DNV·GL

FINAL REPORT

Massachusetts Commercial and Industrial Gross Impact Evaluation Framework

Massachusetts Program Administrators and Energy Efficiency Advisory Council

Date: May 19, 2017

Table of contents

List of figures

List of tables

1 PURPOSE

This report serves as a statewide framework to refine and more fully reexamine and document the approach to be used to determine which Massachusetts commercial and industrial (C&I) impact evaluation studies to undertake, at what level of rigor, and when. This framework represents a fundamental reexamination of traditional impact evaluation methodologies, the changing needs of programs and program administrators, and the role of evaluation in responding to those needs.

This effort is an outgrowth of the substantial impact evaluation planning and documentation work undertaken to date by the Massachusetts Program Administrators (PAs), the Energy Efficiency Advisory Council (EEAC) Consultants, and the non-residential team. The specific issue of prioritizing and methodizing impact evaluations and related efforts has been a key topic of interest for years in: annual planning meetings since 2011, in the development of Massachusetts statewide electric and gas energy efficiency evaluation plans since 2013, and in impact evaluations of Massachusetts energy efficiency programs for more than two decades.

Throughout its history, Massachusetts impact evaluation decisions have utilized sound logic, but the framework behind these decisions has never been formally documented. The PAs and EEAC Consultants recognized the need for a documented decision-making framework to maximize the value of impact evaluation studies now that the statewide C&I programs and evaluation framework have matured.

This framework is a guidance document that focuses on impact evaluation planning, structuring, and decision making. This guidance applies to all C&I measures irrespective of fuel (electric vs. gas), delivery track (prescriptive vs. custom), delivery mechanism (upstream vs. downstream), and other program elements.

The primary audience for this document includes parties who oversee, plan, and execute C&I impact evaluations for Massachusetts PAs. The framework authors foresee a potential secondary audience in PA implementation staff seeking to understand evaluation decision making, structures, and how impact evaluations will affect program savings.

Framework Organization

This framework is structured in such a way to answer the fundamental objectives of this study, which is to explore and document any refinements to impact evaluation in Massachusetts and to document a systematic approach to impact evaluation planning. The framework has three primary content sections yet there is some inherent overlap between them. Below, we briefly describe each section and its primary focus.

Section 3: Refinements to Massachusetts Impact Evaluation answers the four fundamental research questions outlined below that drive to the issues of timing, new evaluation structures, how to handle baseline, net-to-gross, and measure life, and early EM&V involvement. Each of these questions are explored and specific recommendations are provided to make each actionable.

Section 4: Systematic Impact Planning Process zooms out from the detail of Section 3 and focuses on higher level impact planning processes. While some of the details from Section 3 are incorporated here, the purpose of this piece is to document how impact evaluation planning can be done, including the structural challenges present, research categories and key indicators of interest to impact evaluation.

Section 5: Impact Evaluation Tool box provides an overview of ways in which this framework can maintain a repository of program data and various evaluation methods. The dataset, which is the backbone of the tool box, can be used as an aid for impact evaluation planning decision making. Using program participation data, impact evaluation results, expected changes in measure life and/or baseline, and other indicators, stakeholders can explore possible gaps in research. This tool box is not intended to replace the decision-making process, but to inform it. Depending on need, these data can help identify where impact evaluation work should be done, if ISP research is needed, or how changes to measure life might impact savings.

Interaction with the Baseline Framework

The Massachusetts C&I Baseline Framework (Baseline Framework) is focused on baseline assessment and characterization as part of impact evaluation efforts. The Baseline Framework provides guidance on determining the appropriate measure baseline against which ex post gross impacts should be calculated, and is thus an important resource for establishing a necessary part of the gross impact equation.

This Massachusetts C&I Gross Impact Evaluation Framework (Impact Framework) offers insights into how impact evaluations research and handle baseline and measure life, and sometimes trigger further research.

Maintenance of Living Framework

This Impact Framework is intended to be a living document that can be updated on an as-needed basis and used annually as part of the impact evaluation planning process. The authors recommend that this framework and associated tools are reviewed by stakeholders each calendar year to incorporate new material and revisions. The depth of review can depend upon needs and priorities.

2 ROLES AND RESPONSIBILITIES

This Impact Framework involves and serves multiple parties and stakeholders, including EEAC Consultants, evaluation-focused PAs,[1] implementation-focused PAs, and Evaluation Contractors. A cursory overview of these parties and their primary responsibilities follows.

- **EEAC Consultants - Evaluation**: These are third-party expert consultants who have primary responsibility for working with the PAs to plan and implement high-quality evaluation, measurement, and verification (EM&V) in Massachusetts on behalf of the EEAC. They also represent the EEAC's oversight role to ensure objectivity, independence, consistency, timeliness, and credibility.[2] Special insights include priorities of the EEAC, regulatory, and policy issues, and national practices.
- **Program Administrator – Evaluation (PA-E)**: The PA-E has primary responsibility for ensuring high-quality EM&V in Massachusetts on behalf of the PAs. The PA-E provides access to data and serves as an intermediary between Implementation personnel and Evaluation Contractors. Special insights include priorities of the PAs, EM&V in the context of the PAs' regulatory reporting requirements, and PA evaluation history.
- **Program Administrator – Implementation (PA-I)**: The PA-I manages program delivery on behalf of PAs, provides documentation on efficiency installations, and occasionally serves as liaison between the customer and the Evaluation Contractor. Special insights include upcoming changes in customers, markets, program offerings, technologies, and delivery.
- **Evaluation Contractors**: Evaluation Contractors design and perform evaluation and related research and reporting, including impact evaluation, baseline studies, market research, technology assessments, and demand analysis. Special insights include trends, industry best practices, and innovative techniques from other jurisdictions.

[1] While a clear division between evaluation and implementation is typical, not all PAs have such a clear division.

[2] "2016-2018 Massachusetts Joint Statewide Three-Year Electric and Gas Energy Efficiency Plan," October 15, 2015, p.244,

Figure 1. Roles, responsibilities, and relationships

Since its inception in 2008, the EEAC has been charged with managing the EM&V process of energy efficiency programs in the Commonwealth of Massachusetts. In about 2009, the number of active stakeholders in evaluation increased to three: EEAC Consultants, Evaluation-focused PA staff, and Evaluation Consultants.

The 2016-2018 Massachusetts Joint Statewide Three-Year Electric and Gas Energy Efficiency Plan ("MA Three-Year Plan") stresses the importance of programs being evaluated "in a way that provides confidence to the public at large that the savings are real and in a way that enables the Program Administrators to report those savings to the Department with full confidence."[3] This is a key reason why PA Implementation traditionally has not had an active role in evaluation planning. Historically, the PA-E serves as a liaison between the PA-I and the Evaluation Contractors to funnel information while maintaining objectivity.

[3] Ibid.

3 REFINEMENTS TO MASSACHUSETTS IMPACT EVALUATION

This framework seeks to refine two primary elements of impact evaluation:

- The structure and timing of impact evaluation studies and related research
- The portfolio-level impact evaluation planning process

The EEAC Consultants clearly indicated that the first item would be the priority for this framework. The other matter of systematizing the evaluation planning framework at a process level is of secondary importance. Within both of these general objectives reside additional research interests that are also reflected in this Impact Framework.

The research stages for this project involved the following:

- A questionnaire that preceded the Stakeholder Engagement Workshop, which sought to understand desired workshop outcomes and set priorities for that meeting
- A stakeholder engagement meeting that was focused upon those priorities, pursuing core research questions, gathering information from attendees, and maximizing the value of the event
- Secondary research on historical impact evaluations in Massachusetts, recent Massachusetts C&I tracking data, and impact evaluation practices from other jurisdictions

Soon after project initiation, the EEAC Consultants posed four specific research questions that they wished this Impact Framework to pursue. The questions effectively captured the central themes of the Impact Framework while guiding the research and potential outcomes. The four questions are:

Q1. In what ways can impact evaluation research be structured to reduce the length of time that parts of the portfolio go unevaluated?

Q2. Is there an opportunity to integrate staged, "rolling," or reconnaissance style evaluation into the impact evaluation framework?

Q3. How should impact evaluations research and handle baseline and measure life?

Q4. Are there opportunities to incorporate ex ante M&V and other kinds of early involvement into impact evaluation?

These four questions served as the backbone for much of the research detailed herein.

3.1 Research tasks

3.1.1 Pre-workshop questionnaire

In May 2016, DNV GL implemented a pre-workshop questionnaire designed to improve understanding of stakeholder priorities for this Impact Framework as well as desired workshop outcomes. Researchers used the preceding questions in this informal survey of stakeholders. All Massachusetts PAs were invited to answer the questionnaire, and the stakeholders included personnel from all four categories listed in Section 2, Roles and responsibilities. A total of 28 requests were sent out, yielding 15 responses.

Respondents were asked to rate the priority of each of these research questions on a scale of 1 (lowest) to 4 (highest). Due to the low number of responses and informal survey design, detailed analysis was not feasible; however, broad insights did emerge (see Table 1). The top priority (average = 3.13) was Question

4 regarding ex-ante M&V and early involvement. The lowest of the four priorities (average = 1.93) was Question 2 about integrating some specific evaluation styles into the impact framework. Questions 2 and 3 received similar average priority scores.

Table 1. Pre-workshop questionnaire responses

Key research question	Distribution of responses				Average of priority
	Lowest Priority=1	Priority=2	Priority=3	Highest Priority=4	
Q1 – Evaluation Structure and Timing	2	6	4	3	2.53
Q2 – Staged, Rolling or Reconnaissance	8	2	3	2	1.93
Q3 – Baseline and Measure Life	4	4	4	3	2.40
Q4 – Ex Ante and Early Involvement	1	3	4	7	3.13

The survey concluded with some brief open-ended questions to identify desired workshop outcomes and to learn about aspects of Massachusetts impact evaluation that currently are working well or warrant some improvement.

Survey respondents expressed satisfaction in these two areas:

- Effective communication and collaboration between evaluators, PAs, and EEAC Consultants
- Comprehensiveness, rigor, diligence, and reliability in evaluation results

The constructive suggestions for improvement included the following:

- Improve evaluation documentation and communication with non-evaluators
- Reduce evaluation lag, increasing timeliness and quicker feedback
- Improve consistency in the handling of baseline and measure life
- Shorten the lengthy planning process
- Improve applicability of results to future program delivery
- Address concerns about applicability of results to smaller PAs

All of these themes received attention at the subsequent stakeholder engagement workshop.

3.1.2 Stakeholder workshop

On May 24, 2016, the DNV GL team facilitated a stakeholder workshop on key issues in C&I program impact evaluation at the Columbia Gas office in Westborough, Massachusetts. The focus of the workshop was on optimizing the structure, timing, and staging of evaluation and related research. Participants included representatives of PAs, the EEAC, and the DNV GL team.

The workshop agenda was structured around discussion stations for each of the four core research questions listed above, plus a fifth station to capture additional important topics raised by stakeholders. Each discussion station was facilitated by an evaluation consultant and a scribe. Each topic was reported out to the full participant group, and the half-day workshop closed with a prioritization exercise to identify key findings and next steps.

On June 15, 2016, the DNV GL team submitted “Refinements to Gross Impact Evaluation Framework (P63) Workshop Summary Memo” to the PAs, EEAC, and other stakeholders that summarized the priorities expressed in the event. There was a general clustering of ideas relating to implementation involvement and

working as more of a partner with implementation. There were also several relating to incorporating a rolling approach, and some relating to ex ante metering. Other ideas involved useful life research, being articulate/clear and reaching a practical/actionable end point, speeding things up, and baselines.

Additional detail on the stakeholder engagement meeting is presented in APPENDIX A of this document.

3.1.3 Literature review

The literature review was reasonably broad in scope, and focused on identifying existing frameworks, guidance documents, papers and other resources that are aligned with the research objectives of this Impact Framework. It was hoped that an investigation of the impact evaluation practices conducted in other jurisdictions would shed some light on refinements that could be made in Massachusetts to facilitate more timely and useful evaluation results.

Relevant findings from other jurisdictions were minimal.

- Many of the impact evaluation framework or guidance documents currently in existence are of a different nature than this Impact Framework. The largest documents are EM&V protocols that detail specific evaluation methods and techniques that are permissible in the jurisdiction.
- Existing energy-industry guidance documents are largely deficient on some of the key elements sought, such as rolling and continuous evaluation, interactivity of evaluation and baseline or measure life research, and ex-ante involvement.

The following two quotations from the literature review support themes within this Massachusetts Impact Framework:

> "While the value of program evaluation is well established, the questions of who should do what, how (rigor level and consistency) it should be done, and when (rapid versus after-the-fact feedback as well as recurring studies) are far less well defined."[4]

> "The evaluation process should be integral to what is typically a cyclic planning- implementation-evaluation process. Therefore, evaluation planning should be part of the program planning process so that the evaluation effort can support program implementation, including the alignment of implementation and evaluation budgets and schedules, and can provide evaluation results in a timely manner to support existing and future programs."[5]

At least one jurisdiction has experimented with rolling/concurrent evaluation sampling, and a Massachusetts evaluation contractor indicates that the next New York Evaluation Guidance document will add brief sections on rolling sampling and ex-ante evaluation involvement. The updated New York document is months from issuance and the guidance will be of limited specificity, although it will encourage both practices. However, an evaluation contractor who has implemented some rolling/concurrent sampling in New York has contributed those insights to this Massachusetts Impact Framework.

A listing of the literature review resources and some additional quotations are contained in APPENDIX B of this document.

[4] Ontario Power Authority, "Evaluation, Measurement and Verification (EM&V) Protocols and Requirements" (2015), p. v.

[5] Steven R. Schiller and Charles A. Goldman, "Developing State and National Evaluation Infrastructures - Guidance for the Challenges and Opportunities of EM&V" (2012), p.9.

3.2 Exploration of core research questions

Program evaluation as a research discipline predates energy efficiency by decades, and much can be learned from the evaluation of social and educational programs. Michael Scriven's "The Methodology of Evaluation" (1967) characterized evaluations into two types: summative and formative. Summative evaluations are largely quantitative, tend to focus on outcomes, investigate "how many/much," and culminate in a retrospective judgment of an intervention's value. Formative evaluations, on the other hand, are more qualitative, tend to focus on improvements, pursue "why," and serve to inform adjustments and corrective action.

A body of evidence from repeat summative feedback—multiple snapshots of evaluated performance over time—informs a formative assessment with which stakeholders can adjust and improve programs. The shorter the period between evaluations, the more actionable and prospectively relevant the findings become. This view of energy efficiency program evaluation in the context of summative and formative terminology reinforces the value of shifting toward more frequent, if not continuous, program feedback.

3.2.1 Evaluation structure

The first research question explores whether impact evaluations can be structured differently in pursuit of reduced time intervals between evaluations. At the workshop, stakeholders also discussed unevaluated or under-evaluated measures within this conversation topic.

> *Q1. In what ways can impact evaluation research be structured to reduce the length of time parts of the portfolio go unevaluated?*

There are two aspects to this question.

- How can evaluations start sooner and/or the length of the evaluation period be shortened to conclude closer to the installation of the measures?
- How should lower priority elements of the program be incorporated into planning?
    - Some lower priority measures or end-uses have not been evaluated in many years, if at all. Some overarching issues like persistence have received lower attention as well. While perhaps this approach is deliberate and reasonable, the planning process should address and document these elements nonetheless.

When first considering this question, it is common to contemplate the maximum allowable time frames for a particular item to go unevaluated. This has been fundamental part of the Massachusetts impact evaluation planning paradigm for a long time. Some evaluation stakeholders have an intuitive range of years across which a given grouping of energy efficiency offerings should be evaluated: extremely stable, known, predictable measures can go 5 years or longer between evaluations, while offerings with significant uncertainty have been thought to warrant annual examination. This "turn-based" approach remains an important consideration and is incorporated into the Impact Evaluation Tool Box.

But sometimes there are specific research questions of paramount interest to PAs for which answers are needed sooner than the standard reporting cycles permit. In those cases, it is not about impact planning or schedules, but rather the issue of an information and feedback gap relative to expectations or some optimal scenario.

Barriers to timeliness

Workshop participants identified two significant issues that contribute to time delays but may be able to be addressed with attention. ***Resource availability*** was cited as a constraint when there are insufficient specialized staff available to perform work at the desired time and/or pace. This is partly a product of the limited size and specialized nature of the EM&V industry, but also highlights the need for a personnel availability overlay in the evaluation planning process. This constraint is not exclusive to evaluation contractor personnel and also can include PA staff, equipment, or customer availability. ***Data delays*** can be very significant within the MA evaluation planning process. Currently, there is nearly a 6-month lag for impact evaluation activities due to the need for formalized close-out of program year tracking data, acquisition of participant and measure-level data from all PAs, and aligning disparate data in a common frame. ***Cyclical evaluation*** has been the predominant evaluation structure inherently driving slower evaluation starts and longer durations. One way to reduce the time between studies may be to depart from annual evaluation (not necessarily planning) cycles in situations that warrant more rapid feedback. More long-horizon targeting of evaluation completion dates and "working backward" is likely necessary to reduce the time between installation and evaluation results. Current targeting tends to focus more on regulatory filings. Prospective targeting may focus on evolving program offerings and timing of beneficial feedback.

Research gaps

Measure life and baselines were identified as two areas where the current evaluation framework has ***research gaps***. Impact evaluation is a logical vehicle with which to collect information and seed further research within these parameters. Some market conditions that affect program performance are evolving rapidly, including ***changes in technologies and offerings***. These changes are happening at quite different rates across various sectors and technologies and do not always align with a regular or repeating evaluation cycle. Implementers may find evaluation feedback extremely valuable in driving mid-course program corrections if this feedback is timely. Finally, although process, impact, and market studies are all part of the current/recent past picture, the concept of ***feedback cadence*** may have been lost. For feedback from evaluations to be most useful to the programs, it has to align with the corresponding program roll-out cycle, or "cadence." Many energy-efficiency stakeholders desire faster feedback, and the program evaluation industry is working hard on solutions. The point of cadence, perhaps irrespective of speed, is to plan and time the feedback such that it arrives in actionable form and at opportune moments in a program cycle to incorporate it and drive program improvements.

Sometimes research questions arise that need immediate/short-term answers to feed both program management and further evaluation coordination and planning. There are also mid-planning cycle term questions as well as long-range research needs. Revised evaluation structures may be able to accommodate these needs in parallel with the current focus on delivering evaluated results for major savings categories in a timely manner.

Paths forward

Workshop participants considered some potential approaches to expedite evaluation timing, including:

Rolling evaluation. Instead of drawing sample points after the close of program year, a smaller rolling[6] sub-sample of projects could be evaluated each time period, leading up to a finite total sample evaluated projects. The rolling sample helps to disengage the field and analytical planning work from the

[6] Rolling evaluation is defined and discussed at greater length in the following Section 3.2.2: Staged, rolling, or reconnaissance style evaluation.

regulatory reporting timetable, which could allow it to become an ongoing enterprise working away efficiently and independently in the background. In additional to evaluated measure performance, a rolling sample of projects could serve as a continuous feed of observations and insights that may be useful to program administrators on topics such as measure life, persistence, and baseline. Such information, even if observational or anecdotal, has value when the feedback is more timely.

Decouple impact population from annual data aggregation and reporting cycle. One reason for impact evaluation delay is the year-end data transfer into the Massachusetts annual customer data profile, which strives to match the PAs' annual reporting savings numbers. The need for final tracking data may be an unnecessary bottleneck for impact evaluation. Impact samples can be developed using less-than-perfect PA tracking savings data and final samples can be adjusted at a later data. The opportunity to draw impact samples much earlier or more frequently may outweigh the risk of non-finalized data changes, which can be mitigated with post-stratification.

Align evaluation with program cycle. Evaluation can provide more actionable program feedback if better aligned with the program life-cycle. Informally, best practice studies strive to align with program design or pilot development, and process evaluations are often timed to coincide with early launch or the first year or two of program delivery. Likewise, impact evaluation results align well with mature delivery and likely program update intervals. Some PA implementers have indicated that the historical evaluation framework is not optimized toward generating timely, or well-timed, feedback and actionable findings.

Deploy focused studies. One suggestion from the workshop is the introduction of a new "focused study" type that is detached from the traditional reporting cycle and instead deployed to provide rapid, interim feedback directly to PAs on a variety of focused topics. The studies could have narrow objectives or also feed into the regular cycle of impact studies. A fundamental need exists for research that is focused on individual parameters or measures rather than overall impacts. A focused or specialized study type could be introduced that facilitates more rapid *dispatch* – referring to the collaborative design and decision making process. When timeliness of results is critical, specific and narrowly-scoping focused studies may be beneficial, otherwise focused studies could combine various objectives to address multiple research questions simultaneously.

Increase the planning horizon. Multiyear sampling benefits from a longer planning horizon in which stakeholders make a level of commitment to the evaluation method and a plan for pooling results over time. This does not necessarily imply that a decision to employ a given sampling strategy is final or irreversible, but it does represent a practical implication of researching smaller evaluation samples more frequently (e.g. annually) versus larger samples at longer intervals (e.g. triennially). Another benefit of a longer planning horizon is a more proactive and manageable pace for work plans to be considered and approved, and to be ready for tracking data as it is available. A more preemptive approach to upcoming program and technology changes can improve research lag for evolving program offerings and add value to our implementation stakeholders, e.g. industry standard practice and measure life studies. A longer time horizon, in theory, could even out the workload for the all the parties and help reduce bottlenecks. Consulting companies typically have business processes to manage resource assignments and availability, and a longer view of evaluation needs would help reserve consulting staff and identify shortfalls in time to acquire additional resources.

There is general optimism that Massachusetts evaluation stakeholders can shorten the time between studies through a variety of these and other alternative approaches. The next research question elaborates on this concept and explores the potential to utilize some specific evaluation structures.

3.2.2 Staged, rolling, or reconnaissance style evaluation

The second research question follows from the first question on restructuring evaluation to reduce duration between implementation and evaluation results. Several evaluation structures that are not widely employed in Massachusetts are suggested in the following inquiry:

> *Q2. Is there an opportunity to integrate staged, "rolling," or reconnaissance style evaluation into the impact evaluation framework?*

This question introduces some terminology that warrants definition. Since researchers did not locate much documentation on these methods, definitions were developed by veteran evaluators with statistical consultations.

Working definitions

Traditional evaluation in this framework refers loosely to the current, typical approach to impact evaluation performed in Massachusetts, particularly but not exclusively with respect to commercial and industrial measure offerings. For the sake of this discussion, which uses this terminology mainly as a reference point to the next definitions of Staged, Rolling, and Reconnaissance, Traditional evaluation is fundamentally a single-year approach. Once the PAs have finalized a full calendar year of participation, evaluators request, obtain, and clean the data, develop a research sample and work plan, and initiate EM&V on that year's cohort of participants once the prerequisite steps are complete. This approach is repeated for that type of participant (end-use, program, etc.) at different intervals. Traditional evaluation is depicted schematically in Figure 2; specific months of activity may vary.

Staged evaluation involves the periodic gathering of data after a short interval with an independent sampling conducted for that interval. The independent samples from multiple intervals can be combined using propagation of error methods. Traditional cyclical impact evaluation in Massachusetts may be viewed as "staged" wherein the period is one year. Figure 2 depicts staged evaluation of one participation year with quarterly data intervals.

The periodic nature of staging lends itself to more applications than just staged M&V. Staging could involve desk reviews with periodic examination of findings to trigger field M&V steps.

Staging could also involve EM&V on various efficiency offerings, e.g., examination of custom HVAC followed by custom lighting. This too is fairly traditional in Massachusetts but there are opportunities to be more creative with the selection of research intervals than just using calendar years.

Staging could also pertain to the deployment of different studies. Consider an approach that begins with market baseline research, that informs new efficiency projects, tracks participants through efficiency treatment, samples for evaluation, and that concludes with impact evaluation and lifetime research. Each stage may be a separate deliverable within a continuous research stream.

Quarterly staged evaluation is typically less efficient than other techniques for the combined results given the data overhead. Requesting, fulfilling, and compiling data from multiple program administrators, multiple times per year, may be a significant enough drawback that it outweighs the benefits.

Multiyear staged evaluation is a modified form of staged evaluation that offers several advantages, one of which is relative ease of adoption, and the challenges of introducing this method are readily surmountable. Not to be confused with *Rolling* evaluation which uses a sampling forecast, *Multiyear* uses traditional sampling techniques with lesser annual precision targets that are pooled across multiyear stages of EM&V activity. More similar to quarterly Staged, just with different time intervals, in Multiyear the periodic data

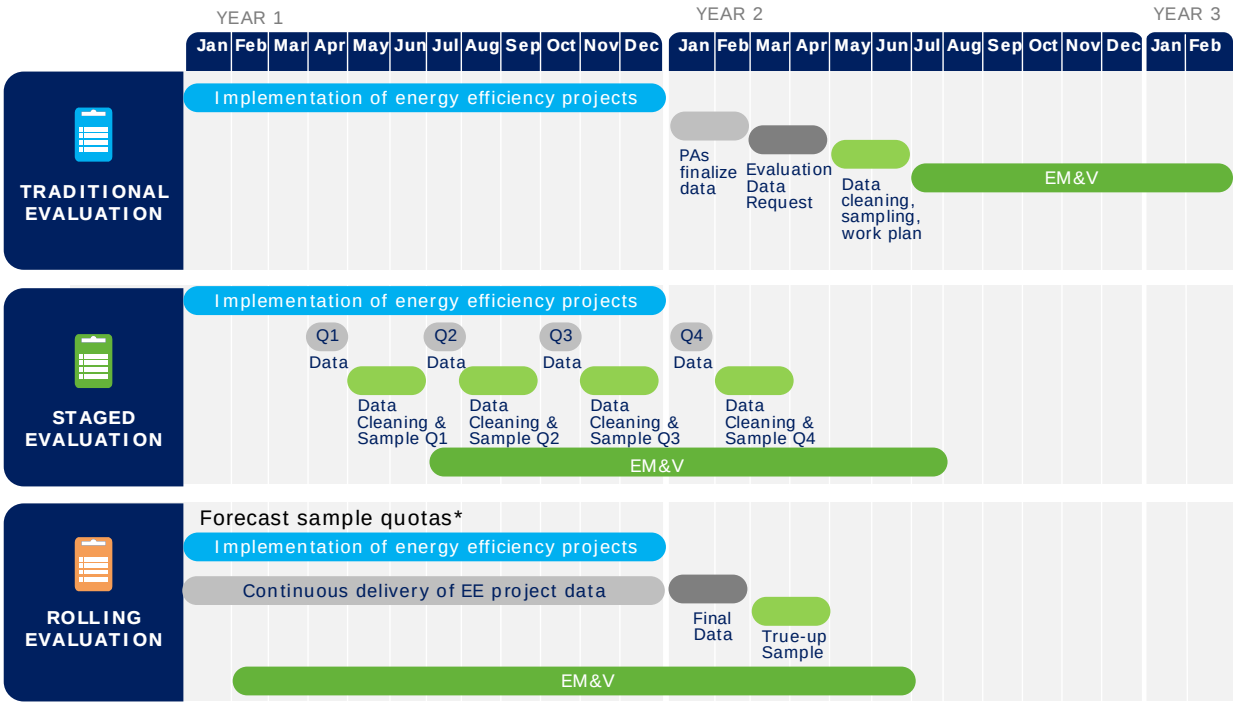
interval is a single calendar year, and the EM&V duration spans several years. This has the distinct advantage of eliminating burdensome quarterly data processes, reducing EM&V intensity and surges, and extending the duration of field data collection activities. Once a multiyear evaluation scheme is established, the multiyear window can "slide" to bring in a new participation year while an older year drops out of the results pool. Multiyear evaluation is not presented in Figure 2 because it would extend the time scale several years and impede readability.

Rolling evaluation also involves the periodic gathering of data after a short interval(s); however, it employs a single sample design across the interval(s). This approach requires a sample design using a proxy population forecast that leverages proxy data or other intelligence. Sites are drawn using probability of selection, and a final adjustment would occur once the full period of interest is complete. This approach lends itself to hybrid designs such as "continuous rolling" (across multiple years) or a rolling-staged hybrid with quarterly true-ups. This can be a more efficient sample design than a staged one, but it requires a reasonable proxy population as its basis.

Reconnaissance evaluation involves the gathering of data, either during or after an interval, and does not necessarily require an upfront or robust sampling framework. Evaluators can use exploratory samples to develop a statistically *insignificant* picture of performance that stakeholders can use to inform both evaluation and program decisions. In evaluation, a key potential use of reconnaissance work would be helping to inform a prioritization of studies and selection of rigor. Reconnaissance research adds opportunity to preemptively reveal unexpected or underperforming measure savings that may otherwise have gone unnoticed for a longer duration. Random site visits or desk reviews can also just spot check hypotheses for tolerance to expected performance or other ad hoc inquiries. Reconnaissance evaluation is not shown in Figure 2 because it is by definition exploratory and any depiction would be very arbitrary on a time scale.

Figure 2 presents a visual representation of these the three primary evaluation structures to differentiate the main differences, particularly with respect to when the EM&V cycle begins relative to the implementation year of interest. This graphic depicts the basic steps and interdependencies within each of the Traditional, Staged (quarterly), and Rolling evaluation schemes. Please note that this illustration is schematic in nature and does not depict specific expectations for the amount of time that each step takes.

Figure 2. Visual representation of three evaluation structures



*Prior to start of Year 1, develop sample quotas based on prior year

The graphic above does not reflect important considerations such as timing EM&V activities to capture seasonal, post-installation performance. The EM&V cycles depicted for the Staged and Rolling evaluation types may need to extend to capture a sufficient amount of summer weather. Finally, quarterly (Staged) or continuous (Rolling) data transfers will be challenging for program administrators and evaluators alike until efficient data processes are developed.

Comparison of evaluation structures

The five structures defined above all have merits as well as some potential disadvantages. The following table strives to capture the primary pros and cons of each evaluation structure. Since most of these structures are untested in Massachusetts, some of listed attributes are theoretical or informed by stakeholder experience in other markets. This living framework will document a more detailed and comprehensive list as newer structures are tested or deployed.

Table 2. Some high-level advantages and challenges of each evaluation structure

Evaluation Structure	Advantages	Challenges
Traditional	• Established and relatively smooth process. • Based upon final PA data; fewer adjustments; more efficient data cleaning. • Timing can permit post-installation, summer EM&V. • Duration of EM&V activities can be shorter than other structures. • Post-EM&V statistical results require fewer or no adjustments.	• EM&V does not start until midway through following year • PAs finalize data at different rates; evaluation dependent upon latest data delivery. • Site-specific EM&V planning and approvals can delay field work until late summer. • Time-intensive EM&V activities can create resource constraints on other projects. • True-ups remain in instances of sample point substitution or drop outs.
Staged (Quarterly)	• Evaluation cycle may begin and end sooner than Traditional. • Sampling can begin as soon as data is available for a quarter. • EM&V can begin about midway through an implementation year • Longer duration EM&V improves resource usage and reduces surge. • Quarterly data may incorporate new PA cleaning from prior quarter, lessening year-end data cleaning.	• Staged approach is yet untested in Massachusetts and adds data complexity. • Quarterly data delivery may be challenging for PAs; some risk of over/under sampling. • EM&V may not begin in time for sufficient summer metering. • Participation 'hockey stick' may create EM&V surge in Q3 and Q4 regardless. • Q4 data cleaning and sampling may be time consuming to cross-check against final year end PA data.
Staged (Multiyear)	• Multiyear staged evaluation is easily evolved out of current methods. • Based upon final annual PA data; lesser annual precision targets. • A multiyear EM&V strategy can facilitate more seasonal EM&V. • Longer duration EM&V improves resource usage and reduces surge. • Statistically valid results each year, albeit with lower precision, e.g. 25% precision @ 80% confidence. • Once a multiyear cycle is built, the oldest year can drop from pooled results as a new year joins.	• Results before end of multiyear period may not yet achieve precision targets. • Some risk of over/under sampling in a given year with respect to a multi-year objectives. • First year EM&V may not begin in time for sufficient summer metering. • Unless redesigned, it will take the full multiyear period to yield targeted precision, e.g. 10% precision @ 80% confidence. • A foreseeable complication to the "sliding" multiyear cycle may involve how to handle anomalous activity or results.

Rolling	• Evaluation cycle may begin and ends sooner than Staged. • Sampling plan can be done well in advance thus not delaying EM&V. • EM&V can begin shortly after the first projects are implemented. • Rolling evaluation may facilitate some ex-ante metering. • Longer duration EM&V improves resource usage and reduces surge. • A single set of year-end PA data may result in more efficient data processing than Staged.	• Rolling approach is also untested in MA and will require continuous data delivery. • Forecast-based sampling carries risk of under/over estimating participation. • EM&V may permit some post-installation summer metering but maybe not enough. • Ex-ante involvement will require new collaboration between stakeholders. • Participation 'hockey stick' may create EM&V surge near year end regardless. • True-up EM&V samples are likely unless participation is accurately forecasted or oversampled.
Reconnaissance	• An exploratory technique for verifying performance or probing for non-conformities • Fast dispatch since robust sampling design unnecessary. • Can start midstream and end once objectives are met. • Low cost, low overhead preemptive tool to discover otherwise unnoticed issues.	• Low statistical rigor not well-suited for standalone evaluation of significant portfolio segments. • Can be statistically insignificant or of minimal value beyond specific projects.

The risks and rewards of these different strategies are somewhat academic and untested at this juncture and should be codified in the living impact framework. Strategies can be tested, refined, and discarded if necessary.

Additional pros and cons of non-traditional structures

A major driver of these "non-traditional" evaluation structures is time: reduced time between evaluations and exploring opportunities to shift some evaluation results closer to the time of installation. In addition to the specific advantages outlined in Table 2 above, a few additional benefits seem to cut across most if not all of the non-traditional evaluation structures:

Correctable course. Programs might have the opportunity to make course corrections when issues are identified in early samples. Implementation and demonstration of corrections create the opportunity to increase program delivery's perceived ability to influence the realization rate.

Knowledge availability. Review of project materials at a time closer to project completion would reduce the need for evaluation to ask questions related to options and decisions made many years after projects are completed. Under the current paradigm, evaluators encounter situations where there are no personnel at the site who were involved with the energy efficiency project. Shortening the time gap between project completion and evaluation allows for gathering information while it is still fresh.

Similarly, some complications may also stem from the new structures. These are predominantly time related as evidenced by the table above. Difficulties, albeit surmountable ones, will arise with respect to midyear data availability, data quality, and implications of early findings on tracking savings.

Communicating early results. Evaluation structures with interim samples affords evaluators an opportunity to present evaluated results to implementation every interval; this is both an opportunity and a

risk. While this type of early feedback is beneficial, such interim results are not final and the realization rate will change through the program period. Thus, conveying preliminary information from evaluators to implementers may require some change in the communication paradigm.

Structure Selection process

Building upon the advantages and challenges of each evaluation structure, this next discussion outlines a stepwise process for selecting an evaluation structure. The first application of this approach is likely to generate lessons which in turn feed into an updated impact framework. If this framework is to fulfill its objective of fundamentally reexamining impact evaluation methodologies, the changing needs of programs and program administrators, and the role of evaluation in responding to those needs, then a process for selecting a structure must begin with high-level examination of some very fundamental questions:

1. **What do we have?** Large population of interest
2. **What do we need?** Objectives, results, confidence, precision, deadlines
3. **What else do we know?** Risks, threats, priorities, limitations, insights

These preliminary questions frame the populations of interest, the research and parameters sought, and identify influential issues that might affect the way the population is structured. This gross impact evaluation framework was charged with proposing some new approaches that improve, expedite, or otherwise reaffirm the status quo. To do so, one must begin with a fundamental reexamination of our historical research populations, take stock of what we know and what we need. When approaching the three questions above, we suggest that the population of interest is defined as large as the scope permits. In this case, let's use all Massachusetts Electric and Gas Commercial and Industrial energy-efficiency program offerings.

The next question of how to segment this large population is potentially the most challenging and should be the most time consuming. Currently in Massachusetts, some legacy dimensions dominate, and implementation stakeholders have indicated that some evaluation research no longer aligns with their efficiency offerings. This topic is discussed in greater detail in Section 4.1: *Research categories and evaluation indicators*. The segmentation of research population into discrete evaluations should include input from both evaluation and implementation stakeholders and include many considerations, including:

4. **How should we segment?**
   - Functional: groupings that relate back to program delivery, specific technical offerings or customer segments.
   - Homogeneity: similarity within a research population reduces variability and improves sampling efficiency, isolate and contain anomalous performance and dissimilar offerings.
   - Knowledge: groupings with well-understood performance may warrant preservation, known CVs or Error Ratios are useful but not vital.
   - Customers: group even heterogeneous measures at key customers, examine multi-offering participation for ways to segment evaluation that reduce customer burden.

It should be noted that Step 4 research segmentations and Step 5 structure selection, discussed below, are interdependent and could potentially be reversed. However, we suggest that segmentation must come first in order to fundamentally reexamine evaluation groups such that they serve the needs of all stakeholders better. Prioritizing segmentation is a step toward prioritizing useful feedback that benefits the delivery organization. Accordingly, a true framework ought not specify evaluation segments but rather outline guiding principles and suggestions toward dividing the program or portfolio mega-population into pieces that are efficient, effective, and statistically valid.

Specific, but not firm or inflexible, recommendations for evaluation re-dimensioning include:

- **Combine Custom.** Consider a shift from discrete evaluation efforts by Custom end-use (Custom HVAC, Custom Lighting, Custom Process, Custom Gas) and combine into a single segment of Custom projects. Evaluators can still set precision targets in sub-segments and report out results and precision by end-use.
    - This emphasizes homogeneity of savings method, deemphasizes end-use as key driver,[7] and may group customers in a way that reduces some evaluation burden and increases EM&V efficiency.[8]
    - This segment may remain well suited for Traditional evaluation, or it may be adapted into a longer duration structure (Multiyear staged or continuous Rolling) that still facilitates annual result generation for regulatory and other needs.
- **Except Custom CDA**. Comprehensive Design Approach is a worthwhile exception to the prior advice given its whole building approach to savings development.
- **Combined Heat and Power**. Similar to Custom CDA, CHP warrants continued isolation from other Custom offerings due to methodological differences, significant savings, and uniqueness as an offering.
- **Prescriptive Offerings**. Methodological consistency and fairly well-understood performance are trademarks of Prescriptive offerings. As such, Prescriptive measures are likely to warrant a separate evaluation segment from Custom. Like Custom above, sub-segmentation of this grouping into end-uses may in fact sub-optimize evaluation efficiency at the sacrifice of customer burden.
- **Upstream Offerings**. Recent evaluations of Upstream Lighting and Upstream HVAC initiatives have confirmed that the segment has some special characteristics that warrants some different research techniques, evaluation skillsets, and more importantly a longer horizon view of installation rates. Accordingly, segmentation of Upstream offerings such as Lighting, HVAC, and gas water heating as a multiple year evaluation will facilitate adjustment from follow-up visits.
- **Small Business**. Recognizing that this customer segment is not mutually exclusive to the aforementioned divisions, this group is known to possess some different characteristics. As a generalization, this segment is sizable in numbers, diverse in business type, sometimes less

[7] A sample design can employ different sampling assumptions within a large population; one can employ different error ratios for sub-segments.

[8] By evaluating multiple measures concurrently at a given customer. Evaluation costs are time dependent, and field EM&V has considerable overhead associated with preparation, scheduling, and travel. This represents a step in the direction of sampling an entire site and evaluating all measure installation at that site while there.

efficiency savvy, and often predisposed toward faster, less complex participation paths. Small Business samples tend to be larger, and site-specific EM&V more straightforward and efficient.

- Lighting so dominates this class of customers that it is tempting to give it its own evaluation segment. Non-lighting electric and gas evaluation within Small Business is easily incorporated into a longer-range evaluation strategy. One such strategy may be initiating a multiyear Small Business evaluation with year one Lighting only work (which may look indistinguishable from a Traditional evaluation). Following up with non-lighting in year two with a small supplemental sample of lighting would lay the groundwork for a three to five-year multiyear effort where years three onward would employ lesser samples to comprise a 'sliding multiyear window' set of evaluation results.

Clearly, segmentation is a complex issue, and no single dimensioning scheme makes universal sense. The electric custom/prescriptive and end-use dimensions have a very long history in Massachusetts, and they may still prove practical. The overarching theme of this segmentation discussion is to:

1) Phase out groupings that no longer relate to the ways programs are delivered,
2) Combine divisions that don't add individual value apart from breaking large segments into more manageable chunks,
3) Extend the duration of these resultant large segments to distribute effort while preserving ability to subdivide results statistically,
4) Prioritize customer burden by evaluating multiple installations at a given site concurrently, and
5) Be willing to deviate from dominant segmentation schemes to perform atypical evaluation in the form of focused studies, verification only (desk review or non-metered) evaluations, and reconnaissance exploration.

In the next section, we propose guidelines and specific examples for choosing an evaluation structure for a selected research segment. Again, it should be reiterated that the Step 4 research segmentations and Step 5 structure selection are interdependent and could be reversed or examined in unison. As illustrated by some of the discussion above, it is indeed useful to consider effective evaluation structures when selecting segmentation. However, at least at this particular evolutionary stage of Massachusetts evaluation where a paradigm change is sought but disruption is not, the sequence provided in this framework seems prudent for effective change management.

Choosing an evaluation structure

Having defined some population segments, the final step in this process is to select an evaluation structure. This process may be somewhat iterative, which is to say that a) multiple structures may be suitable and b) narrowing down to one or two structures might spawn reexamination of the segmentation in the prior step. This is reasonable, and the stepwise process is not intended to strictly prescribe single-pass decision making.

5. What structures are suitable?	Structures: Traditional, Staged, Multiyear, Rolling, Reconnaissance Timing: Quarterly, Annual, Multiyear, Continuous

Results: Segment, Sub-segment, Interim[9], Annual, Pooled[10]

As indicated here, there is a bit more to this than selecting a structure. Consideration must be given to the timing of data interchange that occurs between PAs and evaluators, the timing of evaluation results, and the level of detail of said results. To the last point, evaluation results need not be limited by the segment or timeframe. With some planning and basic statistical methods, evaluators can drill down or roll up results across/within various dimensions or time periods.

Table 3 presents a matrix of the aforementioned evaluation structures that describes some, but not all, practical applications as well as some specific population segments that are good candidates for the evaluation method. The indications in the table would benefit from further exploration and discussion. As the non-traditional evaluation structures are tested, a version of this matrix will prove useful to guide future evaluation decisions.

The candidate segments column is populated with recommendations that are considered suitable for near term evaluation. The entries are not mutually exclusive, that is, the same segment may appear in multiple rows. Particularly for structures dependent upon higher frequency participation data (Staged quarterly or Rolling), the candidate segments are less firm recommendations and more akin to suggestions for trying the non-traditional method.

[9] In this context, "Interim" refers to results that are developed for reporting partway through an evaluation or perhaps partway through a calendar year. It is listed here simply as an alternative to Annual or end-of-evaluation reporting.

[10] "Pooled" here loosely refers to analytical techniques of combining results and error/precision from smaller studies to a combined set of findings.

Table 3. Suggested applications and segments by evaluation structure

Evaluation Structure	Practical Applications	Candidate Segments
Traditional	• Rigor more important than fast feedback • Summer metering possible, but spring and early summer tends not to be unless EM&V period > 12 months. • Segments not evaluated for several years but candidates for multiyear or continuous rolling; start with a traditional evaluation.	• Custom (gas and electric) • Comprehensive Design Approach • Combined Heat and Power • Year 1 of "multiyear" segments
Staged (Quarterly)	• Segments where faster feedback is desired with some statistical validity, e.g. within a given year. • Tracking system and program personnel able to provide quarterly data at high quality. • Savings without dominant seasonal effects	• Prescriptive Lighting • Prescriptive Compressed Air
Staged (Multiyear)	• Stable segments where year to year performance has been shown, or is expected, to be stable. • Evaluated performance known at 80/10 or able to be deferred for more than a year. • Previously evaluated on a three to five-year cycle.	• Small Business (all) • Prescriptive (all) • Upstream Lighting • Custom (gas and electric)
Rolling	• Future performance is sufficiently predictable to forecast participation and savings prior to year start. • Non-seasonal measures, since rolling evaluation fundamentally evaluates once complete.	• Prescriptive Lighting • Prescriptive Compressed Air
Reconnaissance	• Segments where changes may have occurred, or validation of measure performance is otherwise sought.	• Desk reviews on Custom CDA • Specific Custom Process measures undergoing change, IMMs • Upstream Lighting installation rate follow-ups

Rolling sampling

The following elaboration of rolling sampling draws upon the successful experience of ERS (Energy and Resource Solutions) using this technique in New York.

A rolling sampling strategy selects sites for M&V throughout the program implementation period, with the last sample drawn right after the conclusion of the program year's closing. With a rolling sample design, program-level results can be reported continuously through the implementation period with the final results available relatively quickly after the conclusion of the program year. The sampling strategy consists of three components: the proxy sample, interval sampling, and final sample selection.

Proxy Sample Design. Since the installed population will not exist at the time of the sample design, a proxy sample is conducted using tracking data from a similar population[11]. Typically, the proxy sample is designed using the program's prior year of tracking data using the stratified ratio estimation design as outlined in the California framework. The design incorporates the precision targets and error ratios from prior evaluations. The two main outcomes of the proxy design are the stratum cut-points for determining which stratum a site falls within and a probability of selection for each stratum. The probability of selection is based on the proxy sample design quota, adapted for a rolling sample to account any expected large changes in program size from the proxy population and also to result in a slight under sampling.

Interval Sampling. Actual sampling occurs at the conclusion of each interval (for example each quarter). Tracking data is gathered from all of the PAs from the previous interval. Each participant is assigned a random number and binned into the correct stratum using the cut points established in the proxy sample design. Each site with a random number that falls below the random number threshold (the probability of selection for that stratum) is selected for on-site visits. The actual number of sites selected for M&V in an interval varies depending on the level of program activity in the interval and the random numbers themselves. It is possible no sites will be selected within a stratum in an interval.

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

Issues and Limitations. The method depends on receiving timely tracking data from the PAs shortly after the interval period has ended. This data need not be perfect, although substantial changes to individual site values could affect the final efficiency of the sample. This sampling method is somewhat costlier than a traditional evaluation for two reasons. First, the sample design is inherently less efficient than a traditional evaluation and may require a few additional sites over a traditional sample design which has 'perfect' knowledge of the population. Second, each interval requires an additional cycle of compiling and analyzing tracking data which increases analytical labor.

Recommendations

In summary, there is general agreement amongst stakeholders that an opportunity exists to evolve toward non-traditional evaluation structures in Massachusetts. Recent impact evaluation discussions have suggested interest in a "multiyear" approach for some program elements such as Small Business and Custom Gas. Opportunities and receptiveness also exist to incorporate "rolling" or "reconnaissance" style impact evaluation in the C&I research area. There is a strong desire to improve the timing/feedback loop, and an approach of rolling/staging impact evaluations has both interest and promise toward this objective.

For some technologies and program dimensions, the current framework evaluates on a three to five year cycle. The general feeling is that this is too long a duration because as much as five years may have passed between the installation of measures in the first year and the final published results. There are both

[11] Populations without a suitable proxy, or with high or unpredictable variation year to year, may not be the best candidates for rolling sampling. The final adjustment phase (below) is intended to resolve such differences, but a large true-up may very well erode some benefits of the approach.

opportunities and challenges involved in evaluating more often, and PAs and evaluators are motivated to explore positive changes to the current evaluation paradigm.

This Section 3.2.2 contains many suggestions on how to segment populations and select an evaluation structure. A first step could be to identify upcoming significant and stable changes to program offerings so we can build the "new" evaluation structure accordingly. Specific recommendations for refining gross impact evaluation by selectively implementing these proposed structures include the following actions:

Initiate a Traditional evaluation of Custom measures. As of this writing, a Custom Process evaluation is nearly complete, but the last Custom HVAC and Custom Lighting evaluations reflect installations that almost five years old or longer. As stated earlier, a combined Custom segment may be adapted into a longer duration structure (Multiyear staged or continuous Rolling) that still facilitates annual result generation for regulatory and other needs.

Alternatively, perform Reconnaissance on Custom HVAC and Custom Lighting to investigate whether the previous evaluation results appear to be stable.

Lay groundwork for Multiyear staging with a year one effort for Small Business Lighting and Custom Gas. Recent conversations confirm interest in examining Small Business Lighting. A Multiyear evaluation strategy could involve year one lighting-only work followed up with non-lighting in year two plus a small supplemental sample of lighting. In concert, these evaluations would lay the groundwork for a three to five-year "sliding window" multiyear evaluation.

Test quarterly data delivery in a Staged evaluation. Two prescriptive measures – lighting and compressed air – are good candidates for quarterly staged evaluation due to minimal savings seasonality. We recommend testing the feasibility of quarterly tracking system delivery and midyear sampling with one or both of these measures. This would serve as a proof of concept of the Staged evaluation structure and is likely to generate lessons for future updates to this impact framework.

Explore feasibility of continuous data delivery for Rolling evaluation. Concurrent with the quarterly data delivery test immediately above, explore the manner in which continuous data delivery may be feasible for the same population segments of prescriptive lighting and compressed air. Massachusetts evaluators should have sufficient data and experience with these offerings to forecast and develop sample quotas, but the greatest challenge may be the continuous data delivery aspect of rolling evaluation. This is a good opportunity to feasibility test.

Identify candidates for Reconnaissance work. Ideas for reconnaissance evaluation can and should come from a variety of sources. Implementation personnel are a valuable resource for insights on changing offerings, technologies, and customer trends. Reconnaissance research is a valuable and efficient technique to preemptively reveal unexpected or underperforming savings that may otherwise have gone unnoticed for a longer duration. Suggested candidates include variable speed drives, motors, and prescriptive gas.

Revisit desk reviews and develop a protocol. Once a cornerstone of impact evaluation in the Northeast, and still a dominant tool in other parts of the country, file reviews are a cost-effective way to examine measure performance. A large sample of file reviews with a nested, smaller sample of field M&V is a

reliable, tried-and-true method. Reviewing project documentation of a larger sample would allow evaluation to increase the perceived value of feedback to program delivery and provide an early indication of systemic issues. An evaluation file review protocol built into the implementation process could be the basis for unbiased tracking correction.

Remain open to the full spectrum of available M&V methods. Massachusetts gross impact evaluation most often employs IPMVP Option A and B (retrofit-isolation) approaches. Building simulation (Option D) is used in select instances, and billing analysis (Option C) is less frequently employed. All methods have their merits and places, so choosing an M&V approach might go hand in hand with the structuring and timing issues laid out above. Evaluators should be willing to take a step back and higher view of the data, and then match the tools and options to the evaluation and reporting needs.

3.2.3 Baselines and measure life

The third question entails additional issues that arise in or are related to impact evaluation:

Q3. How should impact evaluations research and handle baseline and measure life?

Stakeholders agree that impact evaluations can both inform and trigger baseline, industry standard practice (ISP), measure life, measure cost and incremental cost research. The issue thus is how and when impact evaluation teams can help with these determinations and refinements.

Baselines

The Baseline Framework is a separate document created to guide evaluators in consistently characterizing an appropriate measure baseline in impact evaluation. This Impact Framework does not contemplate baseline definitions and defers to the Baseline Framework for determinations. Thus, one literal component of the answer to Q3 is that impact evaluations handle all baseline matters as outlined in the Baseline Framework.

As of this writing, the Baseline Framework directs the evaluators to one of two distinct research paths:

1. If the combination of measure and application is unique, site-specific data must be the basis of the baseline, with assessment regarding the options available to the particular applicant. The gross baseline is the condition that would have existed absent the installed measure.
2. If there is a recognizable market for the measure at the time of installation, it is not unique, and there is no relevant code or standard, then the evaluation should rely on a population-based ISP study to define baseline.

Of course, two obvious opportunities remain for impact evaluation to contribute to improved handling of baselines. First, impact evaluation can *inform* baseline research by capturing information while on customer premises in a systematic manner that can potentially be leveraged for future use. Second, while not ideal, impact evaluation can *trigger* baseline research, including ISP and measure life research, by identifying in-field inconsistencies, abnormalities, or other market factors worthy of further inquiry. The long-term goal would be to have ISP research completed prior to conducting impact evaluation work to save on time and costs of evaluation. However, in the short-term, there will likely be cases where impact evaluation will require additional ISP research.

At the stakeholder workshop, participants considered some potential approaches for timing baseline research, such as:

Cyclical. As part of a regular evaluation cycle, standard practice research could be conducted followed by a cycle of impact evaluation. In this concept, impact evaluation and baseline research would not be conducted concurrently. A promising idea from a workshop breakout session was to alternate baseline studies with impact evaluations to give implementers a chance to incorporate new baselines before the next evaluation occurs. This approach does not improve the overall timing of evaluation as it could take a year or more for baseline recommendations to take effect. Though this was part of the workshop discussions, due limitations of this approach, the evaluation team doesn't believe this to be a realistic option.

Ad hoc. Baseline needs could be identified through the C&I management committee or subcommittees or through evaluation activity. Another idea was to have a "hotline" for implementation to be able to dispatch ad hoc research services that can be turned around quickly, perhaps as a "focused study" as proposed earlier in Section 3.2.1. The Impact Evaluation Tool Box, which is presented in in Section 5, is designed to both document where research is needed and to help prioritize it.

Exploratory. This approach could involve periodic or ongoing examination of equipment trends amongst program participants to proactively identify technologies that would benefit from standard practice research.

Sometimes timing issues cause controversial baseline determinations in C&I impact evaluation. Challenging situations arise when a customer with a unique project indicates one "standard practice" alternative measure to an implementer and a different practice to an evaluator some time later. In a recent example, this circumstance spawned subsequent conversations with evaluation and implementation in which the customer felt pulled in multiple directions in pursuit of the "right" baseline. This highlights a benefit of shifting impact evaluation timing closer to implementation: reducing the deleterious effect of time on the customer's memory and perceptions. In cases involving non-unique measures, an established ISP addresses this issue as it is more disconnected from a particular customer's set of alternate choices.

Finally, site-specific baseline assessment is also confounded by the occasionally fine line between gross and net baselines. Impact evaluation that is a) timelier, b) better coordinated with implementation, and c) targeted at net savings, could streamline the evaluation and reduce controversy and customer burden.

To the last point, there has been discussion in recent years about bringing at least some net savings research back into the C&I research area. Particularly with some of the complexities that have surfaced with regard to baseline assessment for custom measures, some stakeholders are advocating for assessment of participant free ridership and spillover concurrent with gross impact evaluation. Doing this could help ensure delineation between gross and net effects, avoiding overlap or double counting. The Baseline Framework speaks to this issue as well.

Recommendations

Net-to-Gross Research. The DNV GL Team recommends including some net savings research as part of C&I impact evaluation scopes for unique projects. This would require coordination with the MA Cross Cutting research area to define roles and develop survey instruments designed to assess gross and net savings using the same site-specific baseline.

For custom impact evaluations, the evaluation team would perform net assessments as part of the impact evaluation. The error ratio of the NTG ratio is typically higher than that of the gross realization rate, which impacts the sample size. The typical approach would be to conduct NTG surveys on the gross sample, add supplementary participant NTG surveys to achieve the precision targets, and use pooling and chaining to

combine the results. Net-to-gross surveys should reference either site-specific baselines or code/ISP baselines depending on the classification of the measure as either unique or non-unique, respectively. For this reason, it is recommended that all NTG surveys for custom impact evaluation be conducted by technical staff or engineers most familiar with the technology being evaluated.

For prescriptive and upstream studies, presumed to be non-unique or commodity type measures, the net and gross baselines would be determined by an ISP study or code. For these impact evaluations, evaluators would conduct gross savings results only. Close coordination with the MA Cross Cutting team would be required so that both gross and net savings estimates are based on the same baseline.

ISP Research. It is also recommended that the evaluation team develop an ISP repository to collect ISP studies that have been conducted and to vet those that implementation uses in their savings estimates. A first priority should be for the evaluation team to identify where ISP research is needed, prioritize those which will have the most impact under the new Baseline Framework, and conduct ISP studies under the focused study track.

The Impact Evaluation Tool Box includes several years of program savings tracking data and can be used to help prioritize ISP study needs. This tool is populated with electric and gas program savings from 2011 through 2015 by end use and technology and can be queried to identify where ISP research would be needed most. We recommend ISP studies be conducted on an ad hoc basis and not concurrent with impact evaluation.

Measure life

Measure life has a direct and important impact on measure cost-effectiveness for all projects as well as dual-baseline savings estimates for early replacement measures. The Baseline Framework states that "for retrospective use in impact evaluation the evaluator should use the RUL [remaining useful life] value of one-third of the EUL [effective useful life] unless evaluators previously have developed a program- or measure-specific RUL or the evaluation is of a unique measure that has exceptional available RUL data."[12]

Similar to baselines, impact evaluation can both inform and trigger EUL and RUL research. Stakeholders expressed concerns about using literature research for equipment life determination, with strong preference for primary data collection.

The primary means of collecting impact data in C&I impact evaluations (ex-post on-site M&V) are not a reliable or appropriate way to collect measure life, since the evaluation engineer cannot observe the age of the previous equipment, nor remaining life of the new equipment. The Baseline Framework encourages collection of life data for informational purposes only and guides evaluators to not use site-specific life in impact calculations. RUL and EUL measure life primary data collection will require a different approach.

In addition to verbal statements of equipment age, some additional sources of measure life data available to the evaluators may include:

[12] As cited in the TRM, the Massachusetts Common Assumptions default remaining useful life (RUL) is one-third of the effective useful life (EUL). This is a reasonable compromise to balancing research cost and improving lifetime savings accuracy. This basis also has been used in California. See *Summary of EUL-RUL Analysis for the April 2008 Update to DEER*, KEMA Inc., and more recently, SCE/CPUC's *Early Retirement Using Preponderance of Evidence*, v1.0, July 14, 2014. The MA TRM uses the default for most retrofit measures. Selected measures use other adjustments based on technology-specific research.

Manufacturer warranty information or insurance data. Such facility-specific information could be interesting, definitive, and comprehensive. Acquiring this information may be fraught with difficulties about sensitive company information. Nonetheless this might warrant further investigation.

Incentive applications. Applications sometimes include nameplate photographs and model/serial numbers of existing equipment. However, photographs of nameplate data are often inaccessible and/or illegible. More rigorous documentation of pre-existing equipment would be a valuable information resource if attainable.

Market characterization studies. Useful data may exist in market characterization studies for "mainstream" measures. While many such studies are implemented across the country each year, concerns sometimes arise as to whether equipment, customers, buildings, or operating conditions are sufficiently comparable to be transferrable across service areas. At minimum, researchers should explore the underlying data in the *Massachusetts C&I Market Characterization On-Site Assessments and Market Share and Sales Trends Study* (P41) for measure life insights.

The preceding involves mining, transferring, or somehow leveraging existing data sources. Yet opportunities also exist to generate measure life data via changes to implementation procedures, if we ask questions such as:

Can custom studies require and enforce the identification of key values like age of existing equipment?

Can inspectors attempt to identify the age of the existing equipment in a pre-inspection?

A more concerted effort to collect existing equipment would need to be made a priority for these questions to possibly be answered.

Measure life research needs to be shared with implementation and incorporated in a consistent manner for calculating lifetime savings. Currently, the existing life of retrofit equipment is slightly discounted to account for remaining life. Research needs to clearly identify any such discounting to prevent any gaps or overlap in savings assessment. This approach also needs to be coordinated with dual-baseline applications, which should not have discounted measure lives, but be based on a full EUL and an RUL.

Recommendations

Measure Life Research. The DNV GL team recommends that a prioritization of equipment needing measure life research be completed. With the move to dual-baseline savings, measure life is of greater importance. The Impact Evaluation Tool Box provides a framework for being able to determine the impact of EUL and RUL changes on lifetime savings. Measure life can be entered and adjusted to identify which end uses and/or technologies are affected most by the switch to a dual-baseline approach. It can also be used to assess which end uses and/or technologies are most sensitive to changes in measure life.

We recommend the evaluation team start to collect age of equipment through a variety of ways, including those described above. While none of these data sources could be used to estimate measure life on their own, a collection of data through multiple channels could benefit a larger measure life study. We also suggest the development of a systematic approach to ex-post on-site data collection of equipment age to help bolster these other data sources. This could include expanding the on-site visit to include questions about age of replaced equipment or age of similar, but still installed equipment.

3.2.4 Ex-ante M&V and early involvement

The fourth and final question investigates the potential to deviate from an overwhelmingly ex-post evaluation paradigm.

Q4. Are there opportunities to incorporate ex-ante M&V and other kinds of early involvement into impact evaluation?

This question drives at multiple aspects of ex-ante evaluation involvement, including its benefits, feasibility, and practicality. Pre-treatment conditions are a critical component of many early replacement or retrofit measure impact assessments, and evaluators often struggle with developing savings based upon high-rigor data collection ex-post but less rigorous data ex-ante. That imbalance necessitates assumptions and reliance on verbal reporting and memory, which erode the confidence and accuracy of savings estimation.

Additional benefits of ex-ante evaluation involvement may include:

- Reducing downstream uncertainty on site specific realization rates via early agreement between implementer and evaluator on baseline characterization
- Inspecting pre-retrofit conditions and characteristics that might not be accessible or recalled post-installation
- Educating the implementer about evaluation methods and savings considerations
- For projects with implementer M&V, the evaluator may have an opportunity to review the M&V plan and make suggestions to gather desired data
- Obtaining timely insights to customer motivations and decision making for net-to-gross assessment

While potential benefits are numerous, ex-ante involvement would require improved collaboration between implementation and evaluation while addressing some adversarial strains in their historical relationship. One of these strains has been the perception that ex-ante M&V will delay measure installation and achievement of savings goals.

Stakeholders indicate that these tensions can be overcome, and there is likely ample mutual benefit to doing so. Workshop comments generally indicated that a fundamental lack of understanding sometimes exists between the two parties.

There are more opportunities for pre/post metering as well as other types of ex-ante involvement not related to metering. There also are prospects for ex-ante activities that benefit both evaluation and the customer, such as the installation of permanent cloud-based metering. Other types of early involvement might include baseline consultation, peer review of measure feasibility and calculation accuracy, and commissioning support. Some technologies or customers might be better candidates for early involvement than others given their size, complexity, knowledge level, knowledge interest, and/or relationship with program personnel.

Barriers

Bias. There were concerns about whether early involvement can influence evaluation results, particularly if it only touches a portion of the participant population. One solution to this would be to assign ex-ante customers to their own stratum in the ex-post evaluation sample design. If all participants had early intervention, then statistical isolation likely would not be needed.

Acceptance. There could be some resistance from implementation to include evaluation early enough in process so as not to disrupt the customer or impact the implementation of projects. It is possible that

implementation and evaluation may not always agree and that evaluation recommendations, including baseline selection, aren't accepted by implementation if they view them as being unrealistic or unfair to the customer. A process could be established where both sides must agree to the ex-ante review prior to any ex-ante involvement by evaluation on a case-by-case basis. The result of the ex-ante review is binding to the evaluation barring any extra-ordinary circumstances, as noted in the Baseline Framework.

Logistical challenges. In 2010, evaluators initiated a pre/post study of prescriptive variable speed drives that was very challenging logistically. While this particular project occurred several years ago, it provided insights to issues that can arise early in the incentive application process such as incompleteness, preliminary or flawed assumptions, or equipment targeted for installation that do not ultimately occur. Inserting an evaluation visit in the process flow of a customer application or technical assistance study adds communications complexities and delays. Such issues will need examination if evaluators are to intervene successfully with pre-metering. Pre-metering is currently being done in other jurisdictions, primarily for custom projects. Custom projects are better suited for early M&V involvement due to the need for custom engineering estimates by the implementation vendor and engineering review by the PA. There is an opportunity to integrate pre-metering into the custom workflow before or during PA engineering review.

Facility access. The customer can be a barrier to ex-ante involvement if evaluation on-site visits are included in the scope. Most facility visits are disruptive to their operations, and security concerns are on the rise.[13] Customers have a vested interest in audits that are essential to receiving incentives, but evaluation is not perceived to be part of that process. Evaluators know this phenomenon full well: after rebates are paid, there is considerably less willingness to cooperate. At the stakeholder workshop, some discussed ways to get customer buy-in for early involvement, including finding some added value for the customer. Financial incentives are not uncommon to facilitate cooperation with evaluation or research; however, most municipalities simply cannot accept incentives, and some commercial businesses have policies against it as well. For projects that may be candidates for ex-ante involvement, we would suggest engaging evaluation as early as possible to make them part of the process. This may generate more collaboration and be seen as less of an adversarial role.

Alternatives

Non-evaluator ex-ante M&V. Logistics and facility access are barriers to additional customer visits by non-implementer personnel, so there may be an opportunity to train vendors that provide technical assistance or measure installation in pre-metering. There may also be an opportunity to require or incentivize additional metering as standard practice for certain measures, such as controls, and collect data in a way that improves evaluation reliably. It would be important for data to be collected in a manner that is compliant with ISO New England M-MVDR requirements, although some data, even if imperfect, is better than no data.

Paths forward

When stakeholders considered ways to facilitate ex-ante M&V and early involvement, the conversation overwhelmingly centered on making the relationship between implementation and evaluation more

[13] Advanced billing data analysis, or M&V 2.0, which is yet unproven, may hold some promise for addressing these concerns. Similarly, remote data collection may provide some solutions as more equipment is being controlled or accessed via the internet. However, the requirement remains that data must be collected in a manner that is compliant with ISO New England M-MVDR requirements.

collaborative. Some speculate that the roles have become unnecessarily adversarial through evolution as a result of some past experiences or perceptions that have been imprinted into staff on both sides of the divide. Evaluation leaders expressed interest in serving as a helpful resource to implementation in pursuit of common objectives of effective and successful energy efficiency programs.

Massachusetts examples do exist wherein implementers have assisted with the collection of data for evaluation, but such examples are few. Some ways to facilitate mutual understanding and benefit include:

Use consistent terminology. One implementer remarked that evaluators talk about "programs" that do not actually exist. It is true that some evaluations are structured according to legacy groupings of program tracks, end uses, or delivery mechanisms that may not be relatable to implementers and current offerings. To find common ground, a good start would be to use the same language.

Request feedback. Conversely, it has been suggested that some implementation sales staff have limited use of evaluation findings or "lessons learned" due to the timing of studies. Evaluators should welcome frank feedback on the sort of information, and timing of it, that implementers would find most useful and actionable.

Shift the tone. Evaluation personnel would do well to reposition evaluator input as a constructive feedback situation rather than a critical assessment. Since ex-ante involvement requires improved collaboration, evaluators can alter their tone so that instead of pointing out instances of failure, they are identifying ways in which implementers can mitigate risks in order to succeed.

Identify liaisons. Both implementers and evaluations may benefit from carefully selected intermediaries to liaise and champion more constructive communications and outcomes. Both parties want efficiency offerings to be successful and customers to be satisfied with minimal disruption.

Recommendations

Ex-ante Baseline Review. The DNV GL team recommends coordinating with implementation to determine if they see value to ex-ante EM&V baseline review. The review would focus on the selection of early replacement or end-of-life and the efficiency of the baseline. Since the purpose of ex-ante EM&V baseline review would be to benefit implementation, they should drive the decision to engage in this strategy. Should implementation see the value in an ex-ante EM&V baseline review arrangement, the recommended process would involve the following steps:

1. Implementation calls evaluation in on certain projects to seek evaluation's interpretation of baseline. These would likely be the larger, custom projects.

2. Following the initial outreach and based on the information provided, evaluation chooses to provide a review at this time or defer the review to a future impact evaluation when there may be more information available.

3. If evaluation chooses to engage in the ex-ante baseline review, the agreed upon baseline becomes binding to the evaluator and will not change if the same project turns up in a future evaluation. This should be clearly documented in the PA's project file and documented by the evaluation team. If evaluation defers the review to a future impact evaluation, the selected baseline is eligible for revision at that time. The results of the ex-ante baseline review are not binding to the implementer.

4. During future impact evaluations, projects that were reviewed by evaluation in the ex-ante period would be given a case weight of 1, or included in a separate stratum in a sample design to represent themselves.

It is important to note the voluntary nature of this arrangement for both implementation and evaluation is what allows the ex-ante baseline review to be binding. It is also important that a mechanism be put in place to support timely engagement and response from the evaluation team. At any point, either party can opt-out of the engagement, which would mean that the ex-ante baseline selection is eligible for future evaluation review.

Ex-ante Measurement and Data Collection. The DNV GL team recommends coordinating with implementation to develop a strategy to systematically collect pre-measurements for certain types of projects, notably controls projects. The DNV GL team has evaluators with experience performing pre-installation metering and data collection in MA and other jurisdictions. They understand the challenges laid out above and how to handle these. Any pre-installation measurement approach will have to be coordinated with implementation in order for it to be successful. Evaluation seeks a collaborative strategy that will address implementation's concerns of customer disruption and installation delays, while providing evaluation with data critical to estimating savings for controls type projects.

3.3 Summary of Recommendations

Throughout the research of this project, two overarching themes on how to refine impact evaluations emerged: 1) increase the usefulness of evaluation results, and 2) increase the timeliness of evaluation results. These two elements are strongly linked, since timelier results will increase the opportunity for results to be actionable. However, it was also noted that regardless of when results are received, opportunities remain for these results to be more relevant and useful to program delivery and subsequent improvements.

Enhance coordination

Stakeholders acknowledged that the relationship between implementation and evaluation has been a bit adversarial. They also suggest that each stakeholder group is often lacking a fundamental understanding of the interests of the other. For example, implementers are focused on how their offerings will be changing and on future structuring and just meeting their installation goals, while evaluators are largely retrospective and sometimes focused on offerings and delivery paths that have already changed. With some reservation, it is clear that the stakeholders anticipate ample mutual benefit from enhanced coordination.

Actions items suggested by stakeholders that could lead to enhanced coordination include:

Improved communication. Terminology is sometimes inconsistent between evaluation and implementation, and implementers view the current evaluation structure as opaque and burdensome. A strong desire was expressed for an evaluation framework that is flexible, transparent, interactive, practical, and actionable. It might also be beneficial for evaluation staff to explicitly request regular feedback and input from implementation staff in a structured fashion.

Closer engagement. Benefits would result from earlier and more detailed engagement between evaluation and implementation staff. Potential mechanisms include: earlier file reviews for a larger sample of projects; review of project material closer to project completion; and engagement in pre/post metering, peer review,

or commissioning activities. Implementation staff might consider a review process to ensure that evaluation findings are considered in a timely fashion, and that return feedback is provided to evaluation staff.

3.3.1 Deploy non-traditional evaluation techniques

Research for this Impact Framework discovered widespread support for expanding options for evaluation approaches and timing. One of the drivers for this is the rapid evolution of market conditions that affect program performance, including changes in technologies like LEDs. These changes are happening at different rates across various groupings of customers, offerings, and technologies and do not conveniently align with regular evaluation cycles.

Popular mechanisms to increase the usefulness and timeliness of evaluation activities included the use of a rolling sample, ongoing or continuous evaluation, and decoupling impact evaluation from the annual update of customer profile data. Stakeholders are interested in aligning the evaluation cycle with the program cycle and developing a mechanism for focused ad hoc studies to meet specific needs and to feed into the formal impact evaluation cycle.

Sections 3.2.1 and 3.2.2 contain much information on what this report terms “non-traditional evaluation structures” along with a stepwise process for rethinking the paradigm of population segmentation and guidance on the selection of suitable evaluation techniques. Multiple techniques are bound to be suitable for a given portfolio element, and thus the final evaluation decision should involve stakeholder input and careful weighing of pros and cons, regulatory needs, and beneficial program feedback.

Transitioning evaluation paradigms may be difficult but there are many good reasons to pursue it. As stated earlier, the current paradigm may be generalized as evaluating on a three-five year cycle, and the general feeling is that this is too long a duration because as much as five years may have passed between the installation of measures in the first year and the final published results. There are both opportunities and challenges involved in evaluating more often, and PAs and evaluators are motivated to explore positive changes to the current evaluation paradigm.

Framework authors recommend the following actions as part of the next steps in refining gross impact evaluation and selectively implementing some new evaluation structures (more details in Section 3.2.2).

- Initiate a Traditional evaluation of Custom measures soon, and consider evolving it into a longer duration structure such as Multiyear staged or continuous Rolling.
- Alternatively, perform Reconnaissance on Custom HVAC and/or Custom Lighting to investigate whether the previous evaluation results appear to be stable.
- Lay the groundwork for Multiyear staging with a year one effort for Small Business Lighting, followed by non-lighting plus a small supplemental sample of lighting in year two, and subsequent lower precision samples to build a three to five-year “sliding window” multiyear evaluation.
- Test quarterly data delivery in a Staged evaluation using prescriptive lighting and or compressed air measures as a proof of concept of the Staged evaluation structure.
- Explore the feasibility of continuous data delivery for Rolling evaluation current with the aforementioned quarterly data delivery test.
- Identify candidates for Reconnaissance work, such as variable speed drives or prescriptive gas measures.
- Revisit desk reviews as an evaluation technique and consider developing an evaluation file review protocol to define the manner in which file review findings may or may not be the basis for unbiased tracking correction.

Integrate baseline and measure life research

Stakeholders agree that impact evaluations can both inform and trigger baseline, industry standard practice (ISP), measure life, measure cost and incremental cost research. This research identified the need for specific actions to address baseline and measure life questions. Action items stemming from the discussion of both baseline and measure life included coordination with the Baseline Framework, interviews with PA implementation management on these issues, secondary research on inputs used in other jurisdictions, and investigation of alternative data sources

Net-to-Gross Research. With some of the complexities that have recently surfaced with regard to baseline assessment for custom measures, the framework recommends integrating some participant free ridership and spillover analysis with gross impact evaluation. We recommend this be done for all custom impact evaluations. Doing this could help ensure delineation between gross and net effects. This would require increased coordination with the MA Cross Cutting research area to define roles and develop survey instruments designed to assess gross and net savings using the same site-specific baseline.

ISP Research. ISP research should be completed as part of an ad hoc or "focused study" under this impact evaluation framework. The impact evaluation tool box, along with input from both implementers and evaluations, should be queried to identify and prioritize ISP research first and foremost.

Measure Life Research. The DNV GL team recommends that a prioritization of equipment needing measure life research be completed. In addition to traditional measure life research techniques, measure life data may also be found in alternative information sources, such as company depreciation schedules, insurance data, disposal or recycling contractors, market characterization studies, ex-post on-site data collection and incentive applications. The last item recommends establishing a method to generate measure life data during existing implementation process but with some additions to the data that are collected on pre-existing equipment.

This, as with many findings in this Impact Framework, ties to the potential benefits of closer coordination between evaluation and implementation. Much common ground was found while preparing this Impact Framework, and there is mutual benefit to leveraging many of the opportunities identified herein.

4 SYSTEMATIC IMPACT PLANNING PROCESS

Since 2010, C&I impact evaluation planning has been performed annually as part of a group planning session with key stakeholders. The timing of these planning sessions has been somewhat inconsistent, as they are set up to serve other research areas in addition to the impact evaluation research area.

One of the objectives of this framework was to document and systematize the C&I impact evaluation planning process. While Massachusetts impact evaluation decisions have been based on sound logic, the framework behind the decisions made has never been formally documented. The PAs and EEAC Consultants recognized the need for a documented decision-making framework and "roadmap" to maximize the value of impact evaluation studies now that the statewide C&I programs and evaluation framework have matured.

Historically, evaluators have considered key indicators such as proportion of savings by program or measure category, proportion of program spending, uncertainty, the duration since the last evaluation, and other influential factors to prioritize elements of the C&I efficiency portfolio for impact evaluations along with the targeted level of rigor for each study. Impact evaluation has traditionally been conducted at the end-use level for all electric and prescriptive gas measures and at the sector level for custom gas measures. The dimension of reporting has varied as well, with results generally developed at the statewide level for all prescriptive measures and at the PA level (for the larger PAs) for most custom end-uses and programs.

4.1 Research categories and evaluation indicators

This framework document provides an examination of the research categories and the key indicators that trigger evaluation.

4.1.1 Structural challenges

In the development of impact evaluation studies, there are some structural challenges that need to be considered. These challenges can be related to timing constraints, accuracy requirements and the desire to produce meaningful and actionable findings and recommendations.

Currently, there is a tension between the ways program delivery and evaluation is structured. Implementation is generally structured for timely and effective delivery of program offerings, while evaluation is structured for highly accurate quantitative savings results.

Evaluation is challenged to find a balance between a) the desire to report in dimensions that are meaningful to implementers, and b) structuring research into homogeneous groups with known or estimable error ratios for efficient sample designs and cost-effective research. Historically, impact evaluation has been more (b), while the stakeholder workshop identifies the need for more (a).

4.1.2 Application of results

Recently, the issue of how to apply results of impact evaluation studies has come up in both C&I electric and C&I gas evaluations. The framework document identified the need to develop universal guidelines that address the issue of how to apply results. As a general principle, decisions of how results will apply should be made and documented in the planning stage when designing each evaluation study. All evaluation work should be planned with the end result in mind so that the evaluation team knows how the results will be used. There is also a prevailing need for evaluation results that are applicable to both retrospective and prospective program activity.

The principles and protocols that have been applied in recent years are as follows:

Custom program offerings (e.g., projects that are in the custom track and include engineering studies that may differ from PA to PA): Realization rate results are calculated at the PA level for large PAs and also at the statewide level for use by smaller PAs. In order for each PA to use its own results, it must meet an accuracy target defined in the work plan of the study. The accuracy target is usually based on the importance of the end-use being studied. If a PA result is not better or equal to this target, the statewide result may be used.

Prescriptive/upstream program offerings (e.g., projects with common savings algorithms and assumptions across PAs): Here, a statewide result is calculated for use by all PAs. PA-level results are not calculated for program offerings that are common across all PAs.

As evaluation results come in, it will be important to review and validate those results and their applicability for their intended use. If modifications must be made to how they are applied, the reasons for this must be documented in the final reports.

4.1.3 Research categories

Impact evaluation research categories, since EEAC's inception, have been grouped by end use with prescriptive vs. custom subsets, e.g., custom HVAC. There have been exceptions to this model for measures or delivery paths that were anticipated to perform differently. Custom gas has also deviated from the end use level analysis. The evaluated performance of efficiency offerings relative to tracking savings is an important consideration, since optimal statistical sampling prefers evidence-based error ratios and homogeneous groups. The results of these studies are typically applied retrospectively and/or prospectively to the impact category researched. This provides PAs with accurate results at reasonably granular levels within their energy efficiency portfolios.

As program offerings grow in complexity or in the number of incentive paths, evaluators need to account for potential interactions between different research efforts, e.g., in baseline standard practice research, or between different delivery mechanisms, e.g., downstream vs. upstream.

There are also evaluation activities, like the framework development, that fall outside or span multiple programs. The proposed track of "focused studies" described earlier in Section 3.2.1 provides an avenue for this type of research.

4.1.4 Key indicators

There are many indicators and drivers of impact evaluation, but simpler is better for a framework and decision making.

Researchers suggest synthesizing (not limiting) the indicators into three primary categories:

- Relevance - magnitude of savings, historical and expected; germane to future program activity
- Uncertainty - of savings, measure performance, prior evaluation results; variability and precision
- Priority - regulatory or other requirements; vintage of prior evaluation results; market changes or change in program delivery

Each category should be developed using a combination of both analytical and intuitive key indicators. Analytical indictors include historical tracking data, trends in savings over time, and results and age of prior

studies. Intuitive indicators include knowledge of upcoming programmatic changes, regulatory requirement, political sensitivities, and intuitive uncertainty in an area of a program or existing evaluation.

4.1.5 Backburner allocation

The methodology described in 4.2 presents a process for ranking and scoring impact evaluation activities for measures and programs, which is the fundamental mission of the evaluation. However, this process is unlikely to score frameworks or smaller end-uses for evaluation. This allocation could be used as an available margin for queueing up the "focused studies," which are the more creative, overarching, or neglected activities that also need to be included in the mix.

4.2 Decision making and scoring

DNV GL developed a methodology for decision making and scoring that uses a combination of research category and key indicators to prioritize evaluation research. The method, which is provided as part of the Impact Evaluation Tool Box in 5.2, separates the research categories and key indicators into two dimensions.

- Dimension 1 are the key indicators: uncertainty, relevance and priority.
- Dimension 2 are the "program" groups, including fuel, incentive path, end-use and technology type impacted. These groups can be aggregated to the segments being evaluated as described in Section 3.2.2 above or split out to the more granular technology type level.

Each of the program groups in Dimension 2 can be given a score of 1-5 (very low, low, moderate, high, very high) with a zero representing "not applicable" for each of the dimension 1 indicators. Each of the indicators can also be given a relative weight of 1-10 (very low priority to very high priority). The tool is populated with historical tracking data, study results, and any intuitive inputs for each program group in Dimension 2.

The output of this tool is a prioritized list of each program group, which could be an input for impact evaluation planning discussions. The tool is flexible enough to be able to create new groupings that may be of interest to key stakeholders and doesn't have to stick with the traditional dimensioning done in the past. The power of the tool is its ability to be updated on a continuous basis as findings from studies and input from PAs are added to the tool. Additionally, the tool is not dependent on an annual planning process in order to be used for decision making. Rather, the tool can be "pinged" on a regular (e.g., quarterly) basis, if desired.

One other important use of the tool is its ability to be adapted to help identify other research needs such as ISP or measure life. It can be used as a sensitivity analysis tool, which provides input on uncertainty. For example, if evaluators wanted to prioritize technologies that would benefit most from new or updated ISP research, a field can be added to test the impact that baseline changes could have on certain technologies. Similarly, with a shift to dual-baseline savings estimates for early replacement projects, the tool can be used to run sensitivity checks on the savings impacts of changes to EUL. This can also be used to prioritize measure life research as this becomes increasingly important.

4.3 Portfolio-level impact evaluation planning refinements

Another objective of this Impact Framework was to propose refinements to the portfolio-level impact evaluation planning process. A strong theme that emerged from the stakeholder workshop involved the mutual benefit of evaluators improving communications and engagement of implementation personnel. This framework document seeks to define a long-term planning process that can address the desire both to refine

the planning approach and to meet the needs of the varying impact evaluation structures defined earlier in this document.

This planning paradigm establishes a broad framework for impact evaluation planning that will form the basis for regular planning. The elements of the planning process include the following:

- The potential range of activities that should be encompassed in each planning cycle
- Establishment of a uniform process and terminology to assess evaluation needs
- How to determine the appropriate evaluation method and level of rigor for each study

4.3.1 Long term planning horizon

As a first step, a long-term evaluation map needs to identify when major evaluation outcomes are required (when do PAs need results from the next upstream lighting or CDA evaluation, for example). Ideally, this planning horizon will stretch forward to span the longest interval for regularly planned activities (like market characterization), which may be on the order of ten years. This initial mapping should become more specific with each three year cycle, and refined frequently using the tools described in Section 4.2.

4.3.2 Ongoing impact planning

A shift in evaluation structure will require advance and continuous planning. Adding to this is the need to identify and execute on non-traditional impact research like ISP and measure life. The impact planning process should include the steps outlined in Table 4.

Table 4. Impact planning process

Step	Description	When	Who
1	Determine appropriate program segmentation for evaluation purposes (see Section 3.2.2 for recommendations)	Start of planning cycle	PA-E, EEAC consultants, evaluation contractors
2	Determine the appropriate evaluation structure for each segmentation	Start of planning cycle	PA evaluators, EEAC consultants, evaluation contractors
3	Establish long term evaluation plans for each segmentation, including the sampling approach, evaluation methods and how results will be applied	Start of study	Evaluation contractors
4	Coordination with implementation and regular review of program participation data to identify other research needs for “focused study”	Continuous	PA-E, PA-I, EEAC consultants, evaluation contractors
5	Regular reporting on findings from impact evaluations and focused studies	Continuous	Evaluation contractors
6	Continuous update of Impact Evaluation Tool Box	As findings come in	PA-E, Evaluation contractors

While we still see that impact evaluations may be driven by seasonality and reporting deadlines to some extent, there are other types of research (notably, NTG research) that are applied only prospectively. These types of studies, as well as ISP and measure life research, can be done on a very different timeline, in order to even out the workflow.

5 IMPACT EVALUATION TOOL BOX

The Impact Evaluation Tool Box (Tool Box) is designed to supplement the impact evaluation planning process by providing a repository of information including a list of impact evaluation methodologies, the spreadsheet scoring tool, the impact evaluation calendar, and the documented history of tracking savings and impact evaluations conducted in MA. The Tool Box is intended to be a living set of resources that can be updated and added to over time. This section describes each of these resources, which are all contained in separate spreadsheets that accompany this document.

5.1 Repository of impact evaluation methodologies

DNV GL created an initial list of eight impact evaluation methodologies as described in Table 5 below. This list is not exhaustive and includes those methodologies currently in use as well as potential alternatives. Non-traditional and innovative evaluation techniques should be considered at this stage. This list draws from the experience of the DNV GL team, PAs, and EEAC Consultants both inside and beyond Massachusetts.

Table 5. Description of impact evaluation methodologies

Methodology type	Description
Desk review	Project file review to verify savings assumptions
Verification - phone only	Verification of installed EEMs and operating characteristics through phone survey
Billing analysis[14]	Normalized comparison of pre- and post-installation billed usage for both participants and non-participants. Note, billing analysis is a tool often applied as one of the on-site M&V techniques and programmatically for some prescriptive natural gas impact evaluations.
Verification - on-site/no metering	On-site verification of installed quantities and types through visual inspection only
Verification - On-site with post metering - prescriptive	On-site verification of installed quantities, types and operation through visual inspection and post installation metering
Verification - On-site with metering - custom	On-site verification of installed quantities, types and operation through visual inspection and post installation metering
Whole building simulation	On-site verification of installed quantities, types and operation through visual inspection and metering with calibrated whole building model analysis
Pre/Post data collection	On-site verification of existing and installed quantities, types and operation through visual inspection and metering for both the pre- and post-installation conditions (Alternate method: disabling controls as proxy for pre-measurement)
Focused Study - ISP, NTG, Measure Life	Literature review, interviews, on-site data collection
Ex-ante Baseline Review	Project baseline review, interviews with TA vendor/PA engineer

Figure 3 identifies the research questions that these strategies could address. This is not a complete list of research questions, but rather it includes those that could help evaluation stakeholder decide what type of

[14] A variant of billing analysis, sometimes referred to as M&V 2.0, is an approach using billing data (particularly high resolution interval data) to support implementation and potentially provide evaluation grade impact results. The method is unproven as an electric savings evaluation method in the C&I sector, but is actively being investigated and tested. See http://www.neep.org/broadcast/press-releases/neep-and-partners-win-state-energy-grant-us-department-energy.

methodology to choose when planning studies. For reconnaissance style evaluation, desk review would help answer many questions that may ultimately lead to a full M&V style evaluation. This approach was tested, and used, for recent custom gas impact evaluations.

Figure 3. Summary of research questions addressed by each methodology

Verification - On-site with Metering - Custom	•What are the normalized annual energy and peak demand savings? •What are reasons for differences from ex ante? •Other field observations possible (like lighting quality).
Whole Building Simulation	•What are the normalized annual energy and peak demand savings? •What are reasons for differences from ex ante? •Other field observations possible (like lighting quality).
Pre/Post Data Collection	•What are the normalized annual energy and peak demand savings? •What are the pre-installation operating conditions? •What are the reasons for differences from ex ante? •Other field observations possible (like lighting quality).
Focused Study - ISP, NTG, Measure Life	•What is ISP for commodity type measures? •What are technology specific measure lives? •What are the particpant free-ridership and spillover rates for unique projects?
Ex-ante Baseline Review	•Does evaluation contractor agree with the proposed baseline? •What information would be needed to determine an appropriate baseline?

The DNV GL team considered the applicability, intuitive accuracy, benefits, and drawbacks of each of the methods presented above. A summary of these factors is presented in Table 6. Each of the methodologies, except for desk review and ex-ante baseline review, can produce analytical results that can be applied both retrospectively and/or prospectively. Desk review results are not intended to produce realization rates or savings factors since no actual monitoring and verification is being conducted using this methodology. The verification – phone only and verification – on-site/no metering methods are able to produce analytical results that can be applied, but these are limited to single parameters such as installation rate. The other methods provide sufficient rigor to be able to produce annual savings and peak demand estimates for retrospective or prospective application.

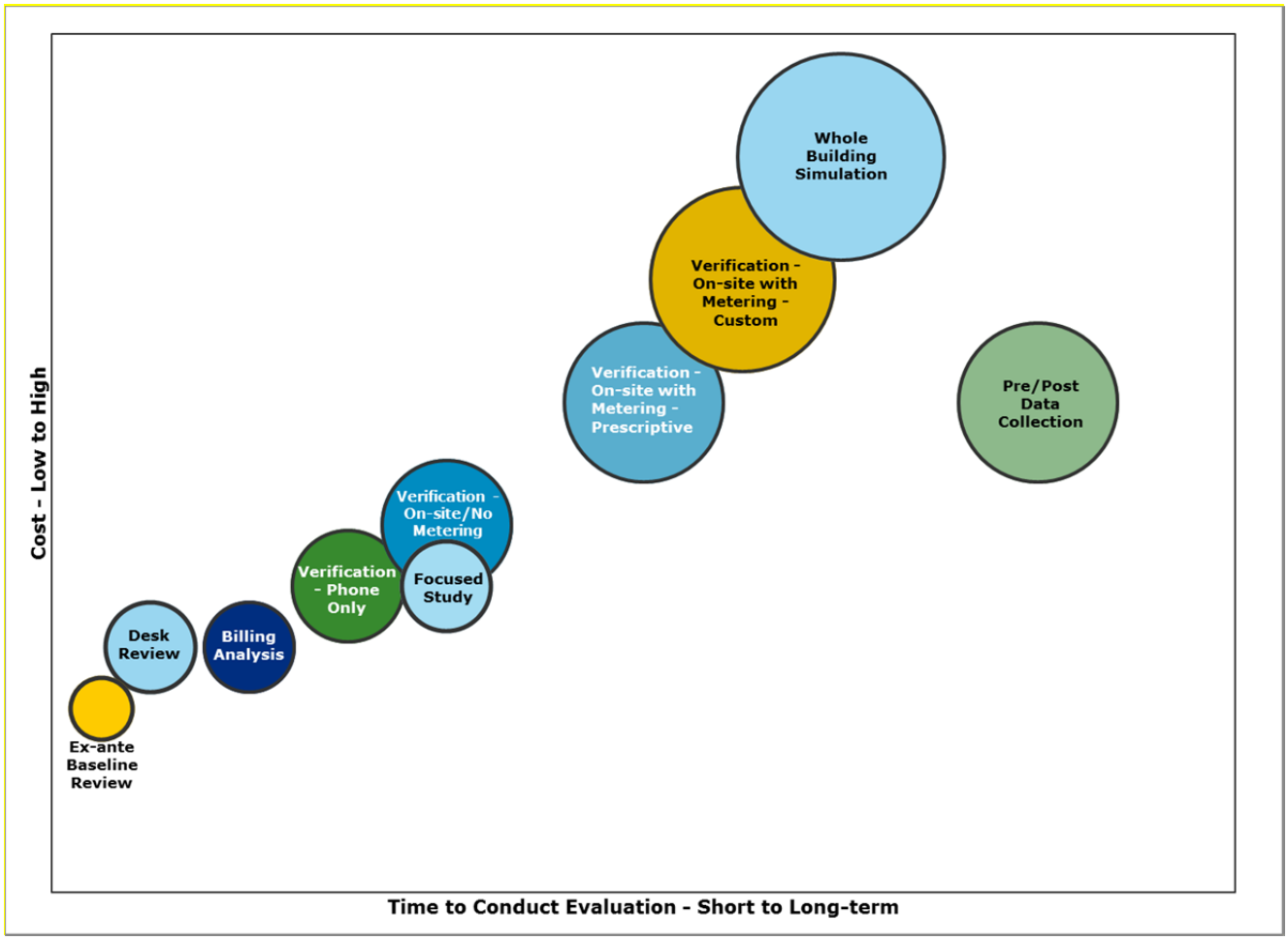
Table 6. Applicability, intuitive accuracy, risks and rewards of each method

Methodology type	Application of results	Intuitive accuracy of results	Rewards	Risks
Desk review	Qualitative	Does not produce quantitative results	Quick, inexpensive, indicator of quality of documentation, reasonableness of savings estimates and baseline.	Does not produce accurate analytic results.
Verification - phone only	Retrospective or prospective	Low	Reasonably accurate installation rates from knowledgeable respondents.	High uncertainty of operating conditions. Risk of respondent not familiar with measures installed.
Billing analysis	Retrospective or prospective	Medium (with large enough population)	Can include large or full populations. Utilizes both pre- and post-installation usage. More opportunity in gas program evaluation.	Not all EE measures can be accurately verified using this method. Savings of less than 10% of billed usage are not easily found. Complications in matching meters. Not applicable to normal replacement measures
Verification - on-site / no Metering	Retrospective or prospective	Medium	Highly accurate installation rates. Reasons for discrepancies.	High uncertainty of operating conditions.
Verification - on-site with metering - prescriptive	Retrospective or prospective	High	Rigorous evaluation done in accordance with many evaluation standards including IPMVP Options A or B. Highly accurate installation rates. Reasons for discrepancies.	More time intensive and higher cost.
Verification - on-site with metering - custom	Retrospective or prospective	High	Rigorous evaluation done in accordance with many evaluation standards including IPMVP Options A or B. Highly accurate installation rates. Reasons for discrepancies.	More time intensive and higher cost.
Whole building simulation	Retrospective or prospective	High	Rigorous evaluation done in accordance with many evaluation standards including IPMVP Option D. Highly accurate installation rates. Reasons for discrepancies.	Highest cost and very time intensive. Requires accurate calibration with actual billed usage and metering, which can be resource intensive.
Pre / Post data collection	Retrospective or prospective	High	Rigorous evaluation done in accordance with many evaluation standards including IPMVP Options A or B. Captures pre-installation usage patterns. Highly accurate installation rates. Reasons for	Very complex to implement this type of evaluation. Dependent on measuring usage prior to implementation or switching measure "off." Requires highly coordinated communication between

Methodology type	Application of results	Intuitive accuracy of results	Rewards	Risks
			discrepancies.	implementation, customer and evaluation.
Focused Study - ISP, NTG, Measure Life	Retrospective or prospective	Low to High (dependent on selected level of rigor and type of data collected)	Can tailor the focused studies to meet the needs of the researchable question. Baseline Framework establishes guidance for low to high rigor ISP research options. Not tied to a larger impact study and can be initiated as needs are identified.	Need to establish how these results get worked into results of impact studies (retrospectively and/or prospectively). Most ISP and ML research would still rely on some secondary data, which adds some uncertainty.
Ex-ante Baseline Review	Retrospective or prospective	Does not produce quantitative results	If evaluation and implementation both agree to a baseline, this result would be binding and would not be overruled in evaluation. Evaluation would have the benefit of getting the same information as implementation at the same time.	If evaluation and implementation do not agree to a baseline, this parameter is at risk for adjustment during evaluation. Potential conflicting views on baseline could add time and cost to evaluation.

While each of the methods has its risks and rewards as describe above, timing and cost should be considered when selecting an appropriate evaluation methodology. Figure 4 presents a graphical depiction of the relative time and cost resources needed to conduct each type of impact evaluation. The size of the bubble in the chart below represents the relative cost of each study. Of course, these timelines and costs do not apply to every study, and exceptions certainly exist. However, in general terms, a whole building simulation evaluation would likely be the costliest option, while a true Pre/post data collection methodology would generally take the most time to complete. This chart, like many resources created for this Impact Framework, can be refined over time based upon Massachusetts evidence and experience.

Figure 4. Impact evaluation methods by time and cost



5.2 Spreadsheet tool of scoring method

The second component of the Tool Box is the spreadsheet-based scoring and prioritization tool. This tool will be one part of any current and future planning sessions and is design to facilitate the development of a prioritized list of evaluation studies. This scoring tool is not the only consideration when it comes to planning, but serves as a consolidated resource that contains historical tracking data and study results along with a matrix of key indicators and program offerings. As part of its development, DNV GL consulted several evaluation stakeholders to gain feedback and input. This collaborative effort culminated in a robust, but flexible, tool that is designed to provide value to all evaluation stakeholders in current and subsequent planning cycles.

Use of the tool follows the process diagram laid out in Figure 5 below. In subsequent planning sessions, Step 1 will include updating the key inputs based on recent findings from research conducted since the previous use of the tool.

Figure 5. Process diagram of spreadsheet scoring tool

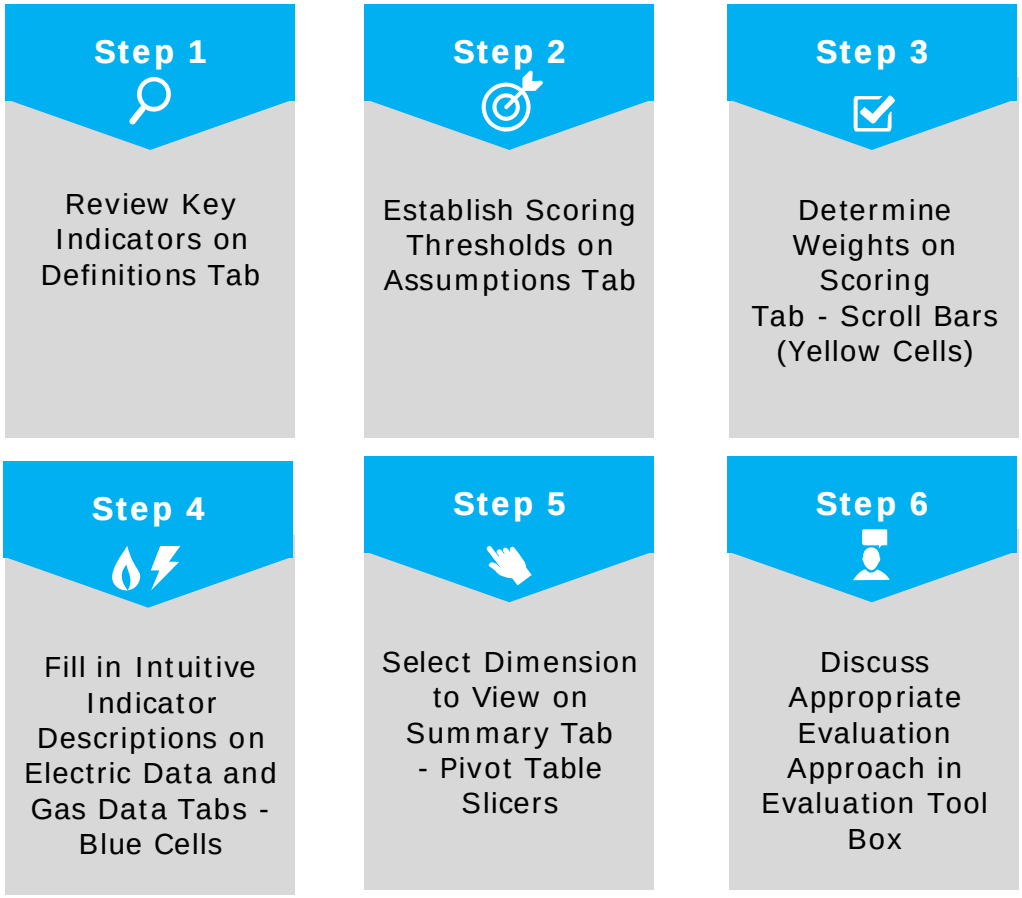


Table 7 presents the key indicators that are reviewed in Step 1 of the tool. These include the high level key indicators (relevance, uncertainty, and priority) identified earlier in this document as well as the sub-indicators and their definitions. Each of the sub-indicators is classified as either analytical or intuitive. Analytical indicators are those that are based on prior data, including, but not limited to, savings size, study results, and study age. Intuitive indicators are intended to include key information about program offerings, uncertainty of current studies, requests from stakeholders, and any requirements or political sensitivities that are less quantitative but also should be considered.

Table 7. Key indicators and sub-indicators for impact evaluation from definitions tab

	Analytical / Intuitive	Classification of metric	Definition
Relevance	Magnitude of savings (kWh and/or Therms)	Analytical	2015 annual energy savings - percent of total C&I portfolio savings
	Historical trend	Analytical	Percent change in energy savings across years with available data
	Magnitude of savings (kW-S)	Analytical	2015 Summer kW savings - percent of total C&I portfolio savings
	Magnitude of savings (kW-W)	Analytical	Winter kW savings - percent of total C&I portfolio savings
	Implementation changes	Intuitive	Any fundamental change in implementation of energy efficiency measure that renders prior results invalid (e.g. new program, baseline, technology, etc.)
	Other	Intuitive	(e.g. have changes taken hold)
Uncertainty	Realization Rate	Analytical	Energy realization rate of most recent impact evaluation
	Precision	Analytical	Relative precision of realization rate at 90% confidence for energy savings
	RR Trend	Analytical	Realization rate trend for projects with more than one completed study
	Intuitive Uncertainty	Intuitive	General uncertainty in results of prior research (e.g. HVAC controls with post-only data collection)
	Market/technology Shifts	Intuitive	Expected or recent market/technology changes that would inspire new research (e.g. baseline)
	Other	Intuitive	
Priorities	Age of Study	Analytical	Number of years that the results have been in use (for retrospective results years = study year - 1, for prospective results years = study year + 1)
	Regulatory Requirement / Political Sensitivity	Intuitive	Any regulatory or political needs for undergoing a study (e.g. FCM requirement that a study be no more than 5 years old)
	Implementation/EEAC Requests	Intuitive	Requests for research by implementation or EEAC consultants
	Other	Intuitive	(e.g. baseline, measure life studies)

Step 2 in the process establishes the scoring thresholds for each of the analytical indicators. Each analytical indicator will be scored on a scale of 1-5 (very low to very high) with a zero representing “not applicable.” The scores for each of the cut points is presented in Table 8. For example, if the savings of the measure is less than 3% per year, it is assigned a low priority.

Table 8. Scoring priorities and cut points from assumptions tab

		Priority Cut Points for Analytical Measures					
		Relevance		Uncertainty			Priority
Scoring	Priority Level	Magnitude of Savings (kWh, kW-S, kW-W, Therms)	Historical Savings Trend	Realization Rate	Precision	RR Trend	Age of Study (years)
0		No Savings Current Year	No Trend	No Prior Study	No Prior Study	1 or No Prior Study	N/A
1	Very Low	1%	-50%		5%	5%	1
2	Low	3%	-20%		10%	10%	3
3	Moderate	5%	20%		15%	20%	4
4	High	10%	50%	between .7 and 1.2	20%	30%	>4
5	Very High	>10%	>50%	<.7 or >1.2	>20%	>30%	No Prior Study

Step 3 involves documenting any intuitive indicators. Users enter descriptions of intuitive indicators into the electric and gas data tabs. An example of an intuitive "relevant" indicator could be an energy efficiency incentive offering moving from downstream to upstream. The change in delivery mechanism is a fundamental change to the program that may warrant new research. Intuitive indicators are likely often given a score of 5 since not every program offering will require intuitive indicators and if something is significant enough to warrant inclusion, it should be scored higher.

Once all indicators have been established, Step 4 sets the relative weight of each key indicator on a scale of 1-10 (low priority to high priority). Figure 6 presents an extract of the electric scoring tab and shows the weights along the top row in yellow. Weights can be selected by moving the sliders up and down as appropriate. Since the tool uses relative weights, the sum of the weights does not have to equal 100, but users are encouraged to think of weighting the key indicators in terms of which should be low, moderate, and high priority. The wider scale of 1-10 is intended to provide more room for differentiation between the indicators.

Figure 6. Electric large C&I scoring tab

			WEIGHTS (0-10)	8	9	8	9	7	9		
				Relevance - SCORE		Uncertainty - SCORE		Priorities - SCORE			
Incentive Path	End Use TRM	End Use Impacted	Technology Type	Magnitude of Savings (kWh)	Program Changes	Precision	Intuitive Uncertainty	Age of Study (years)	Regulatory/ Political Sensitivity	TOTAL SCORE	TOTAL SCORE RANKING
Custom	CDA	CDA	CDA	3	0	3	0	4	0	1.6	7
Custom	CHP	CHP	CHP	5	0	1	0	4	0	1.8	3
Custom	Compressed Air	Compressed Air	Air Compressor	1	5	5	0	4	0	1.8	4
Custom	Compressed Air	Compressed Air	Controls	0	0	5	0	4	0	1.0	45
Custom	Compressed Air	Compressed Air	Drains	0	0	5	0	4	0	1.0	45
Custom	Compressed Air	Compressed Air	Dryer	0	0	5	0	4	0	1.0	45
Custom	Compressed Air	Compressed Air	Filter	0	0	5	0	4	0	1.0	45
Custom	Compressed Air	Compressed Air	Motors	0	0	5	0	4	0	1.0	45
Custom	Compressed Air	Compressed Air	Other	1	0	5	0	4	0	1.4	11
Custom	Food Services	Food Services	Controls	0	0	0	0	5	0	0.4	126
Custom	Food Services	Food Services	Equipment	1	0	0	0	5	0	0.6	103
Custom	Food Services	Food Services	Ice	0	0	0	0	5	0	0.4	126
Custom	Food Services	Food Services	Other	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Aerators and Spray Valves	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Boilers	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Hot Water	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Motors	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Other	0	0	0	0	5	0	0.4	126
Custom	Hot Water	Hot Water	Water Heaters	0	0	0	0	5	0	0.4	126
Custom	HVAC	Building Shell	Building Shell	1	0	2	0	2	0	1.2	23
Custom	HVAC	Building Shell	Insulation	1	0	2	0	2	0	1.3	17
Custom	HVAC	Building Shell	Other	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Boilers	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Chiller	2	0	2	0	2	0	1.5	9
Custom	HVAC	HVAC	Controls	1	0	2	0	2	0	1.3	16
Custom	HVAC	HVAC	DEEC	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Fuel Switch	1	0	2	0	2	0	1.5	10
Custom	HVAC	HVAC	Furnace	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Heat Pump	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	HVAC	1	0	2	0	2	0	1.3	17
Custom	HVAC	HVAC	Motors	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Other	1	0	2	0	2	0	1.3	14
Custom	HVAC	HVAC	Steam Traps	0	0	2	0	2	0	1.0	37
Custom	HVAC	HVAC	Unitary	1	0	2	0	2	0	1.3	14
Custom	Lighting	Lighting	Lighting	4	0	2	0	4	0	1.7	5
Custom	Lighting	Lighting	Other	0	0	2	0	4	0	0.8	75
Custom	Lighting	Lighting	Performance Lighting	0	0	2	0	4	0	0.8	75
Custom	Lighting	Lighting Controls	Lighting Controls	1	0	2	0	4	0	1.1	30

Once the relative weights are set, the tool develops an overall score and ranking for each entry (row). Step 5 in the process allows users to view the results of the scoring method in a summary tab. It is important to note that users can also create custom groupings across rows and view the results of any cut of the data preferred using the pivot table and slicers included in the tool. This provides the starting point for a stakeholder driven process to determine where evaluation resources should be allocated.

Once the priority list of studies has been discussed and agreed upon, Step 6 is to choose the appropriate evaluation methodology from the Impact Evaluation Tool Box presented above. In order to select the appropriate methodology, one may consider where a particular program offering lands on the priority scale. Something with a very high score might require a more rigorous evaluation methodology, while those with a lower score could be candidates for a "focused study."

In the end, the systematic planning process is not going to be spreadsheet driven but will be stakeholder driven to identify the long term and short term evaluation goals. This is and should be a collaborative process that is not formulaic. However, the toolbox, the sensitivity analysis and the structure behind it helps stakeholders put together a well-considered, but flexible allocation plan. The tool is simple and is based on a principle of allocating resources over the long term. The allocations can be adjusted as circumstances require.

5.3 Impact evaluation calendar

The proposed impact evaluation planning process was presented earlier in Section 4.3.1. This planning process integrates lays out the steps that the evaluation group would take beginning with the selection of the appropriate segmentation. Coordinating impact evaluation in step with program implementation is a preferred practice which was revealed in conversations with stakeholders at the workshop and also surfaced as part of the literature review. It will be important to periodically review the planning process as new priorities and objectives arise over time.

5.4 Documented evaluation history

The Tool Box provides a repository of historical evaluations conducted in Massachusetts since 2006. A complete list of studies compiled is included in Table 10. This list is made up of mostly impact evaluations and includes 24 studies that have been completed as part of the MA C&I research area since 2010.

5.5 Summary of Impact Evaluation Tool Box

The purpose of the Impact Evaluation Tool Box is to supplement the impact evaluation planning process by providing a repository of information that can be used as a resource when planning impact evaluation studies. The Tool Box includes a list of impact evaluation methodologies, the spreadsheet scoring tool, the impact evaluation calendar, and the documented history of tracking savings and impact evaluations conducted in Massachusetts. While the Tool Box itself does not provide all of the answers, it is a key component in the planning process. Initially, the Tool Box can be used as a starting point for prioritizing studies. Our intention is that findings and structures from new impact evaluations will be integrated into the Tool Box as the process evolves.

APPENDIX A. RESEARCH FINDINGS: STAKEHOLDER ENGAGEMENT

On May 24, 2016, the DNV GL team facilitated a stakeholder workshop on key issues in C&I program impact evaluation at the Columbia Gas office in Westborough, Massachusetts. The worked focused on optimizing the structure, timing, and staging of evaluation and related research. Participants included representatives of PAs, EEAC, and the DNV GL team (see Table 9 below).

Table 9. Workshop participants

Organization	Name
EEAC EM&V Consultants	Ralph Prahl and Jen Chiodo
CLC	Gail Azulay, Meredith Miller, and Margaret Song
Columbia Gas	Monica Cohen
Eversource	Erik Mellen, Tracy Dyke-Redmond
National Grid	Bill Blake, Tony Larson, and Ezra McCarthy
Unitil	Mary Downes
Apex	Scott Dimetrosky
DMI	Alec Stevens
ERS	Sue Haselhorst, Jon Maxwell, and Betsy Ricker
Itron	Marc Collins and Mike Rufo
Jacobson Energy Research	Dave Jacobson
SBW	Bing Tso
DNV GL	Dan Barbieri, Chad Telarico, Andrew Wood, and Wendy Todd

The workshop agenda revolved around these questions:

Q1. In what ways can impact evaluation research be structured to reduce the length of time parts of the portfolio go unevaluated?

Q2. Is there an opportunity to integrate staged, "rolling," or reconnaissance style evaluation into the impact evaluation framework?

Q3. How should impact evaluations research and handle baseline and measure life?

Q4. Are there opportunities to incorporate ex ante M&V and other kinds of early involvement into impact evaluation?

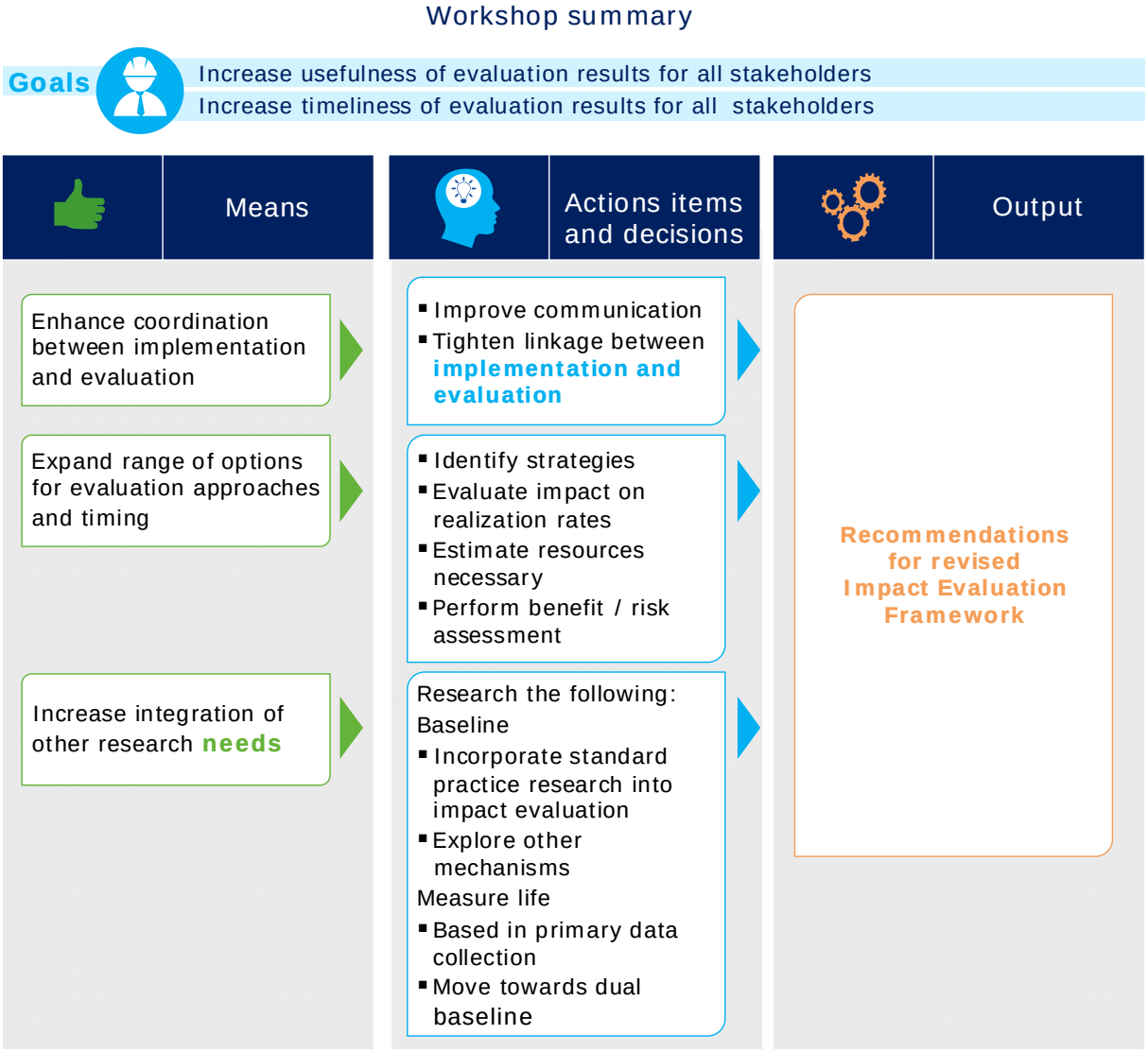
Stakeholders also raised other important topics not elsewhere classified. Workshop activities included discussion stations for each of the topics above, reports by the discussion station to the full participant group, and a prioritization exercise.

On June 15, 2016, the DNV GL team submitted "Refinements to Gross Impact Evaluation Framework (P63) Workshop Summary Memo" to the PAs, EEAC, and other stakeholders which summarized findings and conclusions from the workshop. These have been incorporated into Section 3.2 of this report.

Summary

Figure 7 provides a graphical depiction of the stakeholder workshop.

Figure 7. Stakeholder workshop summary



APPENDIX B. RESEARCH FINDINGS: LITERATURE REVIEW

The DNV GL team gathered information from a variety of sources including:

- Historical impact evaluations completed by the PAs in all research areas
- MA C&I portfolio of tracking data (2011-2015)
- Interviews with individual PAs or subsets of PAs and EEAC Consultants
- Practices for conducting impact evaluations in other jurisdictions, via document review and interviews with knowledgeable individuals

Historical impact evaluations

Impact evaluation in Massachusetts has a long history, and it is useful to reflect upon it and document reasons why the previous specific research was chosen to be undertaken in the past. The DNV GL team reviewed and cataloged information on the prior impact evaluation studies completed by the PAs since 2006. This spreadsheet based catalog includes the source location of the study, year the study was complete, author/s, name of the study, sponsor and sector studied. An extract of the catalog is shown below in Table 10.

Table 10. Historical impact evaluations in Massachusetts

Year	Title of evaluation	Sponsor	Sector
2007	Massachusetts Saving Electricity: A Summary of the Performance of Electric Efficiency Programs Funded by Ratepayers Between 2003 and 2005	MA	Res/C&I
2007	Multiple Small Business Services Programs - Impact Evaluation 2007 - Executive Summary	MA	C&I
2007	Impact Evaluation Analysis of the 2005 Custom SBS Program	MA	C&I
2007	Small Business Services Custom Measure Impact Evaluation	MA	C&I
2007	Impact Evaluation of 2006 Custom Lighting Installations	National Grid	C&I
2007	2005 EI-D2-SBS Lighting Controls Impact Evaluation	National Grid	C&I
2008	NSTAR Electric and Gas, Business and Construction Solutions (BS/CS) Programs Measurement and Verification Final Report	NSTAR	C&I
2008	Business & Construction Solutions (BS/CS) Programs Measurement & Verification - 2006 Final Report.	NSTAR	C&I
2008	Impact Evaluation of 2006 Custom HVAC Installations	National Grid	C&I
2009	2007 Business and Construction Solutions (BS/CS) Programs - Measurement and Verification of 2007 Lighting Measures	NSTAR	C&I
2009	2007 Design 2000plus Lighting Hours of Use & Load Shapes Measurement Study	National Grid	C&I
2009	National Grid USA 2008 Custom Lighting Impact Evaluation	National Grid	C&I
2010	2007/2008 Large C&I Programs, Phase 1 Report Memo for Lighting and Process Measures.	WMECO	C&I
2010	Western Massachusetts Small Business Energy Advantage Impact Evaluation Report Program Year 2008	WMECO	C&I
2011	Prescriptive Condensing Boiler Impact Evaluation	MA	C&I
2011	Impact Evaluation of 2009 Custom HVAC Installations	MA	C&I
2011	Impact Evaluation of 2008 and 2009 Custom CDA Applications	MA	C&I

Year	Title of evaluation	Sponsor	Sector
2011	Impact Evaluation of 2009 Custom Gas Installations	MA	Res/C&I
2011	2007/2008 Large C&I Programs Final Report.	WMECO	C&I
2012	Impact Evaluation of 2010 Custom Process and Compressed Air Installations	MA	C&I
2012	Impact Evaluation of the 2010 Custom Lighting Installations.	MA	C&I
2012	Heat Pump Water Heaters Evaluation of Field Installed Performance.	MA	Res/C&I
2012	Non-Controls Lighting Evaluation for the Massachusetts Small Business Direct Install Program: Multi-Season Study	MA	C&I
2012	Final Report, Small Business Direct Install Program: Pre/Post Occupancy Sensor Study	MA	C&I
2012	Impact Evaluation of the 2011-2012 ECM Circulator Pump Pilot	MA	C&I
2012	Prescriptive Gas – Draft Final Program Evaluation Report	MA	C&I
2012	Impact Evaluation of 2010 Custom Gas Installations	MA	C&I
2013	Impact Evaluation of 2011 Custom Gas Installations	MA	C&I
2013	Impact Evaluation of 2011 Custom Refrigeration, Motor and Other Impact Evaluation Final Report	MA	C&I
2013	Impact Evaluation of 2010 Prescriptive Lighting Installations Final Report	MA	C&I
2013	Impact Evaluation of 2011 Prescription Gas Measures	MA	C&I
2013	Combined Heat & Power Program Impact Evaluation	MA	C&I
2013	Impact Evaluation of 2011-2012 Prescriptive VSDs	MA	C&I
2013	Massachusetts Combined Heat and Power Program Impact Evaluation 2011-2012	MA	C&I
2013	Massachusetts Small Business Direct Install 2010-2012 Impact Evaluations	MA	C&I
2013	Onsite Lighting Inventory – Results Final Report	MA	Res/C&I
2014	Retrofit Lighting Controls Measures Summary of Findings	MA	C&I
2014	Upstream Lighting Impact Evaluation	MA	C&I
2014	Impact Evaluation Massachusetts Prescriptive Gas Pre-Rinse Spray Valve.	MA	C&I
2015	Massachusetts Commercial and Industrial Upstream Lighting Program: "In Storage" Lamps Follow-Up Study	MA	C&I
2015	Massachusetts 2013 Prescriptive Gas Impact Evaluation - – Steam Trap Evaluation Phase I, FINAL	MA	C&I
2015	Massachusetts Commercial New Construction Energy Code Compliance Follow-up Study	MA	C&I
2015	Impact Evaluation of 2012 Custom HVAC Installations	MA	C&I
2015	Impact Evaluation of Prescriptive Chiller and Compressed Air Installations	MA	C&I
2015	Comprehensive Review of Behavior and Education Programs	MA	Res/C&I
2015	Comprehensive Review of Non-Residential Training and Education Programs, with a Focus on Building Operator Certification	MA	C&I
2015	2013 Massachusetts Prescriptive Gas Thermostat Evaluation Study & Programmable Thermostat Decision Memo.	MA	Res/C&I
2015	Project 43 Impact Evaluation of PY2013 Custom Gas Installations.	MA	C&I

MA C&I tracking data (2011-2015)

As part of the development of the systematic scoring tool, DNV GL compiled all of the C&I electric and gas portfolio savings between 2011 and 2015. The complete set of data currently resides in the systematic scoring tool spreadsheet that accompanies this framework document. A high-level summary of annual statewide gas and electric tracking savings is presented in Table 11.

Table 11. Summary of annual electric and gas savings.

	Electric tracking savings			Gas tracking savings
Year	Annual kWh	Summer kW	Winter kW	Annual therms
2011	542,809,471	79,497	64,325	9,783,407
2012	611,338,966	42,912	84,935	13,926,522
2013	665,497,677	182,786	159,792	12,167,447
2014	751,484,880	60,664	60,028	15,300,725
2015	916,889,575	80,302	98,408	12,125,568

Practice from other jurisdictions

The issues raised during the scoping for this project and at the stakeholder workshop were not unique to the Massachusetts service territory. For this reason, the DNV GL team performed extensive, but not exhaustive,[15] secondary research into evaluation frameworks in other jurisdictions. The research included jurisdictions in the U.S. and Canada and was limited to sources published within the last decade. Almost 40 documents were acquired and given a cursory review. Of this pool, 26 were determined to be relevant to the research and were reviewed further with reference to the four core research questions. DNV GL classified the documents into three categories: explicit frameworks, papers or articles, and general evaluation resources.

With regard to the key research issues of increasing the usefulness and timeliness of evaluation results, the basic findings of this research are:

1. These issues have been under discussion for more than a decade.
2. Published frameworks have yet to incorporate mechanisms that would achieve the benefits sought.

The earliest succinct statement found during this research in a published framework was from the 2010 Avista Utilities filing with the Washington Utilities and Transportation Commission. It set a scheme for prioritizing EM&V resources that included the following quotations:

> "Timing: Information that would have value in improving an ongoing program would have higher preference" [16]
>
> "Timeliness is an important consideration for planning evaluations. EM&V should be undertaken in a manner that is designed to provide important information in a timely fashion for regulatory reporting, program planning and/or improvement, and other needs"[17]

These themes recur in other frameworks, almost exactly in the same words in some cases.[18] Yet to date no jurisdiction reviewed has developed implementation mechanisms to these ends.

The rest of this Appendix identifies the documents reviewed by category and includes some selected quotes from each category.

[15] The difference between "extensive" and "exhaustive" is the degree to which the researchers can express confidence that they have found all relevant sources. This research effort was limited to electronically accessible published material.

[16] Avista Utilities, "Evaluation, Measurement, and Verification (EM&V) Framework" (2010), p.18

[17] Ibid, p.19

[18] See Puget Sound Energy Evaluation Framework, revised August 2015

Table 12. Impact evaluation frameworks and guidance documents

Year	Company / jurisdiction	Document title
2006	NREL - National	A Framework for Evaluating the Total Value Proposition of Clean Energy Technologies
2010	Avista / WA	Avista Utilities: Evaluation, Measurement, and Verification (EM&V) Framework
2011	PacifiCorp/WA	Evaluation, Measurement & Verification Framework for Washington
2013	Texas	Texas rule Chapter 25, Subchapter H. Division 2. (n)(2)(q) Evaluation, measurement and verification (EM&V)
2015	Delaware	An EM&V Framework for Delaware (Presentation)
2015	U.S. EPA	Evaluation Measurement and Verification (EM&V) Guidance for Demand-Side Energy Efficiency (EE)
2015	Pennsylvania	Evaluation Framework for Pennsylvania Act 129 Phase Ii Energy Efficiency and Conservation Programs
2015	Kentucky	Energy Efficiency EM&V Basics and Issues (Presentation)
2015	New York	Case 14-M-0094, Proceeding on Motion of the Commission to Consider a Clean Energy Fund Clean Energy Fund Information Supplement
2015	Ontario	Evaluation, Measurement and Verification (EM&V) Protocols and Requirements: EM&V Protocols v.2.0
2015	Puget Sound Energy	Exhibit 8: Evaluation, Measurement & Verification (EM&V) Framework

Selected Quotes:

"Evaluation will be a key tool to continually test, measure, and adjust the approaches used to engage the market under the CEF. NYSERDA will use the Test-Measure-Adjust platform to identify the effectiveness of pilots and decide whether and how to scale them up, and to continually assess the effectiveness of initiatives beyond the pilot state. The evaluation approach will combine quick-cycle feedback activities along with long term tracking and accountability. A quick feedback cycle will provide actionable recommendations to refine NYSERDA's strategy and rebalance its portfolios."[19]

"Field verification is most commonly an accountability mechanism to ensure accurate and credible energy and emission impacts, but it will be implemented under the CEF in as 'real-time' a manner as possible.

[19] New York State Energy Research and Development Authority, "Clean Energy Fund Information Supplement" (2015), p.25

In doing so, field verification will be designed to identify ways to improve current project level impacts, and to obtain key insights to improve impact projections for future initiatives."[20]

"Before describing the evaluation planning process, it is important to understand how it is integrated with the program planning-implementation-evaluation cycle. This is necessary to align budgets, schedules, and resources. It is also a way to ensure that data collection supports planned evaluation efforts and is embedded with program delivery." [21]

Table 13. Relevant papers and articles

Year	Company / Jurisdiction	Document title
2007	Ontario	Starting Over – Developing an Evaluation Framework and Protocols in Ontario
2009	Ontario	A Comprehensive Framework for Evaluating Demand Response in a Resource Planning Context
2010	New Mexico	Review of 2009 New Mexico Energy Efficiency Program Evaluations and Recommendations for Future Evaluation Infrastructure
2012	USA	Developing State and National Evaluation Infrastructures- Guidance for the Challenges and Opportunities of EM&V
2012	International	International Review of Frameworks for Impact Evaluation of Appliance Standards, Labeling, and Incentives
2012	USA	An Impact Evaluation Framework for Public-Private Collaborations on Research, Manufacturing, Supply Chain and Early Markets
2013	Massachusetts	To Do or Not to Do: Is It Time for Another Impact Evaluation?
2015	USA	How Information and Communications Technologies Will Change the Evaluation, Measurement, and Verification of Energy Efficiency Programs

Selected quotes

"The evaluation process should be integral to what is typically a cyclic planning implementation-evaluation process. Therefore, evaluation planning should be part of the program planning process so that the evaluation effort can support program implementation, including the alignment of implementation and evaluation budgets and schedules, and can provide evaluation results in a timely manner to support existing and future programs"[22]

"This paper presents a novel approach for developing objective criteria to aid in deciding whether to proceed with an expensive full-scale evaluation. The criteria consist of different measurements of the quality of

[20] Ibid., p. 152

[21] Ontario Power Authority, "Evaluation, Measurement, and Verification (EM&V) Protocols and Requirements" (2015), p.27

[22] Steven R. Schiller and Charles A. Goldman, "Developing State and National Evaluation Infrastructures - Guidance for the Challenges and Opportunities of EM&V" (2012), p.9.

the applicant savings estimates and subsequent program administrator (PA) engineering reviews, comparing past program activities (the benchmarks) to the present program on an application-by-application basis. The inference is that if the present program is measurably different from the benchmark, it is prudent to proceed with the full-scale impact evaluation." [23]

"The energy efficiency sector has long sought the ability to measure energy savings as they happen. While this has not been fully realized, we are getting closer. ICT is simplifying the harvesting of savings data, improving the quality of analysis, and increasing the timeliness of reporting."[24]

"Real time is the key term for program EM&V. Rather than evaluating projects by comparing 'before' and 'after' snapshots, programs will eventually be able to track energy savings at the same time as consumption, as they happen."[25] (2015, USA)

Table 14. Additional resources

Year	Company / jurisdiction	Document title
2010	USA	Energy Efficiency Evaluation, Measurement and Verification Resources (Compendium by Steve Schiller)
2010	USA	"Review of Evaluation, Measurement and Verification Approaches Used to Estimate the Load Impacts and Effectiveness of Energy Efficiency Programs"
2010	Ontario	Draft Evaluation Plan Template 2011 - 2014
2015	Ontario	Protocols for Evaluating Behavioral Programs
2016	USA	Supplemental Comments of Environmental Defense Fund on EPA's Draft Evaluation Measurement and Verification (EM&V) Guidance for Demand-Side Energy Efficiency
2016	USA / LBNL	"Planning and Budgeting for the Evaluation of EE Programs" webinar
2016	Virginia	Memo to Virginia re-establishment of EM&V framework

Selected quote:

"(Evaluate) In a timely manner and provide feedback for:

- Ongoing program improvement
- Supporting portfolio assessments
- Support the planning of future portfolio cycles, load forecasts, and energy resource plans."[26]

[23] Susan Haselhorst, Erik Mellen, and Chad Telarico, "To Do or Not to Do: Is It Time for Another Impact Evaluation" (2013), p.1

[24] Ethan A. Rogers, Edward Carley, Sagar Deo, and Frederick Grossberg, "How Information and Communications Technologies Will Change the Evaluation, Measurement and Verification of Energy Efficiency Programs" (2015), p. vii

[25] Ibid., p.26

[26] Steven Schiller, "Energy Efficiency EM&V Planning Basics and Frameworks." Slide 10. Webinar, Lawrence Berkeley National Laboratory, "Planning and Budgeting for the Evaluation of Energy Efficiency Programs," May 23, 2016

About DNV GL

Driven by our purpose of safeguarding life, property and the environment, DNV GL enables organizations to advance the safety and sustainability of their business. We provide classification and technical assurance along with software and independent expert advisory services to the maritime, oil and gas, and energy industries. We also provide certification services to customers across a wide range of industries. Operating in more than 100 countries, our 16,000 professionals are dedicated to helping our customers make the world safer, smarter, and greener.