.

The overriding principle in this Appendix is that the requirements should not be too onerous, but should be sufficiently rigorous so that key stakeholders are able to make informed decisions about programs

In calculating the relative precision for net program-level electric and/or gas savings, the following general guidelines should be observed:

- Follow standard propagation of error formulas for the calculations involving addition, subtraction, multiplication, and division (Cochran, 1977; Taylor, 1997; TecMarket, 2004)
- For direct program gross savings involving engineering algorithms or energy use simulation models (e.g., DOE2), ignore the errors that propagate through the algorithms or simulation models. Only the standard errors (or relative errors) associated with the resulting program gross savings should be considered.
- For spillover savings, ignore the errors that propagate through a given engineering algorithm or energy use simulation model (e.g., DOE2). Only the standard errors (or

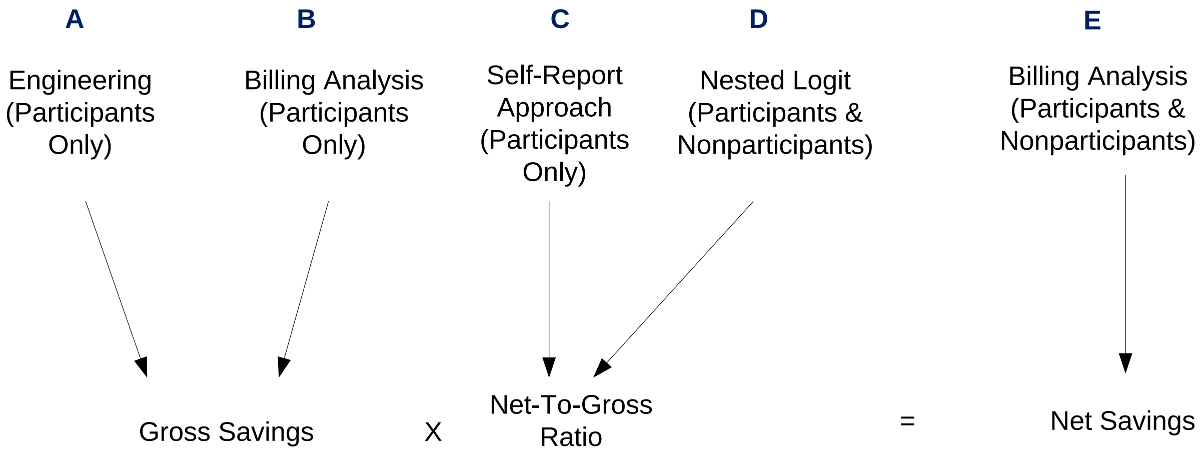
relative errors) associated with the net spillover savings should be considered.

- For net-to-gross ratios based on the self-report method, ignore the errors that propagate through a given NTGR algorithm. Only the standard errors (or relative errors) associated with the program NTGR should be considered.
- When using nested logit models to estimate net-to-gross ratios, use the standard errors (or relative errors) that are produced by statistical packages (e.g., Stata, SAS, Limdep, Gauss, etc.).
- When using regression models to estimate either gross or net savings, use the standard errors (or relative errors) that statistical packages (e.g., Stata, SAS, SPSS) provide.
- Show all data used and each step in the calculation.

Technical Discussion

This discussion has been prepared primarily for the technical analysts who will be carrying out the necessary confidence and precision calculations and is intended only as a framework for thinking about program-level confidence and precision. While we cannot anticipate all the possible methodological combinations for estimating net energy and demand impacts, the most common ones are illustrated in Figure 1.

Figure 1. Methods for Estimating Net Savings


A
Engineering (Participants Only)
B
Billing Analysis (Participants Only)
C
Self-Report Approach (Participants Only)
D
Nested Logit (Participants & Nonparticipants)
E
Billing Analysis (Participants & Nonparticipants)
Gross Savings
X
Net-To-Gross Ratio
=
Net Savings

From Figure 1, one can see that net savings can be calculated in three steps by first estimating the gross savings (using A or B), estimating the NTGR (using C or D), and then multiplying the two. Or, one can estimate net savings in one step using a participant and nonparticipant billing analysis (E).

Whether one uses the three-step or the one-step approach to estimating net impacts, there are sampling errors (random errors) that must be taken into account using propagation of error methods (Taylor, 1997; TecMarket Works, 2004). Propagation of error is the effect of variables' uncertainties (or errors) on the uncertainty of a function based on them. Consider

Equation 1:

Equation 1

Net Savings = NTGR x Gross Savings

Because the NTGR and the Gross Savings are based on samples, they contain some random error, or uncertainty. The uncertainty around the NTGR and the uncertainty around the Gross Savings propagate through the calculation to produce an uncertainty or error in the final answer, Net Savings.

Three-Step Calculation

Net impacts can be calculated in three steps by first estimating gross savings, the NTGR, and then multiplying the two to obtain net savings. The general equation for estimating ex post net program savings is presented in Equation 2.

Equation 2

$$\text{Ex Post Net Program Savings} = \text{Ex Post Gross Savings} \times \text{Ex Post NTGR}_{SA}$$

where

Ex Post Gross Savings = Gross savings estimated by the evaluator
NTGR_{SA} = Net-to-gross ratio adjusted for any possible spillover

The next sections discuss the calculation of relative precision when:

- Using a realization rate to adjust ex ante gross savings (i.e., savings claimed by a program administrator (PA) and recommending what errors can be ignored,
- Estimating the mean gross savings (instead of a realization rate),
- Using a participant billing analysis,
- Using the self-report approach or nested logit, and
- Estimating net savings.

Gross Program Savings

Ratio Estimation

A typical engineering approach is to calculate the ex post gross savings for a sample of projects and divide it by the ex ante savings for the sample of projects. This ratio is referred to as a realization rate. The ex ante gross savings for the population of projects is adjusted by multiplying the ex ante gross savings for each project in the population by the realization rate. When using engineering approaches, such as DOE2 or engineering algorithms, the errors in the individual parameters used to calculate ex post gross savings *can be ignored*. Only the relative error for the resulting realization rates must be calculated. There are standard formulas for calculating the standard errors of realization rates (i.e., ratio estimators) (Cochran, 1977).

Consider the use of an engineering algorithm used to estimate a realization rate for a sample of 100 participants who each installed one CFL in a residential rebate program. Site visits are made to each house to determine the wattage of the light that was removed and attach a lighting logger to each installed CFL. The algorithm for estimating gross savings for CFLs is presented in[35].

Equation 3

$$CFL\ Gross\ kWh\ Savings = \Delta Watts\ \times Operating\ Hours$$

There is some uncertainty contained in $\Delta Watts$ and Operating Hours since both are based on samples. These errors propagate through the equation to the estimate of average gross kWh savings for the sample. It is this error around the CFL Gross kWh Savings that can be ignored. The realization rate is then calculated by dividing the estimated average ex post savings by the average ex ante gross savings. It is the relative precision for the realization rate that must be calculated.

Or, in the case of DOE2 models, there are hundreds of engineering algorithms involving one or more of the four basic arithmetic operations (addition, subtraction, division, and multiplication) and terms around which there is some degree of random error. To track the propagation of error through all of these algorithms that would eventually contribute to the error around the final estimate of the energy savings associated with a particular building is infeasible. Again, the realization rate is then calculated by dividing the estimated average ex post savings by the average ex ante gross savings. It is the relative precision of the realization rate that must be calculated.

Mean Estimation

In certain situations, the ratio estimator might not be the best approach. In such cases, one could estimate the mean savings for a sample and extrapolate it to the population. When estimating the mean savings using engineering approaches (DOE2 or engineering algorithms), one could use the standard error of the mean using Cochran (1977) for various sample designs (e.g., simple random, stratified random, two-stage, etc.).

Participant Billing Analysis

When estimating relative precision for gross savings based on an analysis of participant bills, statistical packages such as Stata and SAS can produce the necessary standard errors for savings.[36]

[35] We recognize that most cases will involve measures for which the propagation of error calculation is far more complex particularly for custom measures. The simple case of CFLs is used only to illustrate the propagation of error principle.

[36] When gross savings are estimated by measure group or end use, the propagation of errors must be taken into account in calculating the error around the sum of the gross savings across measure group or end use.

Estimating Gross Savings by Measure Group

When estimating gross savings, using any of the above three approaches, by measure group or end use, the propagation of errors must be taken into account in calculating the standard error around the parameter of interest (i.e., the realization rate, the mean gross savings or the regression-based gross savings across measure group or end use).

Estimating Relative Precision for the NTGR

Self-Report Approach

If the NTGR is based on the self-report approach, then the standard error should be based on the distribution of NTGR estimates for the sample. If it is based on a more complicated quantitative approach involving, for example, a key decision maker and a vendor, then, depending on the NTGR algorithm, the propagation of error should be taken into account in calculating the standard error of the NTGR.

How to address the spillover rate is more complicated. Equation 4 shows the calculation of the NTGR_{SA} as the sum of the NTGR and the spillover rate. Equation 5 shows the calculation of the spillover rate.[37]

Equation 4

$$\text{NTGR}_{\text{SA}} = \text{NTGR} + \text{Spillover Rate}$$

Equation 5

$$\text{Spillover Rate} = \frac{\text{Net ISO} + \text{Net OSO} + \text{Net NPSOEx Post Gross Program Impacts}}{\text{Ex Post Gross Program Impacts}}$$

where

ISO = Net inside participant spillover
OSO = Net outside participant spillover (kWh or therms)
NPSO = Net nonparticipant spillover (kWh or therms)

Methods for calculating spillover vary in terms of the level rigor (standard and enhanced) and can involve multiple steps in the calculation involving multiple sample-based parameters, each having the potential for error. Again, as in the case of engineering algorithms, the error that propagates into the estimated spillover savings are ignored; only the sample error around the final estimates of spillover savings is considered in calculation of the relative precision. For example, consider 100 participants who received on-site visits to investigate spillover. For 20 of these, spillover was estimated using various engineering algorithms. The mean spillover is estimated for all 100 and extrapolated to the population of all participants and the standard error is calculated based on the sample of 100. Note that for 80 of these participants the spillover is zero while for each of the 20 the spillover is greater than 0. The errors involved in actually calculating the

[37] Core_NTGR = 1 - Free Rider Rate.

spillover savings can be ignored.

The total net spillover is simply the sum of program-level ISO, OSO, and NPSO. In the cases where all three types of spillover are found and that they are simply added, the relative precision for total net spillover is calculated using Equation 6.

Equation 6

$$\delta TotalSO = \sqrt{\delta ISO^2 + \delta OSO^2 + \delta NPSO^2}$$

where

$\delta TotalSO$ =Standard error of the total spillover (kWh or therms)

δISO = Standard error for the participant inside spillover[38]

δOSO = Standard error for the participant outside spillover

$\delta NPSO$ =Standard error for the nonparticipant spillover

If only two types of spillover are found, one type drops out. If only one type of spillover (e.g., ISO) is found, Equation 6 is unnecessary and can simply calculate the standard error for the one remaining type.

Given that Equation 5 is a ratio, the relative precision of the spillover *rate* could then be calculated using Equation 7.

Equation 7

$$rpSR = \sqrt{rp(TotalSO)^2 + rp(GS)^2}$$

where

$rpSR$ =Relative precision of the spillover rate

$rpGS$ = Relative precision of the total ex post gross savings

$rpTotalSO$ =Relative precision of the total program spillover

Given that Equation 4 is additive, the standard error of the NTGR_{SA}, which accounts for both free riders and spillover, is calculated using Equation 8.

Equation 8

$$\delta NTGR_{SA} = \sqrt{\delta NTGR^2 + \delta SR^2}$$

[38] The standard errors for IOS, OSO, and NPSO are each multiplied by 1.645 to yield the 90% standard errors (i.e., 90% confidence interval.

where

δSR = Standard error of the spillover rate

$\delta NTGR$ = Standard error for the core net to gross ratio, excluding spillover

$\delta NTGR_{SA}$ =Standard error for the total net to gross ratio including spillover

Nested Logit Analysis

When NTGRs are estimated using nested logit models, the propagation of errors can be taken into account across the participation and implementation models. Software such as Stata, NLOGIT and Gauss will produce the necessary standard errors (SBW Consulting and Pacific Consulting Services, 1995).

Estimating Relative Precision for Net Savings

The three step methods all involve calculating gross savings, the NTGR$_{SA}$ (either through self-report or discrete choice analysis), and multiplying the two to produce the estimate of net savings.

Since Equation 2 is multiplicative, the relative precision for the gross savings and the NTGR$_{SA}$ are used in Equation 9 to estimate the relative precision for the net program savings that takes the propagation of error into account.

Equation 9

$$RP\text{ Program Net Savings } = \sqrt{(rp(GS_{Program})^2 + \mathrm{rp}(\mathrm{NTGR}_{SA})^2}\,)$$

where

$rp(GS_{Program})$ = the relative precision of the gross savings

$\mathrm{rp}(\mathrm{NTGR}_{SA})$ = the relative precision of the spillover-adjusted NTGR$_{SA}$.

Another possible approach to adjusting for spillover is to calculate the NTGR without adjusting for spillover, multiply this NTGR by the ex post gross savings to obtain net direct program savings, and calculate the relative precision. Next, calculate the spillover savings and its relative precision. The net direct program savings can then be added to the net spillover savings to obtain the total net program savings for which the relative precision can be calculated using standard propagation of error formulas.

One Step Calculation

When estimating net program savings *in one-step* by incorporating participants and nonparticipants in a billing analysis, statistical packages such as Stata and SAS can produce the required standard errors. Note that this one step approach incorporates both the direct program savings as well as any participant spillover savings. Note also that this approach may penalize a PA since nonparticipant energy use might be lower than it would have been because of spillover.

References

1. Cochran, William G. *Sampling Techniques*. New York, NY: John Wiley & Sons, Inc., 1977.
2. SBW Consulting and Pacific Consulting Services. 1995. *1992-93 Nonresidential New Construction programs: Impact Evaluation Final Report*. Submitted to the Pacific Gas & Electric Company.
3. Taylor, John R. An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements. Sausalito: University Science Books, 1997.
4. TecMarket, Works. *The California Evaluation Framework.* Guidelines, Rosemead: Southern California Edison Company, 2004.
5. Train, Kenneth.(1980). Qualitative Choice Analysis: Theory, Econometrics, and an Application to Automobile Demand. Cambridge, Massachsetts: The MIT Press.

Appendix G - EM&V Report Guidance

Components of an EM&V Report

The outline below is designed to serve as a guide for preparing EM&V reports. Consistency in the development and presentation of EM&V reports will aid stakeholders in their review and understandings of EM&V activities across multiple program administrators. Evaluators should include all relevant components in their EM&V reports, however it is noted that the level of detail may differ depending on the scope and magnitude of the EM&V activity conducted.

Section	**Description/Components**
Executive Summary	Succinctly describe the EM&V objectives, methods, and results.
Main Report	Assembled with the following sections to readily enable all readers to understand the focus of the study and the results of the work, with some amount of methodology description: • Introduction • Evaluation Results • Conclusions and Recommendations • Methods
Appendices	Detailed methodologies or other information may be included in appendices, in order to keep the main body of the report brief and accessible to all users. Appendices may include, but are not necessarily limited to, the following detailed information that specific types of audiences or reviewers may wish to understand: • Glossary of Terms • Logic Model • Survey Disposition Information • Detail on Statistical Analyses • Other detailed information, as applicable, based on the study

Reports may vary from the suggested outline above, based on the nature of the work completed and the specific needs of the EM&V activity. Use of visuals, such as tables, graphs and images, along with short descriptions, when possible, is encouraged to maximize the accessibility of key information. Where visuals can be used in place of lengthy text, this approach is preferred.

Discussion of Evaluation Methods

In addition to complying with the methodological standards set forth in this EM&V Guidance, each report should include, where applicable, the following methodological details, which should be planned for in advance:

- **Approach to Estimating Savings.** Each important step in the estimation of key parameters, from data collection, to data cleaning, to construction of analysis datasets, to the analysis, and to final estimates, should be described in sufficient detail so that the analytical process can be understood by another analyst. Such understanding is essential in judging the reliability of the reported results. It is not necessary to discuss how each case was handled with respect to the various methodological issues. For example, with respect to outliers, evaluators should discuss how outliers were defined and, once those cases that met the definition were identified, how they were typically handled (e.g., deletion, weighting, etc.). These descriptions of the methods used to estimating gross and net savings should be included in appendices to the report.

- **Multiple Sources of Error** (See Figure 1, Appendix F). Depending on the data collection and analysis methods used, include a description of the efforts made in the planning and implementation of the evaluation plan to address the multiple sources of error including survey error and non-survey error, such as:
  - **Sampling Error**
    - The sample design (e.g., simple random, stratified random and two-stage)
    - For each key parameter (e.g., energy and demand, realization rates, installation rates, etc.): The achieved confidence, relative precision, and 90% confidence intervals Population size, Achieved sample sizes, Observed variance, standard deviations, and standard errors, and associated formulas. Provide a table containing the detailed disposition of the sample consistent with industry recognized standards, such as *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* developed by the American Association for Public Opinion Research (2009). The following rates may be reported: 1) Response Rate 1 (RR1) and 2) Response Rate 3 (RR3). Evaluators may report any other measures of survey outcomes that they think are important such as refusal rates, cooperation rates, and contact rates. Evaluators may choose to employ other recognized industry standards as well, specifying which is employed in the evaluation study being performed.
    - *If the sample design was stratified,* describe the methods that were used to determine strata boundaries and, if the sample is disproportionate, explain why and how the weights were calculated.
    - If any post-stratification was used, describe the methods used to determine

strata boundaries, how the weights were calculated.

- **Non-Sampling Error**
  - Measurement error: For example, report efforts to develop valid and reliable questionnaire items (e.g., multiple internal reviews, use of proven questions, etc.), pre-test questionnaires, minimize unit and item non-response (e.g., multiple call backs at different times of day, incentives, the use of experienced interviewers, etc.), calibrate instruments for field measurements, etc.
  - Non-response bias: The extent of any suspected non-response bias. There could be unit non-response in which only a subset of those targeted completed the survey. There could also be item non-response in which those who completed survey did not answer all the questions. Any suspected bias should be reported as well as methods and the assumptions underlying these methods to mitigate any bias.
  - Sample frame error: For example, report efforts to construct appropriate sample frames, the extent to which the effort was successful and what the implication are.
  - Data processing errors: For example, describe the development of QA/QC processes to ensure accurate collection and storing of data.

- **Non-Survey Error**
  - Modeler error (statistical and engineering): For example, describe efforts to provide guidelines for calibration of DOE-2 models using customer billing data or efforts to insure that regression diagnostics were routinely conducted by all modelers.
  - Internal and external validity[39]. In studies where the effort is designed to test causal hypotheses, describe how the selected research design addresses both internal and external validity.
  - Self-selection: Self-selection is such an important threat to internal validity that it deserves special mention[40]. Discuss the extent to which self-selection effects are believed to pose a significant threat to the internal validity of key findings, providing both empirical findings and/or theoretical reasoning to support the conclusions reached. If self-selection effects are believed to pose a significant threat to validity, explain how these were addressed.

[39] Internal validity refers to inferences about whether the experimental treatments made a difference in a specific experimental instance, i.e., it addresses the causes of the outcomes that you observed in your study. External validity addresses the ability to generalize the results of a study to other people and other situations (Shadish, Cook and Campbell, 2002).

[40] Self-selection refers to situations in which subjects decide the treatment condition they will enter. Self-selection bias can occur when selection results in differences in subject characteristics between treatment conditions that may be related to outcome differences (Shadish, Cook and Campbell, 2002).

- Choosing an appropriate baseline: Describe the baseline chosen and why it was chosen.
- Statistical conclusion validity: In studies where the effort is designed to test causal hypotheses, describe why the statistics used to establish whether the independent and dependent variables co-vary are appropriate.

Data Documentation: The data associated with statistical and engineering approaches should be documented. Datasets should contain descriptions and the role the data played in the analysis or estimation of savings.

References

1. Shadish, William R., Thomas D. Cook, and Donald T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.
2. The American Association for Public Opinion Research. (2009). *Standard Definitions: Final Dispositions of Cases Codes and Outcome Rates for Surveys: Revised 2009*. http://www.aapor.org.
3. Yarbrough, Donald B., Lyn M. Shulha, Rodney K. Hopson, and Flora A. Caruthers. (2011). *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*. Los Angeles, CA: SAGE Publications.