Evaluation, Measurement and Verification Protocol V4.0

February 2021

Connecting Today.
Powering Tomorrow.

Table of Contents

1. Abbreviations

$m:	Million dollars
AC:	Air conditioning
CDM:	Conservation and demand management
DEER:	Database for energy efficiency resources
DSM:	Demand-side management
EM&V:	Evaluation, measurement and verification
EUL:	Effective useful life
GWh:	Gigawatt hour
HEMS:	Home energy management systems
HVAC:	Heating, ventilation and air conditioning
kW:	Kilowatt
kWh:	Kilowatt hour
LC:	Levelized delivery cost metric
LDC:	Local distribution company
LED:	Light emitting diode
M&V:	Measurement and verification
MAL:	Measures and assumptions list
MW:	Megawatt
NEBs:	Non-energy benefits
NTG:	Net-to-gross
PAC:	Program administrator cost test
PC:	Participant cost test
RCT:	Randomized controlled trial
RDD:	Regression discontinuity design
RED:	Randomized encouragement design
RFP:	Request for proposal

RFS:	Request for services
RIM:	Ratepayer impact measure
SC:	Societal cost test
SEER:	Seasonal energy efficiency ratio
TRC:	Total resource cost
Yr:	Year

2. Introduction

This Evaluation, Measurement and Verification (EM&V) Protocol ("Protocol") describes standard industry best practices to evaluate conservation and demand management (CDM) programs. CDM program evaluations are conducted to assess program performance, support good management practices, help facilitate adjustments to achieve stated goals, ensure that the CDM programs achieve their intended goals, provide value for ratepayers, and identify areas of strength and opportunities for improvement. Evaluations help enhance programs by:

- Estimating the extent to which desired outcomes are being achieved.
- Identifying the necessary improvements to maximize the effectiveness of the stated goals.
- Providing actionable updates on program activities to indicate they are being carried out as planned.
- Determining if customer needs and/or expectations are addressed.

The EM&V Protocol V4.0 is an updated version of the EM&V Protocols and Requirements V3.0 which was published in 2014. The protocols were updated in order to:

- Address the evolution of energy efficiency programming;
- Explore new and emerging EM&V concepts;
- Ensure protocols provide sufficient guidance for evaluating the latest energy efficiency technologies and programs;
- Make the document more user friendly and easy to understand for non-experts and first-time users; and
- Make the document AODA[1] compliant.

The major changes incorporated in the updated EM&V Protocol V4.0 include the following:

- The structure now follows the evaluation process and combines Volume I: EM&V Protocols and Requirements and Volume II: Protocols for Evaluating Behavioral Programs from the previous protocol.
- An emphasis on providing guidance and improving comprehension for a broader audience.
- New and updated examples to better explain evaluation concepts.
- New and emerging evaluation topics such as midstream program evaluation and M&V 2.0 are discussed.

[1] Government of Ontario (2016). *Accessibility for Ontarians with Disabilities Act*, 2005, S.O. 2005, c. 11. Website: https://www.ontario.ca/laws/statute/05a11

2.1. Intended Audience

The Protocol is primarily intended to provide detailed guidance on how to evaluate CDM programs. The Protocol leverages existing best practices and addresses emerging needs using novel approaches; and is of most relevance and use to:

- **Evaluation administrators and managers.** The organization or individual(s) responsible for defining the scope for the program evaluation. The evaluation administrators and managers are also the point-of-contact for EM&V contract management.
- **Evaluators.** The individual(s) or firm(s) selected to develop and implement the evaluation plan based on the scope defined by the evaluation administrator. The evaluation contractor could also be referred to as the "independent, third-party evaluator" or the "evaluator"
- **Program administrators.** The individual(s) or organization(s) responsible for the design, development, and implementation of an energy efficiency, conservation, or demand response initiative. A program administrator may also be referred to as a "program manager" or a "program implementer."

2.2. Evaluation and CDM Program Categorization

There are different impact evaluation methodologies for different type of CDM programs. In this protocol the impact evaluation methodologies are described as it aligns with whether the program targets technology change or behavioural change:

- **Technology-based programs:** These programs achieve energy or demand savings by replacing existing technology with more efficient technology, or by improving the operational efficiency of technology.
- **Behavioural-based programs:** These programs achieve energy or demand savings by utilizing strategies designed to influence energy and demand consumption behaviours by consumers, operators, installers, lenders and other market actors. Behavioural-based programs consist of a diverse set of programs, which incorporate various elements, including outreach, education, competition, rewards, benchmarking and feedback.

CDM program evaluation is aligned with this categorization as summarized in Section 2.3.

Programs can also be categorized based on the points in a product delivery chain to which incentives are directed:

- **Downstream programs.** These programs offer incentives directly to customers. One of the main reasons to deliver these programs is to raise consumer awareness, which most often leads to positive spillover to other energy efficiency purchases. These programs are able to target selected populations.
- **Midstream programs.** These programs offer incentives to distributors and retailers and typically encourage retailers to stock or sell a larger percentage of efficient products. The main objectives of these programs are to influence customers at their point of decision and to help address the lack of availability of efficient products.

- **Upstream programs.** These programs offer incentives to manufacturers. The intent of these programs are usually to encourage manufacturers to streamline their production lines and increase production thus lowering the price. The main advantage of these programs is that they usually have lower transaction costs because it can influence a large portion of the market through fewer actors.

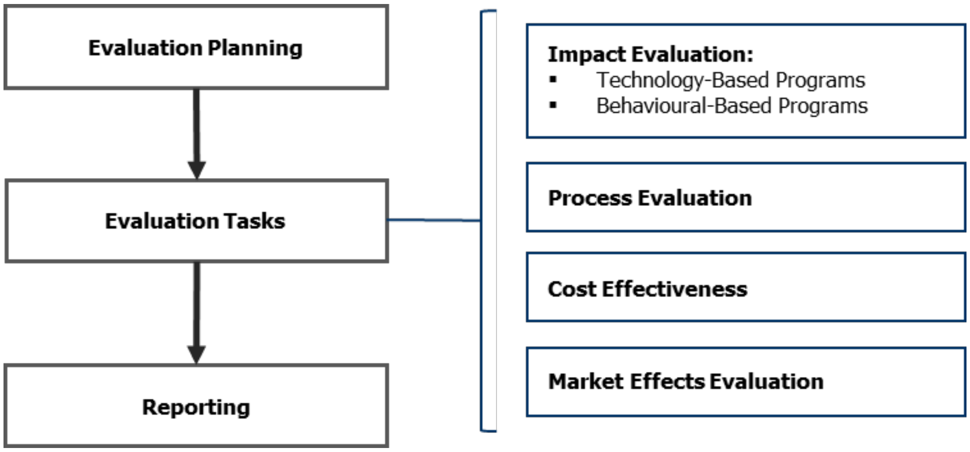
Downstream programs are the most common and widely delivered type of CDM program. The impact evaluation protocol (Sections 4.1 and 4.2) provides guidance to evaluate downstream programs. Evaluation elements that differ for midstream and upstream programs, when compared to downstream programs, are highlighted and discussed in separate sub-sections.

2.3. Structure of the Document

The protocol is structured in key sections addressing the main evaluation task areas as outlined in Figure 2-1 and described below. The body of the protocol is designed to serve as a user-friendly and easy-to-follow guideline for non-experts, while the appendices offer additional guidelines and examples that require a basic level of understanding of evaluations. The key sections of the Protocol are:

- **Evaluation planning.** This section describes the evaluation administrator's development of the evaluation scope of work and procurement of evaluation services, and the evaluator's development of the evaluation plan.
- **Evaluation tasks.** This sections describes the steps to conduct impact and process evaluations, and provides guidance on how to determine cost-effectiveness and evaluate market effect. The impact evaluation section addresses both technology-based programs and behavioural-based programs. Almost all the evaluation tasks are the responsibility of the evaluator.
- **Reporting.** This section describes the development of evaluation reports and is mainly the responsibility of the evaluator.

Figure 2-1 | Evaluation Task Areas

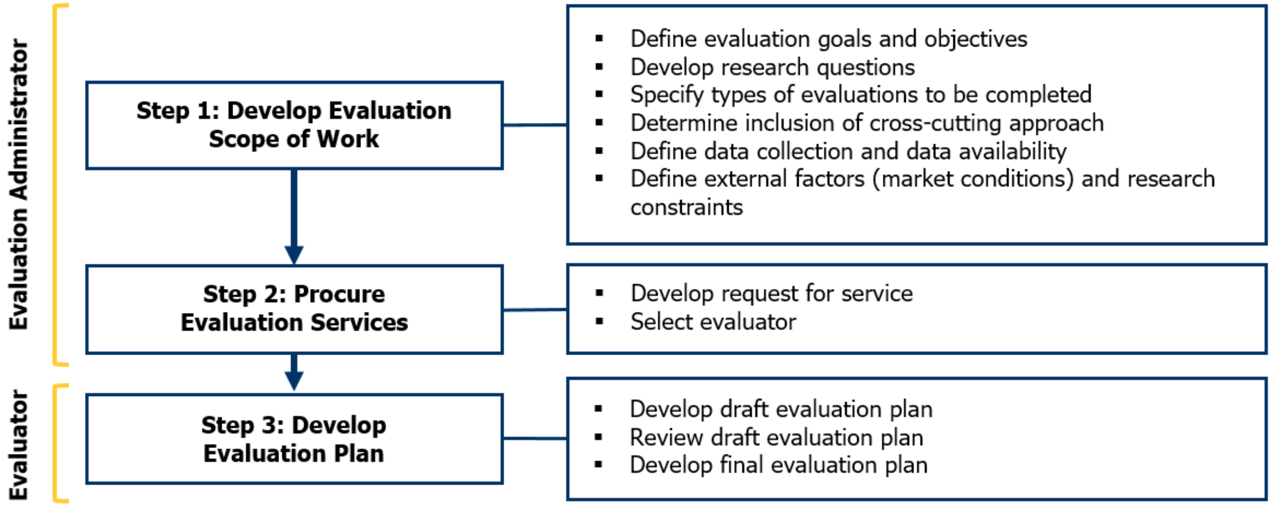


3. Evaluation Planning

Evaluation planning is the process of identifying the goals, objectives, and intended use of the evaluation. The objective of evaluation planning is to ensure that evaluation activities achieve their defined goals and objectives on schedule and within budget. The main steps in evaluation planning are outlined in Figure 3-1 and are described in the remainder of this section.

As depicted in Figure 3-1, the evaluation administrator is accountable for the initial evaluation planning steps, which are developing the scope of work, procuring services and selecting the vendor. The last step, after retaining the evaluator, is the development of the evaluation plan. The evaluator is usually responsible for developing the evaluation plan in consultation with the evaluation administrator. The development of the scope of work is described in Section 3.1, followed by the description of the procurement of evaluation services in Section 3.2 and the development of evaluation plan steps in Section 3.3.

Figure 3-1 | Evaluation Planning Steps


Evaluation Administrator
Evaluator
Step 1: Develop Evaluation Scope of Work
Step 2: Procure Evaluation Services
Step 3: Develop Evaluation Plan
▪ Define evaluation goals and objectives
▪ Develop research questions
▪ Specify types of evaluations to be completed
▪ Determine inclusion of cross-cutting approach
▪ Define data collection and data availability
▪ Define external factors (market conditions) and research constraints
▪ Develop request for service
▪ Select evaluator
▪ Develop draft evaluation plan
▪ Review draft evaluation plan
▪ Develop final evaluation plan

Key items to consider during the evaluation planning process are provided in the information box below. Accounting for these aspects ahead of time will improve outcomes and streamline the evaluation process.

Key Items to Consider During Evaluation Planning

When planning evaluations, the evaluation administrator needs to consider how the evaluation serves as a management tool. In addition, the evaluation provides savings estimates that demonstrate program impact and cost-effectiveness, which may be used for regulatory purposes. Evaluation findings are also often used to improve both short-term and long-term impacts, allowing mid-course revisions to enhance program achievements. To realize these benefits, it is important to recognize that evaluations are not solely intended to be program performance audits.

Evaluation planning includes the allocation of program resources for monitoring, measuring, verification, and reporting of performance and results of individual programs (or portfolios). Deployment of program resources into program evaluations involves simultaneous consideration of:

- The prioritization of the program decisions to which the evaluation will contribute.
- The resources needed to satisfy the evaluation's goals and objectives and what the program can afford (for example, impact targets, market transformation, behavioural changes, and non-energy benefits (NEBs) such as job impacts or the reduction of greenhouse gas emissions).
- The timeline and rigor for all evaluation activities and results.

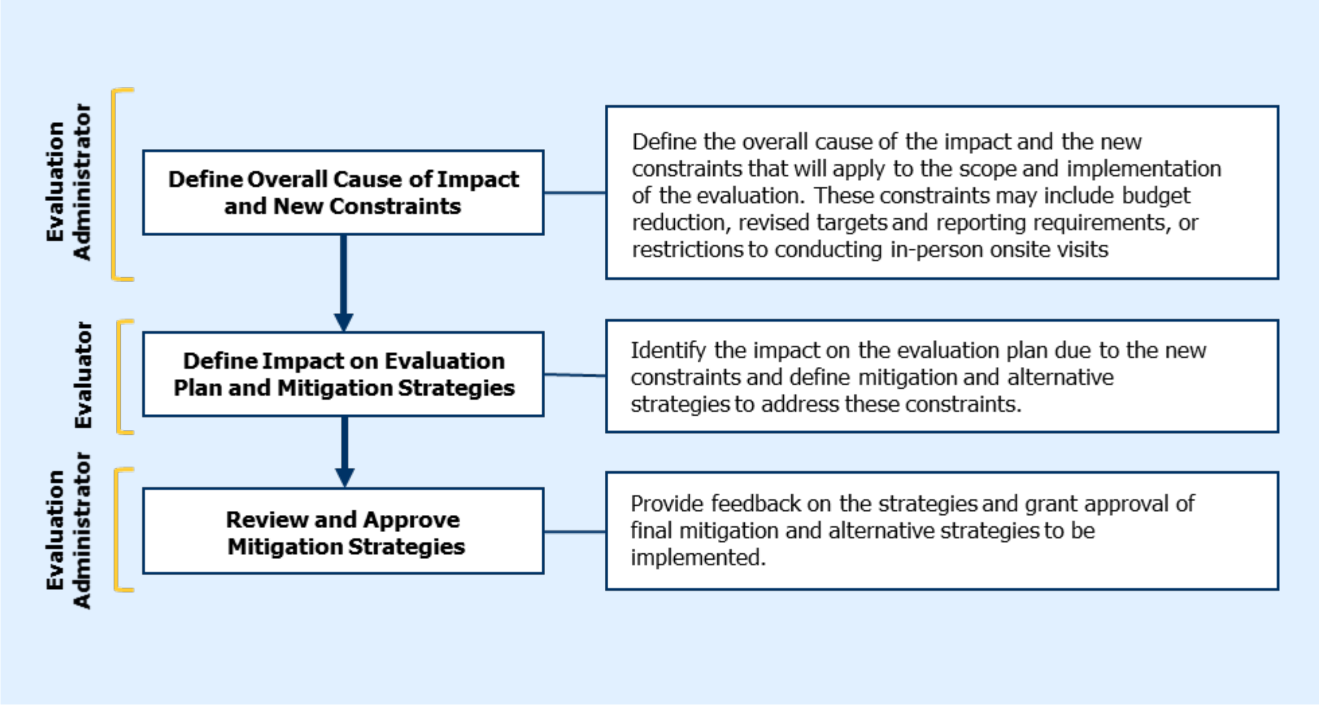
When planning the evaluation, it is important to consider that resolving low priority issues or employing unnecessarily complex methods is often not the best use of valuable resources. When faced with limited evaluation resources, prioritizing the key activities will ensure the evaluation objectives are met without straining resources. To help ensure the usefulness of evaluations, planning is usually performed during the early stages of a program's lifecycle and the following items are to be considered:

- **Integration of evaluation into the program implementation cycle.** 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.
- **Program design.** The draft evaluation plan, or as a minimum the evaluation scope, is prepared as part of the program design and an evaluation budget is assigned during this stage. Upon completion of the program design, the evaluation plan is implemented to ensure data is collected and reported on time, allowing for incremental feedback to guide program managers.

- **Preparing for program launch.** Ideally, the draft evaluation plan is prepared before the program is launched. If it cannot be developed before the program launch, it is recommended to be drafted as soon as possible post-launch. Baseline data is to be collected before, or soon after, new equipment is installed (or new behaviours are suggested) so that market effects resulting from the program offers are documented and their impacts on the baseline are minimized.
- **Defining the evaluation objectives.** Evaluations focus on the linkage between program outputs and the resulting program outcomes. The evaluation guides the program administrator on ways to enhance program efficacy. To this end, program administrators and regulators need to be assured that the type and quality of required information can be generated by the evaluation.
- **Plan for program risks, such as disruptions or changes.** In collaboration with the evaluation administrator, the evaluator can develop a list of potential disruptions based on previous experience and potential scenarios. Each disruption usually includes a mitigation plan to minimize the influence on evaluation quality, schedule, and budget as well as an assessment of the likelihood for each scenario.
- **Program implementation.** Some baseline data collection and program reporting occur throughout program implementation. This incremental data is often reviewed within a pre-determined timeline to inform and update program metrics that can be evaluated every few months or a couple of years. Evaluation administrators can analyze and present performance metrics to program managers as findings from the evaluator while keeping in mind that evaluation activities often continue after the program year is completed.
- **Incentive stacking.** As energy efficiency programs become more diversified and penetrate deeper into the market, their boundaries may become less clear. Therefore, customer target groups and claimed incentives might start to overlap. Incentive stacking occurs when a participant can claim incentives from two different programs. The evaluation administrator needs to identify potential programs where incentive stacking can occur and reach an agreement with program administrators to determine the allocation of savings.

Unanticipated external conditions may have a disruptive impact on evaluation activities and will possibly require the evaluation to be adjusted. These external impacts include, for example, an unexpected end to a program during the evaluation period, changes to the operating hours at a facility, or unexpected policies to restrict in-person contact due to health and safety requirements. These changes in external conditions can have a disruptive impact on the evaluation activities and the evaluation administrator and evaluators need to allow for the flexibility to implement the steps that are summarized in the flow diagram below.

If an adjustment or amendment of the evaluator’s scope of work is required, then the evaluation administrator and evaluator are to follow the terms and conditions for scope amendments, as specified in the evaluation services contract between the evaluation administrator and evaluator.


Evaluation Administrator
Define Overall Cause of Impact and New Constraints
Evaluator
Define Impact on Evaluation Plan and Mitigation Strategies
Evaluation Administrator
Review and Approve Mitigation Strategies

Define the overall cause of the impact and the new constraints that will apply to the scope and implementation of the evaluation. These constraints may include budget reduction, revised targets and reporting requirements, or restrictions to conducting in-person onsite visits

Identify the impact on the evaluation plan due to the new constraints and define mitigation and alternative strategies to address these constraints.

Provide feedback on the strategies and grant approval of final mitigation and alternative strategies to be implemented.

3.1. Step 1: Develop Evaluation Scope of Work

The first step of evaluation planning is the development of the evaluation scope of work, which entails defining evaluation goals and objectives, developing research questions, and selecting the types of evaluations to be completed.

3.1.1 Define Evaluation Goals and Objectives

The first step in developing an evaluation scope of work is defining the intended use of the evaluation findings. The evaluation administrator identifies how the findings of an evaluation will be utilized beyond the determination of verified savings. For example, a program design team may commission a research study to use the findings to assist in the design of a program that estimates measure level effectiveness.

The reasons for the evaluation need to be indicated, which usually is dependent on the use of the evaluation findings. Common examples of evaluation end-use may include:

- **Administrative or compliance**, to verify program savings as per government directive or other regulatory requirement.
- **Experimental**, to measure the effectiveness of a pilot program.
- **Operational**, to determine the effectiveness of a program delivery approach.

In addition to defining the intended use of the evaluation findings, available evaluation budgets need to be considered when developing the evaluation goals and objectives. Evaluation budgets may be constrained, which necessitates a balance between cost and evaluation rigor. Evaluation administrators and program managers usually collaborate to strategically identify which elements are to be evaluated to achieve evaluation goals and objectives while staying within budget.

Upon defining the intended use of the evaluation findings and considering the available budget, evaluation goals and objectives can be identified and documented in the scope of work. This will help provide a sense of how the findings will be presented in the final report and presentations.

3.1.2 Develop Research Questions

Evaluation goals and objectives defined in the preceding step can be used by evaluation administrators to formulate research questions. An example of a research question is, "Are program designs and supporting organizational controls adequate to achieve the objectives of the program?"

General research questions, which are directly derived from the evaluation goals and objectives, are usually created first. Each general research question can then logically produce multiple specific research questions. The specific research questions can be answered through data collection and analysis to deliver insights into the general research question.

Each research question is drafted by considering distinct research factors, including research design, sample sizes, relevant comparison groups, data collection methods, analytical approaches, and others. As a result, there may only be a few research projects that can effectively answer more than a handful of research questions. The narrowing of research questions is a fundamental activity within evaluation planning and is necessary for a manageable study. Evaluation administrators generally narrow the inquiry to a few well-crafted research questions.

Clear and concise research questions help establish consensus among evaluation stakeholders and provide guidance on the areas of investigation, thus increasing the likelihood of valuable findings, insightful conclusions, and useful program recommendations. Example questions are provided for each type of evaluation in the subsequent sections. The following key items can guide the phrasing of research questions:

- The questions flow directly from the evaluation objectives.
- The questions are specific and solicit significant findings.
- The questions can yield actionable answers.
- The questions are answerable within the constraints of the evaluation budget and other resources.

3.1.3 Specify Types of Evaluations to be Completed

Having defined the evaluation goals and objectives and created clear research questions, evaluation administrators can then determine the types of evaluations that need to be completed to ensure the evaluation goals and objectives are achieved. The evaluation administrator lists the required types of evaluations to be completed while developing the statement of work. The common types of evaluations include:

- Impact evaluation
- Process evaluation
- Market effects evaluation
- Cost-effectiveness evaluation
- Outcome evaluation

It should be considered that the analytical methods used in each type of evaluation will depend on the type of program being evaluated. For example, evaluators will use different analytical methods to evaluate and report findings from technology-focused or behavioural programs in Sections 4.1 and 4.2. Each type of evaluation is described below.

Impact Evaluation

Impact evaluations assess both the intended and unintended effects that can be attributed to a program, policy, or project. They are the most rigorous of all evaluations because attribution must be mapped from program outputs[2], through observed outcomes[3] and to the tangible impacts achieved. Such evaluations are most appropriately applied to measure the change in energy consumption and/or demand caused by a program. This can include M&V engineering processes used for developing a new or improved estimated savings.

For an impact evaluation, the contribution of external factors towards the achievement of desired impacts are limited to factors that are reasonable and can be accounted for within the analysis. For example, the installation of building insulation usually considers reasonable external factors such as weather conditions and the interior temperature setpoint. In general, an impact evaluation addresses the following question: what are the verified quantifiable effects (impacts) attributable to the program? Sections 4.1 and 4.2 describe how impact evaluations are conducted.

Examples of Research Questions Used in Impact Evaluations:

- What is the direct impact of the entire program on energy savings and demand reductions?
- What is the direct impact of individual program elements or behaviours on energy savings and demand reductions?
- What is the direct impact of the overall program on non-energy impacts?
- What is the magnitude of observed effects and what proportion of those effects can be attributed to the program?
- What key factors are responsible for the verified savings?

[2] A term used generically with logic modeling to describe all of the products, goods, and services offered to a program's direct customers.

[3] A term used generically with logic modeling to describe the effects that the program seeks to produce. It includes the secondary effects that result from the actions of those the program has succeeded in influencing.

- What could have caused the observed energy-saving behaviours, if they were not caused by the program?
- What behaviours were adopted by program participants when compared to those of non-participants?

Process Evaluation

Process evaluations are conducted to evaluate a program's performance and/or identify lessons learned to help guide future program strategies. This evaluation includes reviews of the program's policies, procedures, practices, and organizational controls that were implemented during the period under review. Process evaluations also assist in identifying the strengths and weaknesses of a program and identify opportunities for improved operational efficiencies. Through these data collection and analysis efforts, process evaluations verify program expenditures, review the effectiveness of the services provided by the program, and document the resulting operational outputs compared to the program objectives. Section 4.3 describes how process evaluations are conducted.

Examples of Research Questions Used in Process Evaluations:

- Are program designs and supporting organizational controls adequate to achieve the objectives of the program?
- Is the program producing the intended outputs?
- How satisfied are the participants with the program's operation?
- Are resources reasonable relative to program objectives?
- How might the program be improved?
- How can the program be modified to improve cost-effectiveness or to enhance the stream of benefits?

Cost-Effectiveness Evaluation

Cost-effectiveness evaluations measure the collection of benefits against the costs associated with program design, implementation, administration, and evaluation. Cost-effectiveness is typically implemented at the program level by leveraging industry-established tests. The details of the tests required in Ontario can be found in Section 4.4.

Where the evaluation administrator or evaluator deems it appropriate, the evaluation may also involve exploring the cost-effectiveness of individual measures, program elements, specific program activities, delivery agents, and/or implementation procedures. Section 4.4 describes how cost-effectiveness evaluations are conducted.

Examples of Research Questions Used in Cost-Effectiveness Evaluations:

- How much did the program spend to achieve the verified energy savings and demand reductions?
- What benefits resulted from individual program activities relative to their costs?
- Was the program cost-effective and does it pass the cost-effectiveness requirements?
- How cost-effective were specific program activities delivered by program delivery agents, and what were the main reasons for differences in cost-effectiveness achieved by the different program delivery agents?

Market Effects Evaluation

Market effects evaluations assess both the short-term and long-term changes to structural elements of the marketplace caused by programs, policies, and projects. This type of evaluation also reviews the cognitive processes and behaviours of key market actors that lead directly to energy and demand savings. For example, conducting a market effects evaluation of available lighting technologies based on manufacturing restrictions placed on higher wattage equipment and comparing the results with existing equipment in storage or available from distributors.

Therefore, the market effects evaluation seeks to attribute transformational impacts on the market, resulting from the application of codes and standards, legislation, innovation, and capability-building initiatives. The outcomes from this type of evaluation will serve as a guideline for market transformation elements of efficiency programs and can contribute to the development of forecasted saving estimates.

Evaluation administrators often utilize market effects evaluations a year or two ahead of program re-designs. This allows program administrators to suggest future changes to target markets, adopt a long-term approach with proposed exit strategies, or to suggest that actors’ behaviours will remain outside the scope of the intervention based on the market effects findings. Section 4.5 describes how market effects evaluations are conducted.

Examples of Research Questions Used in Market Effects Evaluations:

- Have changes occurred in the willingness or ability to produce, distribute, or service new energy-efficient technologies?
- What changes or effects are associated with individual program components/activities?
- How have the behaviours of targeted actors changed over time?
- What external factors are related to the achievement of observed market effects? What is the strength of those relationships?
- How effective has the program been in reducing market barriers?
- Have desired behavioural outcomes continued over time?

Outcome Evaluation

Outcome evaluations are used to document causal links between program outputs and program outcomes. Program outputs are the products, goods, and services offered to a program's direct customers, while program outcomes are the effects that the program seeks to produce. These include the secondary effects that result from the actions of those the program has succeeded in influencing. Outcome evaluations measure outputs and outcomes (including unintended effects) to judge program effectiveness and may also assess program process to understand how outcomes are produced.

Alternatively, they are used to test elements of a complex program theory, where direct program impacts may be difficult to isolate from influences beyond the results from program-sponsored activities. This type of evaluation verifies cognitive and behavioural changes believed to be necessary for the achievement of program objectives. It is typically applied to assess the effectiveness of market transformation initiatives, policy directives, social programs, and other interventions in a multifaceted environment.

Examples of Research Questions Used in Outcome Evaluations:

- What are the secondary and tertiary benefits resulting from the program under consideration (for example, persistence, delayed implementations, spin-offs)?
- What was the nature and magnitude of the NEBs associated with the program or individual program activities?
- What were the causes of any unintended program impacts?

3.1.4 Determine Inclusion of Cross-Cutting Approach

A cross effect, also known as an interactive effects, is when a change caused by one end-use measure affects the energy or demand savings of another end-use measure. Evaluations that consider cross effects are referred to as cross-cutting evaluations.

The evaluation administrator can include a request for the evaluator within the scope of work to consider applying a cross-cutting approach when it is thought to be of value if applied. If the cross-cutting approach is selected to be conducted within the evaluation, the evaluator can describe how cross-cutting techniques will be used to optimize evaluation cost-effectiveness while adding to the reliability of evaluation findings within the evaluation proposal and the evaluation plan. This is due to the fact that different scenarios could theoretically result in either overstating or understating program savings.

Examples of When Cross-Cutting Is Useful:

A lighting program may involve the replacement of incandescent lamps with LED lamps that can provide the same lumen output with greater efficiency. Installation of the LED lamps also means that less heat would be emitted by the light source, which could reduce cooling loads (seemingly adding to efficiency gains when a space requires cooling) or increase heating loads (seemingly reducing efficiency gains when a space requires heating) if the installations occur in conditioned spaces. To account for these cross effects, cross-cutting analytical approaches (such as adding the lighting energy savings to the cooling load and subtracting them from the heating load) must be used where the effects are expected to be substantive.

Energy efficiency initiatives often have some effect on seasonal or peak demand. Therefore, the impact resulting from one or more energy efficiency initiatives affecting the same market as a demand response initiative should be factored into demand savings calculations. Evaluators will often look only at the direct influence of one program on another where a participant in one program has also participated in another program. By failing to use a cross-cutting approach in such a case, the evaluator risks understating savings.

3.1.5 Define Data Collection and Data Availability

The evaluation administrator needs to define, within the scope of work, what data will be available for evaluation, and when that data will be available. The administrator can consider asking the evaluator to propose alternative strategies for collecting the desired data and/or options for collecting similar data. If there are any constraints related to the data acquisition, the evaluation administrator needs to highlight these constraints in the request for proposal (RFP), or disclose them as soon as possible, to reduce the possibility of them affecting evaluation practices. Most evaluators can recommend alternatives for data collection if typical data sources are unavailable.

Where the data constraints are expected to be persistent, the evaluation administrator needs to indicate the steps in the scope of work that can be taken to ensure EM&V best practices are upheld. Timelines required to resolve data constraints are usually set out in the evaluation plan and time should be built into future evaluation cycles to ensure the constraints are resolved.

3.1.6 Define External Factors (Market Conditions) and Research Constraints

Considering the impact of external factors helps evaluators isolate and report on a program's cause and effect. Examination of external, non-program factors that could influence an expected outcome may reveal non-program relationships and suggest alternative justifications for observed outcomes. The process of examining the program's cause and effect, making the logical relationships explicit between varying program components, and considering external influences can suggest the need for changes to a program's design or the evaluation plan.

Where external influences prohibit the study of critical program elements, the constraints prohibiting the analysis need to be stated within the scope of work. The evaluation administrator usually narrows the areas of investigation before the evaluator begins their work. Doing so after-the-fact can jeopardize the evaluator's independence to explore a program's cause and effect. To identify unforeseen scenarios, the evaluator can outline steps in the evaluation to be taken during a disruption of evaluation activities, which aim to minimize the influence of the disruption on evaluation results.

3.2. Step 2: Procure Evaluation Services

The procurement of evaluation services, which includes the development of the request for services (RFS), or request for proposal (RFP), and the selection of the evaluator, is usually the responsibility of the evaluation administrator. These steps are described in this section.

3.2.1 Develop Request for Service

The evaluation scope of work developed in the previous section is used to develop the statement of work that forms part of the request for evaluation services, which can also be in the form of an RFP. This is developed by the evaluation administrator. An evaluation scope of work template is included in Appendix A: Evaluation Scope of Work Template.

The statement of work forms the basis of the request for consulting services. When issuing a request for evaluation services to vendors, information on the program's budget for services are normally not included, due to the following reasons:

- Evaluators will propose alternate methods and approaches to achieve the same end result. Since there is more than one appropriate and acceptable methodology to accomplish most program evaluation tasks, these alternative methods may have different cost implications. It is best to allow the proponents to detail their position as to why their combination of proposed quality and cost should outrank their competitors.

- Evaluation methodologies and best practices are evolving. Therefore, at any time, proposals could pose a new method for measuring performance results. A core purpose of the competitive RFP process is to spur this type of innovative and creative thought process. The evaluation administrator expects bidders to continually strive to provide the best value proposition.
- It will be rare that the absolute best quality approach will get selected or even proposed. A program evaluation is always a compromise between the highest level of rigor and available resources. Managing this balance and deciding which contractor to select is easier when a truly competitive process is followed for both the substance and cost portions of the job.

The RFS/RFP needs to specify the requirements for data security and the protection of privacy. These requirements are usually aligned with the local acts and regulations, and the policies of the evaluation administrator.

The development of the RFS/RFP and the public procurements are expected to comply with government procurement requirements. The overall objective of these requirements is to ensure the acquisition of goods and services are conducted in the most economical and efficient manner.

3.2.2 Select Evaluator

Once a valid RFS/RFP process has been held, a winning bidder can be selected following the government procurement requirements. The selection criteria usually include:

- Experience, skills, and qualifications
- Understanding of the deliverables
- Work plan
- Project management
- Pricing

When evaluating the budget proposed by evaluators, there are general guidelines on the appropriate amount to spend on evaluations relative to the size of a program. Small pilot studies, where very detailed information will help inform and reduce the risk of a potential broader roll-out strategy, could justify spending the same amount as the program itself. In comparison, a program that has been running consistently for several years and that has no new or unusual activity occurring may require only a basic level of field verification and auditing and thus does not require substantial expenditure. The cost to achieve a successful evaluation is also affected by whether multiple evaluation types are required (outcome, impact, process, market, cost-effectiveness) or just one.

Ways to avoid bias or the perception of bias in the selection process include employing a committee to select the evaluator and utilizing grading rubrics to evaluate proposals. It is generally best to form a cross-functional team, representing the varying interests in the evaluation results.

3.3. Step 3: Develop Evaluation Plan

Once an evaluation vendor is selected, an evaluation plan is developed to guide the evaluation activities. The evaluator authors the evaluation plan, which is based on the definition of the scope and the statement of work, as described in the preceding sections. The steps to develop the evaluation plan are as follows:

- Develop a draft evaluation plan.
- Review the evaluation plan with the evaluation administrator.
- Update the evaluation plan to a final version in which the feedback and comments of the administrator are addressed.

An evaluation plan typically includes the outline provided in the example below.

Example of Evaluation Plan Outline

1. **Program description:** Provide a short introduction of the program offer
2. **Evaluation approach:** Describe the details of the approach that will be followed for the evaluations to be included, suchs as impact, process and cost effectiveness evaluations.
3. **Evaluation deliverables and schedule:** A listing of all the physical deliverables that will be part of the evaluation and the schedule to deliver the deliverables.
4. **Reporting:** A description of the reporting activities and an outline of the reports to be developed.
5. **Communication protocol:** A description of the communication to be included as part of the evaluation, such as parties involved in the communication, frequency of communication and form of communication.

3.4. Summary

Evaluation planning is the process of identifying the goals, objectives, and intended use of the evaluation. The main steps in evaluation planning are:

- Step 1: Develop evaluation scope of work.
- Step 2: Procure evaluation services.
- Step 3: Develop evaluation plan.

The evaluation administrator is accountable for the initial evaluation planning steps, to develop the scope of work, which entails defining evaluation goals and objectives, developing research questions, and selecting the types of evaluations to be completed. The evaluation administrator is also responsible for procuring the evaluation services, where they develop a RFS and select an evaluator. The last step, after retaining the evaluator, is the development of the evaluation plan. The evaluator is usually responsible for developing the evaluation plan in consultation with the evaluation administrator.

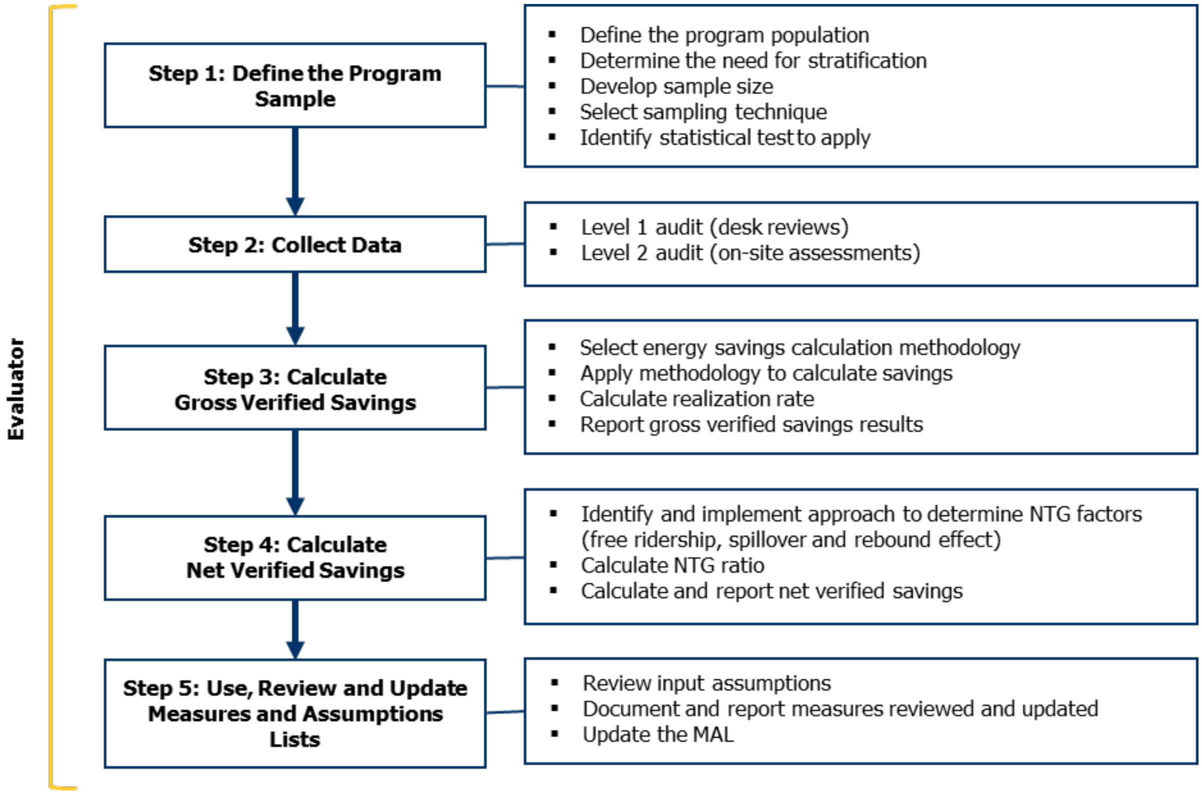
4. Evaluation Tasks

Following the development of the evaluation plan, the evaluator initiates the evaluation tasks as prescribed in the plan. The specific tasks completed during the evaluation depend on multiple factors, including goals and objectives of the evaluation, type of program under study and the types of evaluations to be completed as requested by the evaluation administrator. This section describes in detail the tasks required to complete a CDM program evaluation. First, in Sections 4.1 and 4.2, impact evaluations of technology-based programs, and behavioural-based programs are described, respectively. Section 4.3 presents the process evaluation tasks, followed by cost effectiveness in Section 4.4 and market effects evaluation in Section 4.5.

4.1. Impact Evaluation of Technology-Based Programs

Impact evaluations assess the outcomes of implementing a program, policy, or project. These evaluations are applied to measure a change in energy consumption and/or demand caused by a program or project. Impact evaluations of technology-based programs are discussed in detail in this section. The steps involved in conducting an impact evaluation of a technology-based program are summarized in Figure 4-1 and described in more detail in the remainder of this section.

Figure 4-1 | Technology-Based Program Impact Evaluation Steps


Evaluator
Step 1: Define the Program Sample
Define the program population
Determine the need for stratification
Develop sample size
Select sampling technique
Identify statistical test to apply
Step 2: Collect Data
Level 1 audit (desk reviews)
Level 2 audit (on-site assessments)
Step 3: Calculate Gross Verified Savings
Select energy savings calculation methodology
Apply methodology to calculate savings
Calculate realization rate
Report gross verified savings results
Step 4: Calculate Net Verified Savings
Identify and implement approach to determine NTG factors (free ridership, spillover and rebound effect)
Calculate NTG ratio
Calculate and report net verified savings
Step 5: Use, Review and Update Measures and Assumptions Lists
Review input assumptions
Document and report measures reviewed and updated
Update the MAL

4.1.1 Step 1: Define the Program Sample

When conducting impact evaluations, it is not always viable to study the entire program population (i.e. all participants). Similarly, for a comparison group (or control group), it is strenuous and rather implausible to study the entire range of eligible non-participants. Therefore, appropriate statistical sampling is utilized to select a representative sample of the populations under study whereby evaluators depend on a process known as sampling design.

Sampling design is the basis for defining the program sample by which the evaluator selects a sample that is representative of the population of interest. The sampling design may:

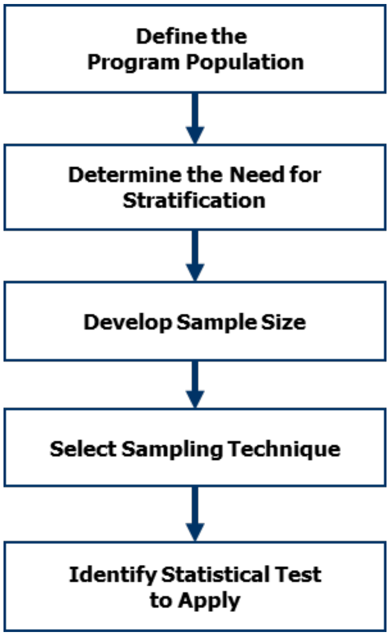
- Be identical to the population (for example, the population is too small to select a sample),
- Be only a part of the population, or
- Have an indirect relationship to the population (for example, the population is a specific lighting measure and the sample is a list of its suppliers).

Non-participant populations are often used as a comparison group in impact evaluations or to gain process evaluation insights. For example, non-participant responses are of value in understanding the challenges and barriers faced by customers in participating in a program, or the effectiveness of marketing to create awareness of the program among non-participants. Typical source of non-participant data include LDC or gas utility customer data sets and membership lists of associations. When emailing non-participants, the evaluator needs to ensure compliance with Canada's anti-spam legislation (CASL),[4] meaning the emails should not contain content that encourages participation in a commercial activity, such as promoting or offering to sell a service.

Steps to undertake a sampling design are summarized in Figure 4-2 and described in more detail below. Additional information regarding the benefits of sampling and elements to consider when designing a sample is provided in Appendix B: Sampling Plan Design.

[4] Government of Canada (2015). *An Act to promote the efficiency and adaptability of the Canadian economy by regulating certain activities that discourage reliance on electronic means of carrying out commercial activities, and to amend the Canadian Radio-television and Telecommunications Commission Act, the Competition Act, the Personal Information Protection and Electronic Documents Act and the Telecommunications Act*, S.C 2010, c.23. Website: https://laws-lois.justice.gc.ca/eng/acts/E-1.6/page-1.html#h-176920

Figure 4-2 | Sampling Design Steps


Define the Program Population
Determine the Need for Stratification
Develop Sample Size
Select Sampling Technique
Identify Statistical Test to Apply

Define the Program Population

When designing a program sampling plan, the initial effort is defining the population of the program. This is required to ensure the sample is representative of the population. For example, when selecting a sample for an impact evaluation of a technology-based program, the population is everyone who participated in the program and received a form of incentive. The level at which the impact results are to be assigned (e.g. provincial, regional, or individual utility level) should also be identified. Defining the appropriate population from which the sample will be selected and the relevant population attributes to be represented by the sample is critical to avoiding bias and ensuring a representative sample. For example, selecting a small rural customer base for a provincial program would not be representative of an urban area; nor is it likely to resemble the province. Similarly, it may not be accurate to estimate savings for this small rural customer base from a broadly scoped study used to establish provincial savings estimates. As such, it is essential to describe the characteristics of the population, including size and variance.

Determine the Need for Stratification

The definition of both the sample and program population informs the type of conclusions that can be drawn from an evaluation. This often requires stratifying (sub-dividing) the population. Stratification is sorting the population into distinct groups/categories based on common characteristics. For example, there might be a need to estimate provincial level savings and allocate these savings to individual groups (e.g. regions, business types). As a result, it may be necessary to sub-divide the population into strata by individual groups or by a stratum of different groups with similar characteristics. This allows for drawing inferences about the sub-populations that would otherwise be overlooked in a broadly defined sample.

It is beneficial to apply stratification if it will achieve the following conditions:

- Variability within the individual strata is reduced,
- Variability between the different strata is maximized, and
- Variables used to stratify the population are strongly correlated with the desired dependent variables.

For the accurate application of stratification, information on the characteristics of the population is required. In the absence of this information, the evaluator may resort to alternate and more advanced statistical methods to define the appropriate strata. The two most common advanced stratification methodologies are shown in Table 4-1.

Table 4-1 | Advanced Stratification Methodologies

Advanced Stratification Methodology	Definition	Example
Over-sampling	Creating biases for the sampling process to address a known about the population, such that the findings better represent the study population.	If it is known that there is a high non-response bias from a certain participant demographic, the evaluator may choose to over-sample this population or sub-population to ensure the number of responses received meets statistical requirements.
Post-stratification	Developing estimates about sub-populations after the data collection is complete. This can be used if the characteristics of the sub-populations are unknown at the time of data collection.	An example of post-stratification in the residential sector is when income levels of participants are unknown. The stratification by income level may be considered important for the evaluation of the program and the information is obtained during the data collection stage. Post-stratification by income level can be done after the data collection.

These advanced techniques are generally reserved for specific situations and used only after careful consideration of basic stratification techniques. Additionally, the use of these methods needs to be documented in the sampling design section of the evaluation plan.

Develop Sample Size

While sampling has several advantages, it will not perfectly represent the entire population under study. Sampling can lead to errors and inaccurate conclusions about the population. Thus, the evaluator needs to ensure that the sampling strategy provides an acceptable and agreed-upon precision level, as required by the program administrator.

Confidence and precision are two factors to consider when developing a sample size. Specific confidence and precision values are often a requirement defined by the evaluation administrator. An evaluation requirement might be for the savings estimates to be ± 5% (precision) at a 95% level of confidence. Repeated sampling of the population would then result in mean savings estimates that is within 5% of the true mean of the population 95 times out of 100.

To determine the sample size required to achieve the desired level of confidence, the evaluator will need to make assumptions regarding the normal variance around the population mean since it is unknown. Typically, the coefficient of variance (CV) is set at 0.5 when other studies are not available to construe the likely variance around the sought population mean. The setting of the coefficient of variance at 0.5 is often acceptable because such a coefficient is indicative of neither a weak nor strong dispersion. In evaluations that span multiple program years, historical CVs usually offer a more reliable measure of relative variability in the program than the standard assumption of 0.5. In consultation with the evaluation administrator, the evaluator may calculate and trend the CV of previous evaluation cycles and consider using an average of the actual CV to accurately reflect the variability in the program data. Additional information about precision and confidence is provided in Appendix B: Sample Plan Design.

Select Sampling Technique

After the population stratification (if applicable) and sample size have been identified, a representative sample of the defined populations can then be selected. It is important to use an appropriate sampling technique to address biases during sample selection. Despite the chosen sampling methodology, it is important to keep in mind that the sample needs to represent the population included in the study. The sample is ideally selected from the entire program population, and not from a population of convenience. For instance, selecting a sample of individuals from a pool of participants who volunteer to complete a questionnaire is not a suitable practice. Selecting these individuals is convenient, but they typically have strong opinions and/or possess more knowledge about the program. Therefore, they are not necessarily representative of the entire population participating in the program.

The most utilized probability sampling techniques are listed in the information box below. There are many other sampling techniques available for use, and the evaluator needs to provide an explanation in the evaluation plan and final report for selecting a specific sampling technique.

Most Common Probability Sampling Techniques

- **Simple random sampling:** This involves randomly assigning members from the study population to the sample. This could leverage software to randomly assign, for example, 15% of program participants to the sample.
- **Systematic sampling:** This involves systematically assigning members from an ordered study population to a sample. For example, every 12th participant entering a program may be selected for the sample.
- **Matched random sampling:** This involves selecting members from the population based on relevant characteristics and assigning them to a group, then randomly selecting samples from within each group. For example, the evaluator may decide to categorize participants by facility size and select a random sample from each group. This technique may be used to select a comparison group when studying a program. Alternatively, the use of a matched control group can be used to normalize estimates obtained for a study population.
- **Quota sampling:** This is when the evaluator is asked to sample a fixed number of members that meet specific criteria and assigns them to a sample. For example, a researcher may be asked to survey 400 rural households and 300 urban households. Quota sampling relies on the researcher's judgement and convenience in sample selection. This deems quota sampling a non-proportional (biased) sampling technique.
- **Panel sampling:** This involves a longitudinal study of a previously defined sample. For example, this approach may be employed to infer how a population is likely to react to an increase/decrease in energy prices.

Some instances may exist where non-probability sampling is required. For example, conducting a study to understand electricity use across a province would ideally have sub-populations under study, such as industrial or manufacturing facilities. These sub-populations are not typically a homogeneous group, as there are fundamental differences between the energy use of the various industrial and manufacturing facilities. In such cases, a random sampling of this stratum could lead to unintended biases, particularly the selection of unusually large or small facilities, whose energy use is not representative of the stratum. To resolve this issue, a non-probability sample may be employed for the non-homogenous strata and a random probability sample can be used for the remaining strata. In such a scenario, it is best to leverage the expertise of a subject matter expert or a sector specialist to define a representative sample of the population. For example, the sector specialist may be able to isolate some of the odd facilities from the stratum and systematically select a sample from the remaining facilities that can represent the entirety of the group. Accurate results can be achieved when using this method, in comparison to what would be achieved using a simple random sample.

Similar to advanced stratification techniques, non-probability sampling must be carefully considered to ensure that sampling bias is explicitly identified and kept to a minimum. Additionally, the details of the non-probability sampling need to be described in the evaluation plan.

Identify Statistical Test to Apply

Statistical testing is generally used to describe a given population, make comparisons against a hypothetical value, or establish predictions based on known values. While there are various types of statistical test models, the most suitable one to employ is one that accurately answers a particular research question(s). Multiple tests may be considered suitable for addressing a research question. In such cases, it is recommended for the evaluator to consult with a statistics professional before selecting and applying a statistical test. Additional information about selecting a statistical test is provided in Appendix B: Sampling Plan Design.

4.1.2 Step 2: Collect Data

Data collection is initiated following the completion of the sampling design by the evaluator. Within the scope of impact evaluations of technology-based programs, data collection activities are usually referred to as project audits. Guided by the sampling plan, the evaluator obtains the project files for the participants included in the sample from the evaluation administrator. Depending on the nature of the evaluated program and availability of data sources, the evaluator may choose to collect data through various means. Impact evaluation data collection methods often include the following:

- **Level 1 audits** (also referred to as desk reviews). Evaluators review project documentation available in the program database, including applications, savings worksheets, and any other relevant documentation needed to recreate savings calculations. These audits can also include financial and eligibility review.
- **Level 2 audits** (also referred to as on-site assessments). These expand upon Level 1 audits and include on-site reviews of equipment installation or telephone interviews of selected participants. These audits are most suitable for projects involving custom M&V approaches. Level 2 audits consider all implemented measures within a sampled project. However, for large projects, where it is not feasible to review all implemented measures (for example, a lighting retrofit where 1,000 fixtures were replaced), the evaluator can consider nested sampling, where a sub-sample of lighting fixtures or equipment is selected for review.

The evaluator plans and executes the project audits to collect and review data and to ensure the data is suitable to achieve the evaluation’s goals and objectives. The main objective of impact evaluations of technology-based programs is to determine the impact on energy use and/or peak demand resulting from an intervention of the program. Therefore, project audits represent a significant portion of the impact evaluation efforts and contribute to the accurate evaluation of programs.

Data Collection for Midstream and Upstream Programs

Unlike evaluations of resource-acquisition programs, evaluations of market-transformation programs, such as midstream programs, require additional upfront coordination between the evaluation and implementation teams. Data collection requirements should be clarified prior to the program launch and should consider the metrics to track program performance and the short-term, midterm, and long-term market indicators. Data collection is required at multiple stages to determine sales levels, including:

- Before the program launch to determine the historical monthly sales levels and define the baseline levels. A formal market assessment can be utilized to establish the baseline levels of the market indicators, including the current market share of the incented products, retailer awareness of energy efficiency levels, and current stocking and promotional practices.
- During the program to determine the sales volume of the qualified product above baseline levels.

Historical data need to be collected for a sufficient period of time to ensure that any seasonal sales patterns are reflected. In many cases, this means collecting at least one year of historical data. However, if it is known that sales do not vary significantly throughout the year, less data may be acceptable.

Occasionally, distributors are hesitant to share their records in an effort to protect their proprietary sales data. A common practice to protect distributors' data is to sign a non-disclosure agreement (NDA). Additionally, data privacy issues can be mitigated by implementing a secure, online, password-protected file-sharing system through which program partners and distributors can transmit and view confidential sales data and other information.

4.1.3 Step 3: Calculate Gross Verified Savings

The data collected from Level 1 and Level 2 audit activities allow the evaluator to recalculate the savings for each sampled project – an effort that is referred to as gross verified savings. Gross verified savings calculations are based on the difference between energy and demand use after the implementation of a program and an assumed set of baseline conditions that estimate what energy consumption and demand would have been in the absence of the program. Equation 4-1 shows the general formula that applies when calculating project or measure level gross verified savings for technology-based programs.

Equation 4-1 | Project or Measure Gross Verified Savings

$$\boldsymbol{Gross\ Verified\ Savings = Baseline\ Use - Reporting\ Period\ Use \pm Adjustments}$$

Where:

- Baseline use is the energy or connected demand consumption that is estimated to have occurred before the implementation of the program. The baseline period is selected to be representative of normal operations. Depending on the type of program under study, the evaluator may employ various methodologies to define the baseline energy and demand use. For example, for a new construction program, minimum code standards are usually used as measures' baseline, or, in some cases, an independent baseline study might be required.
- Reporting period use is the consumption that occurs following program implementation.
- Adjustments account for independent variables that are beyond the controls of the program, implementer or participant. Adjustments are meant to bring the baseline and reporting periods to the same set of conditions (rather than a simple subtraction of pre- and post-installation energy and demand use). Common independent variables that are adjusted for include:
  - Weather normalization
  - Occupancy levels and hours (i.e. hours of operations)
  - Production levels (i.e. operating cycles, shifts)

In addition to calculating energy and connected demand savings, there might be a need to calculate peak demand savings. Key elements to consider when calculating peak demand savings are listed in the information box below.

Key Considerations when Calculating Gross Verified Peak Demand Savings

The concept of peak demand is not simply the highest demand for electricity in 24 hours. Rather the concept relates to energy demanded throughout a pre-defined period, for example, 1 pm – 7 pm, during which the overall demand on the electricity grid tends to be, on average, higher.

The system peak could occur in either the summer or winter seasons. Although in Ontario, summer peak has been dominant, such a pattern may or may not necessarily persist and the system may experience a winter peak.

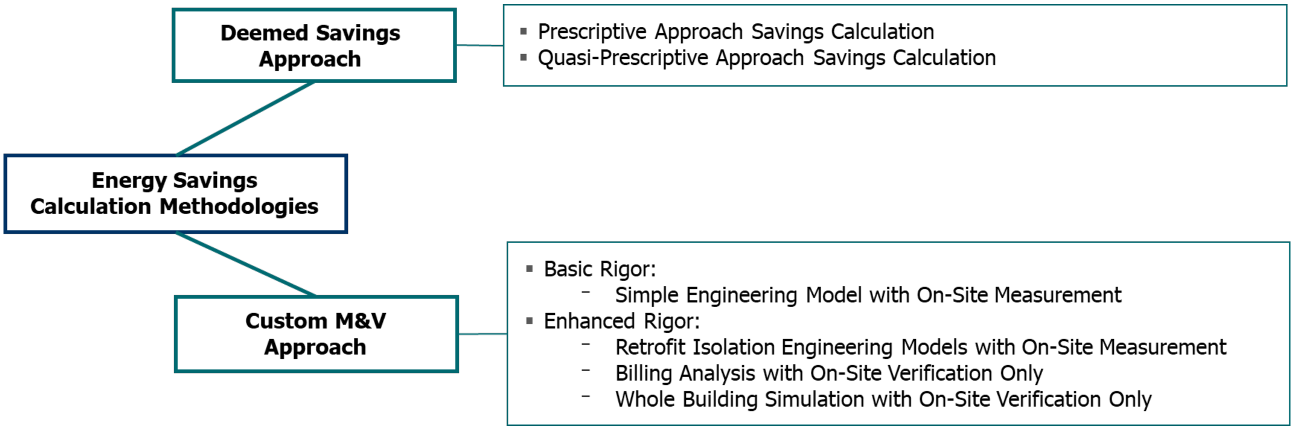
To ensure accurate calculations of gross verified peak demand savings, the following should be considered:

- Determine the pre-defined blocks of hours whereby demand is generally at its highest. The hours that count towards the savings target should be known in advance and remain constant for the full program cycle.
- Peak demand saving estimates are to be based on the average demand reduction across the total number of hours in the appropriate peak summer or winter blocks.
- An alternative method can be used to calculate peak demand savings for facilities with variable load characteristics or weather-sensitive measures. Peak demand savings are calculated based on a weighted average of the maximum demand reduction in each of the three months that occurs within the peak blocks.

To calculate the gross verified savings, the evaluator must first select the appropriate methodology and then apply it. As depicted in Figure 4-3, there are two types of savings calculation methodologies that are applicable to technology-based programs:

- Deemed savings approach
- Custom M&V approach

Figure 4-3 | Savings Calculation Methodologies


Energy Savings Calculation Methodologies

Deemed Savings Approach

- Prescriptive Approach Savings Calculation
- Quasi-Prescriptive Approach Savings Calculation

Custom M&V Approach

- Basic Rigor:
  - Simple Engineering Model with On-Site Measurement
- Enhanced Rigor:
  - Retrofit Isolation Engineering Models with On-Site Measurement
  - Billing Analysis with On-Site Verification Only
  - Whole Building Simulation with On-Site Verification Only

Deemed savings approach uses agreed-upon values for program-supported measures with well-known and documented saving estimates. Deemed savings are determined by the evaluator using prescriptive and quasi-prescriptive assumptions and standard equations for determining gross verified savings. More information about these assumptions and equations are provided in Appendix C: Technology-Based Programs Energy Savings Calculation Methodologies. When applying the deemed savings approach and using the Measures and Assumptions List (MAL) or Technical Resource Manual (TRM), field measurements are not commonly required for determining the savings per measure or project. Gross impacts are determined by multiplying the savings per measure values derived from the MAL or TRM by the verified number of installations. If a quasi-prescriptive approach will be used, it may be necessary to verify additional information, such as facility type or the system's end-use.

Custom M&V approach is typically applied for measures that need accurate measurements to determine savings or with measures that have varied input assumptions. Custom M&V approaches require the tracking of gross verified savings and estimation on a project-by-project basis. Custom projects tend to be more complex than those using prescriptive measures, for example, building equipment retrofits, where equipment load profiles are variable and saving estimates utilize equations that can change on a project-by-project basis. Therefore, project-level M&V is essential for tracking and reporting savings and need to be taken into consideration for all situations requiring a custom M&V. Custom projects that require implementing a custom M&V approach include equipment retrofit(s) and/or operational change(s). Custom M&V approaches are based on the widely recognized International Performance Measurement and Verification Protocol (IPMVP)[5].

Appendix C: Technology-Based Programs Energy Savings Calculation Methodologies provides more detail about the energy savings calculation methodologies. Guiding factors to consider when selecting a methodology are provided in the information box below.

Factors to Consider when Selecting a Calculation Methodology

- The program implementation strategy and the types of data available for collection during program delivery.
- The types of measure(s) supported by the program (for example, simple, mass-market versus complex, commercial, or industrial measures).
- The perceived accuracy of previous evaluations or assumptions, such as those identified in the MAL.
- The amount of energy and demand savings expected to result from the program.
- The time and budget available for the evaluation.
- The professional experience and judgement of the evaluator.

[5] Efficiency Valuation Organization. *International Performance Measurement and Verification Protocol*. Website: https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp

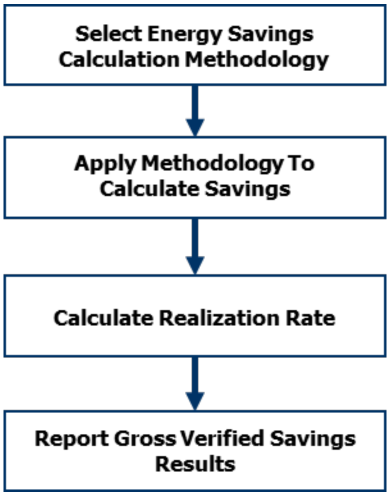
At the program level, the ratio of gross verified savings to the reported savings is referred to as the realization rate. Both gross verified and reported savings values are often reported together. The measure or stratum-level realization rate is the weighted average for all projects in the stratum sample. The stratum-level realization rate can be calculated by dividing the stratum-level total gross verified savings for the strata by the stratum-level total reported savings. Reporting of gross verified savings by the evaluator includes the following information:

- Methodology or methodologies used to verify and assess reported savings.
- Sampling plans and survey instruments used to collect data.
- Confidence and precision of data and results.
- Total gross verified savings and sample calculations.
- Explanations, where possible, of variances between verified and reported savings for the program.

The total gross verified savings for the program reflect the direct impact of the program. These savings do not account for customer behaviour or market effects that may augment or lessen a program's direct results. Adjustments to capture the customer behaviour or market effects are included through tasks carried out to calculate net verified savings, which are discussed in the sections below.

To summarize, the steps to calculate the gross verified savings are outlined in Figure 4-4.

Figure 4-4 | Savings Calculation Methodologies



Select Energy Savings Calculation Methodology

Apply Methodology To Calculate Savings

Calculate Realization Rate

Report Gross Verified Savings Results

Key Considerations when Calculating Gross Verified Savings for Midstream and Upstream Programs

As explained in Section 4.1.2, midstream programs evaluations rely on historical sales data to set an appropriate baseline, which is used to determine program qualified sales.

Subsequent to the collection of historical data, the evaluator needs to establish the non-program baseline of the qualified products (for example, the sales data of the qualified product in the program's absence). Historically, evaluators have taken two approaches to determine the influence of midstream programs; by establishing baseline sales of efficient products through either (1) historical comparison or (2) geographic comparison.

Simple historical comparisons risk the possibility that market changes independent of the program will limit the baseline period's relevance for comparison. This is particularly critical for product categories with short product-refresh cycles and those undergoing rapid technological change, such as lighting equipment. As a result, evaluators need to establish an approach that considers the diffusion of energy-efficient products in the market and any other parameters that may hinder the comparison's accuracy.

To mitigate the risk of using a simple historical comparison, forecasting the market share of efficient products without program intervention is recommended. One approach for baseline forecasting is to create mathematical models to predict efficient technologies' diffusion over time. Another approach is to use targeted analyses comparing the periods immediately before and after limited-time promotions or comparing the stores that received promotions with those that did not. Depending on the technology under study, data availability and budget, the evaluator often works with the evaluation administrator to develop the most appropriate baseline forecasting approach for the program under study.

The forecasted baseline is compared to the actual post-implementation program-period sales data. The difference between the program-period data and the forecasted baseline is the program's net effect, also referred to as the sales lift.

Following the calculation of the program's net effect (the number of the qualified sold products attributable to the program), the gross energy and demand savings are calculated. This calculation includes multiplying the number of qualified sales attributable to the program by the per-unit energy and demand savings. The per-unit energy and demand savings are usually developed in consultation with the program and evaluation administrators.

4.1.4 Step 4: Calculate Net Verified Savings

Net verified savings recognize behavioural factors and represent benefits that are only attributable to, and the direct result of, the program in question. Net verified savings are particularly important for public or ratepayer-funded programs, where the responsible party is interested in the influence of the program when it is producing incremental savings.

Program net verified savings are calculated by multiplying the gross verified savings with the net-to-gross (NTG) ratio. The gross verified savings are determined by multiplying the reported savings with the realization rate, as shown in Equation 4-2. Guidance on how to calculate the NTG ratio is provided in the information box below.

Equation 4-2 | Program Net Verified Savings

$$\boldsymbol{PS_{net} = Reported\ Savings * Realization\ Rate * NTG\ Ratio}$$

Where:

- PS_{net} is the program's net verified savings.
- Reported Savings are savings as presented by the program administrator (kW and/or kWh).
- Realization Rate is the ratio of gross verified savings to reported savings, as calculated in the steps to determine gross verified savings.
- NTG ratio is calculated as described in this section.

Net-to-Gross (NTG) Ratio Calculation

Program net verified savings are estimated by adjusting (discounting or increasing) the gross verified savings through the application of a set of adjustment factors, including free-ridership rates, spillover effects, and rebound effects. The aggregate effect of these factors in a program impact evaluation is represented by the NTG. Free-ridership is the most commonly evaluated adjustment factor, followed by spillover and rebound effects. Deciding which of these factors to account for in an analysis of net verified savings is influenced by the goals, objectives, and constraints of the evaluation. The NTG ratio calculation is defined in Equation 4.3.

Equation 4.3 | Net-to-Gross Ratio

$$\textit{NTG Ratio=1-Free-Ridership+Spillover-Rebound Effect}$$

The value of the NTG ratio can vary drastically. Factors that influence the NTG ratio of a program include:

- How the program is implemented in the marketplace,
- The number of other programs that reach similar customer classes, and/or
- Other market influences, such as codes and standards.

Free-Ridership

Free-ridership is the program savings factor attributable to participants who would have implemented a program measure in the absence of the program. Though they may not be directly attributable to the evaluated program, savings occur as a result of free-ridership, and thus these effects reduce the direct impact of the program. There are generally three types of free-riders:

- Total free riders are participants who would have implemented the program-promoted measures in the same way and timeframe as in the absence of the program.
- Partial free riders based on efficiency and/or quantity are participants who would have implemented program-promoted measures but would have installed less efficient equipment and/or a lower quantity.
- Deferred free riders are partial free riders based on timing. These are participants who would have implemented program-promoted measures but at a later time.

Spillover

Spillover is a reduction in energy consumption and/or demand caused by the presence of an energy efficiency program. This extends beyond the program-attributed gross verified savings of the participants and without financial or technical assistance from the program. Additionally, these savings are not counted as part of another program within the same portfolio. Spillover can manifest in participants who take action beyond the program (for example, a small business owner who replaces inefficient non-lighting equipment with more efficient equipment due to participation in a small business lighting program). It can also manifest in active non-participants (customers who apply to but do not ultimately participate in a program) or true non-participants who adopt energy-efficient measures or behaviour due to program influence (for example, after being exposed to program marketing or acting on the recommendation of another participant).

To be able to determine spillover rates, questions regarding the installed measures and the impact the program had on the decision to implement the project should be addressed. Depending on the program, these non-program installations may include lighting, lighting controls, air conditioning, motors and motor drives, HVAC equipment, or appliances. For each installed measure, the NTG ratio data collection efforts will obtain details and specifications that will facilitate estimating the quantity of savings that the upgrade produced.

Rebound Effect

The rebound effect (also known as take-back) is the decrease in energy savings associated with the use of measures installed through a program. Some participants who experience lower energy costs because of the installation of the program measure may take-back some of those savings by using more energy. For example, after installing high-efficiency program-incentivized air conditioners, some participants may be inclined to operate them at cooler temperatures or more frequently compared to their former, inefficient air conditioner.

There are three common approaches to determine free-ridership, spillover and rebound effect to determine the NTG ratio for a specific program:

- **Self-reporting and enhanced self-reporting surveys.** These ask participants a series of questions to determine what actions they would have taken in the absence of the program.
- **Econometric methods.** These are mathematical models that use statistics, and energy and demand data from participants and non-participants to derive accurate NTG ratios.
- **Agreed on NTG ratios.** These can be used when the cost of conducting more detailed analyses of program NTG ratios is a barrier, or when the accuracy of the results is not paramount.

All three approaches can generally be used with any type of program. Econometric methods require large numbers of participants. Agreed on NTG ratios is the least costly approach, followed by self-reporting surveys and enhanced self-reporting surveys. The selected approach to determine the NTG ratio is usually discussed during the evaluation planning stage and described in the evaluation plan. Additional information on these approaches are presented below.

Self-reported Surveys and Enhanced Self-Reported Surveys

Self-reported surveys ask participants a series of questions to understand what actions they would have taken in the absence of the program. Estimates of spillover effects can be developed by surveying program participants and non-participants. Surveys can be web-based, distributed in hard copy, or administered by telephone. Self-reporting surveys are the lowest cost approach to estimating free-ridership and spillover rates for specific programs that support particular technologies or measures. It is preferred to conduct self-reported surveys as early as possible after the program implementation, to ensure respondents still have a relevant and accurate recollection of their decision-making process.

A word of caution about situations where respondents self-select for participation in the survey: self-selection bias can skew the results because those with strong opinions or higher degrees of knowledge about the subject tend to be more willing to take the time to participate in a survey.

A typical self-reporting survey asks a series of questions and may present respondents with an answer scale, rather than allowing for simple yes or no responses. A sample set of survey questions is provided below:

- Did you require financial assistance in order to go ahead with the install?
- Did you have previous experience with the energy efficient technology?
- Had you already planned to install the measure without the program/incentive?
- Did the program/incentive influence your decision to install the measure?
- Would you have installed the same number of measures without the program/incentive?
- Would you have selected the same level of efficiency without the program/incentive?
- When would you have installed the measure without the program/incentive?

Enhanced self-reporting surveys are used to improve the quality of information used to provide NTG ratios derived from self-reporting survey methods. Multiple additional data sources and techniques can be used to get at the rationale for decisions to install energy efficiency measures or to adopt conservation behaviours. Some of these techniques include:

- **In-person surveys.** Surveys conducted in person can improve the quality of the survey results because personal views and information can assist in understanding the influences and motivations that determine the role of CDM programs in participant and non-participant decision-making processes.
- **Project analyses.** These analyses consider specific barriers to energy efficient measure installations and document participants' rationale for proceeding with the measure or project. For example, since most barriers to energy efficiency are related to the costs of installation, conducting a financial payback analysis on a project may reveal the likelihood that the customer would have proceeded with the project in the absence of the program if the project is shown to have a very short payback period. Feasibility studies, engineering reports, and internal memos are examples of other documentation that may provide insights into whether a customer would have proceeded with a project regardless of the program.
- **Non-specific market data collection.** This involves collecting information from other programs to estimate an appropriate NTG ratio or a reasonable range to apply to the program being evaluated.

Key Consideration when Conducting Self-reported Surveys and Enhanced Self-Reported Surveys

Achieving Confidence and Precision Targets

Net-to-gross estimation efforts often have pre-defined confidence and precision targets (for example, 90% confidence at 10% precision) set by the evaluation administrator. The evaluator designs NTG estimation methods and draws an appropriate sample to achieve these confidence and precision targets. However, in some cases, such as a low response rate to surveys or population size constraints, the evaluator is unable to achieve the confidence and precision targets and needs to work with the evaluation administrator to find an alternative approach to produce an acceptable NTG ratio. These risk factors must be considered during evaluation planning to ensure that confidence and precision targets can be set at an achievable level. If an evaluation fails to meet the pre-defined precision and confidence targets, some potential alternatives include:

- Allowing for additional time to collect more survey responses
- Adjusting the confidence and/or precision targets to more achievable thresholds. For example, suppose the evaluator is not able to achieve the 90/10 thresholds. The target can be reduced to 90/15 or 85/15 as long as these parameters meet the needs of the program

- Application of the deemed NTG ratios from previous evaluation cycles for the same (or similar) programs. NTG from previous evaluations can also be referenced as a comparison to validate the current NTG ratio for new program cycles

Market Actors Surveys

When estimating net savings, it is important to consider all program influence points . For example, it is beneficial to collect information from contractors, delivery agents and trade allies involved in delivering the program.

Contractor free-ridership (FR) is generally calculated at the program level and incorporates market-based insights about the program's equipment. These contractor questions focus on market penetration to estimate if the participant would have purchased similar efficiency equipment in the program's absence. The decision to apply participant FR, contractor FR, or a combination of the two is dependent on questions from the participant survey querying how they decided on the efficiency level of their new equipment and the level of input provided by the contractor.

Contractor FR is only recommended for programs with a large and diverse pool of participating contractors who complete projects inside and outside the program. As the number of contractors associated with the program decreases, it can be difficult to provide contractor FR with an acceptable level of certainty. Suppose the program is dependent on a limited number of contractors, or the program only offers a few pieces of equipment. In that case, the contractor's influence becomes intertwined with program performance and less benefit is derived from a contractor-focused FR estimate. In consultation with the evaluation administrator, and depending on the available data and evaluation budget, the evaluator decides whether using contractor surveys is of added value to the program evaluation to collect additional insights.

Econometric Methods

Econometric methods are mathematical models that use statistics and energy and demand data from participants and non-participants to derive accurate net-to-gross ratios. Applying econometric methods are the most costly way of estimating net-to-gross factors and require large numbers of participants and comparable non-participants to make accurate estimates.

Any of the above methods can be combined with participant and non-participant surveys to estimate free-ridership, spillover, and rebound effects. When non-participants are included in the NTG ratio, care must be taken to select a group that is comparable to the participant group.

Agreed on Net-to-Gross Ratios

In some jurisdictions, agreed on net-to-gross ratios may be set by regulatory boards or commissions to be used by program administrators. Agreed on NTG ratios can be used when the cost of conducting more detailed analyses of program net-to-gross factors is a barrier or when the accuracy of the results is not paramount. Agreed on NTG ratios are often periodically updated based on reviews and evaluations of net-to-gross factors. For example, there is a consensus among evaluators and program stakeholders that NTG ratios for most low-income programs are unlikely to be significantly different than one (1.0), particularly when the person making the participation decision is the low-income customer. It is perceived that there is little to no free riders among low-income program participants, in instances where it is assumed that participants would not procure the energy-efficient equipment/service in the absence of the program. Asking a panel of other industry experts, such as contractors, trade allies, and builders to recommend, and arrive at a consensus on an appropriate NTG ratio for a specific program is another example of an agreed-upon NTG ratio approach.

The evaluator selects and implements the approach to determine free-ridership, spillover and rebound effect. Once these factors are determined, the NTG ratio is calculated. The NTG ratio and realization rate are used to calculate the net verified savings. The evaluator compiles the results and reports the net verified savings.

Key Considerations when Calculating the Net-to-Gross Ratios for Midstream and Upstream Programs

Midstream programs pay incentives directly to distributors and include different types of incentive pass-through requirements. As such, the process is seamless to a degree that often makes the customer unaware of the program’s existence. As a result, conventional self-report surveys cannot be used to accurately evaluate midstream programs’ attribution. However, it does not necessarily mean that the program did not influence the end-users’ decision to buy the product only because they did not receive a direct incentive.

In addition to influencing consumers’ decisions, midstream programs aim to change the overall market behaviour. Therefore, evaluation methodologies need to consider all the points of influence on target-market actors (for example, distributor, retailer, customers). Midstream program evaluators need to measure market transformations—the changes in retailer, distributor, and contractor behaviours that can accelerate the adoption of energy-efficient equipment—and account for increased stock. This includes redefining the definition of free riders and spillover to include how the intervention influenced decisions and behaviours of the manufacturer, supplier, distributor, retailer, and consumer.

Given the data available, incentive pass-through requirements and evaluation budget, the evaluator usually works with the evaluation administrator to develop the most appropriate NTG estimation approach. For example, suppose customers’ contact information is not available. In that case, the approach might have to prioritize collecting information about the program’s influence on distributor behaviour with respect to selling energy-efficient equipment.

4.1.5 Step 5: Use, Review and Update Measures and Assumptions List

Measure-level input assumptions are provided by program administrators for inclusion in a program. These input assumptions are included in the Measures and Assumptions List (also referred to as a Technical Reference Manual). The Measures and Assumptions List (MAL) is a database of energy efficiency measures, which are typically substantiated with documented credible results or third-party verification, testing, or certification. The input assumptions that are included in the MAL may be updated as new knowledge, information, or technology emerges. Evaluators use, review, and update the MAL. When reviewing the MAL, the source of the assumptions for measures should be documented in the recommended revisions submitted to the program administrator.

Caution is required when using a MAL or measures assumptions that were developed for use in other jurisdictions, especially where there are different codes, standards, or market conditions. Since the impact evaluation reviews implemented projects, it often provides information that is more relevant and specific.

Data provided in a MAL typically includes:

- Definitions of the baseline and high-efficiency cases or technology
- Energy and demand savings resulting from high-efficiency technology
- Other resource savings (for example, natural gas, water)
- Incremental cost data (for example, the cost differential between baseline equipment and high-efficiency equipment)
- Effective useful life (EUL) of equipment and assumptions about persistence
- End-use load profiles

Free-ridership rates and other NTG adjustment factors are not considered in a MAL. Such factors are a function of program design and operation and need to be determined and regularly accounted for through program evaluation research. Reviewing and updating the MAL includes the following steps, which are described in detail in the remainder of this section:

- Review input assumptions
- Document and report measures reviewed and updated
- Update the MAL

Review Input Assumptions

The evaluator reviews the input assumptions of each measure for accuracy, relevance, and applicability. Where there is insufficient data to update input assumptions or substantiate new measures, the evaluator can gather technical information through various means, including literature reviews, program evaluations, case studies, third party testing, verification, or certification relating to the specific measure being investigated.

Additionally, the review needs to include an hourly (8760) annual load profile created from metered data or a verified operating schedule. If an annual load profile is unavailable, a description of the operating hours during the weekdays and weekends for different seasons can be considered.

Document and Report Measures Reviewed and Updated

The evaluator needs to list the measures covered in the review, the results of the literature search, methods used to identify uncertainties, and methods used to estimate the range of savings specific to the measures in the program.

Update the Measures and Assumptions List

The evaluator submits recommended revisions to the program administrator. The program administrator uses a standardized template, which lists the information to be included when submitting recommendations. An example of this template is included in Appendix E, which is the Measures and Assumptions Substantiation Form used by the IESO. Evaluators are encouraged to use the template, or at least consider it as a guideline upon submission.

4.1.6 Summary

Impact evaluations assess the outcomes of implementing a program, policy, or project. These evaluations are applied to measure a change in energy consumption and/or demand caused by a program or project. The main steps included in impact evaluation of technology-based programs are:

- Step 1: Define the program sample
- Step 2: Collect data
- Step 3: Calculate gross verified savings
- Step 4: Calculate net verified savings
- Step 5: Use, review and update measures and assumptions list

4.2. Impact Evaluation of Behavioural-Based Programs

Behavioural-based energy efficiency programs achieve energy or demand savings by utilizing strategies designed to influence energy and demand consumption behaviours by consumers, operators, installers, lenders and other market actors. Behavioural-based programs consist of a diverse set of programs, which incorporate various elements, including outreach, education, competition, rewards, benchmarking and feedback.

Behavioural-based programs result in changes to habitual behaviours (for example, turning off lights) or occasional behaviours (for example, deciding to request an energy audit). Additionally, these programs may target purchasing behaviour (for example, the purchase of energy-efficient products or services), and are often used in combination with other programs. Other program designs of this class target behaviours related to the selection, installation, and operation of building systems.

Behavioural-based programs are often categorized into three program types:

- Training/capability building programs
- Feedback programs
- Education/awareness programs

Each category differs in behavioural outcomes of interest and the mechanisms used to trigger impacts. As a result, the details of the measurements and approaches that are applied to assess impacts may differ between program types. The different types of behavioural-based programs are discussed in greater detail in the information box below.

Types of Behavioural-Based Programs

Training/Capability Building Programs

Training and capability building programs are designed to induce energy savings by providing training to customers, energy managers, installers and building operators. These programs enable the specified individuals to develop their capabilities in technical aspects of energy efficiency, conservation and demand response. Outcome measures of interest for training and capability building programs include:

- Subscription rates for training courses (for example, how many students are enrolled in training courses).
- Results of standardized tests used to assess the ability of students to recall the material covered in training courses.
- Pass or certification rates for students taking training courses.
- Observation of skills required before and after training.

Feedback Programs

Feedback programs provide information to participants to compel a change in their behaviour. Examples of feedback programs and strategies include:

- Reports indicating normative comparisons of energy usage – periodic (monthly, semi-monthly or quarterly) reports presented to customers comparing their energy use and costs with that of customers who are labelled as neighbours (in these reports, neighbours refers to customers with similar usage and house characteristics , rather than the customers living next door) or similar to the target customer. Home Energy Report and Business Energy Report programs are examples of these types of programs.
- In-home displays – devices that communicate with advanced meters through Wi-Fi or cellular network to display electricity and/or gas consumption in various formats in near real-time.
- Home Energy Management Systems (HEMS) – devices that allow customers to control thermostats, lights and motor loads in their homes and businesses using the internet and smartphone applications.
- Smart thermostats – share similarities with HEMS, though they are designed to analyze customer demands for heating and cooling based on responses to the thermostat setting changes. Optimizing thermostats allow the discovery and scheduling of the optimal operating schedule based on occupancy and observed temperature preferences.
- Bill alerts – messages sent by email, text, or bill inserts, informing consumers that their usage is abnormally high or will exceed a designated value they identified in advance.

- Web-based feedback – information provided to customers on the web about usage and tips for reducing their consumption.

Education/Awareness Programs

Education and awareness programs have been critical in encouraging energy conservation and the efficient consumption of energy for decades. These programs typically involve a highly structured approach for developing and transmitting specific messages to target populations, through well-developed communication strategies. They usually involve:

- Planning – this consists of defining the goals and objectives of the education/awareness effort, assessing resource requirements, obtaining cooperation and resources from organizational leadership and assembling a project team.
- Design and implementation of an informational campaign, including:
  - Identification of specific opinions, perceptions and behaviours that will be influenced by the campaign
  - Formulation of specific messages that will be delivered using surveys, focus groups and other measures to evaluate message content intended to change behaviour
  - Identification of channels used to deliver messages
  - Determination of actions needed to deliver the information campaign
  - Management of the campaign
- Evaluation of program impacts, including estimation of changes in behaviour. This can be completed by comparing survey responses from the target population, before and after exposure to the information campaign and change in energy use when possible.

4.2.1 Research Design

Behavioural-based programs are designed to induce changes in energy consumption related behaviours by individuals and organizations. The main challenge when evaluating behavioural-based programs is that the evaluation is incapable of predicting how much behavioural change or change in energy consumption will occur without testing the effect of the program on the target individuals or organizations. To measure the impact of behavioural-based programs, it is important to consider the outcomes in the absence of the program. By comparing the behaviour in the presence of program interventions against the behaviour in the absence of program interventions, the magnitude of change in the outcome (behaviour or energy consumption) can be determined.

The most robust strategy for assessing the impact of behavioural programs is to implement an experiment in which it is possible to:

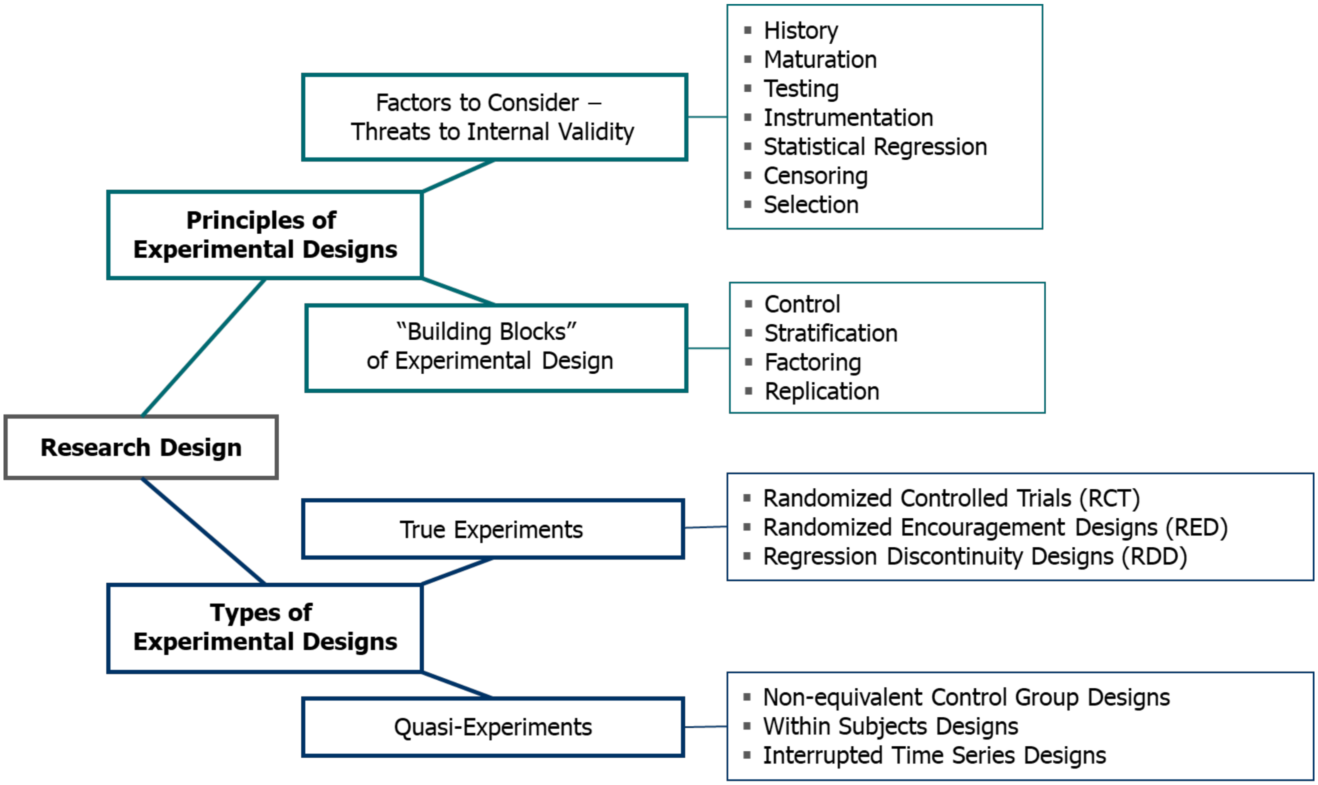
- Ensure the program intervention occurs before the behavioural change.

- Ensure that no other non-program factors may have produced the change in observed behaviour.

True experiments are not always possible, since it is not always possible to control the assignment of observations to treatment and control conditions. In cases where a true experiment cannot be conducted, alternative methods can be used to assess the impact of behavioural programs. These methods are referred to as quasi-experimental techniques and are less rigorous in comparison to true experiments. To reach valid conclusions, these techniques require more skill and talent from evaluators. In cases when it is suitable to do so, true experimental designs involving random assignment of target market actors should be used. When this is not possible, quasi-experimental techniques should be used instead.

Multiple research and evaluation techniques are applicable when evaluating the variety of programs under the behavioural-based program classification. To evaluate this wide range of programs, evaluators should be knowledgeable in the different types of experimental design, and their basic elements. A summary of the most relevant principles and types of experimental design is provided in Figure 4-5. These elements and techniques are described in detail in Appendix D.

Figure 4-5 | Principles and Types of Experimental Design


Research Design
Principles of Experimental Designs
Factors to Consider – Threats to Internal Validity
History
Maturation
Testing
Instrumentation
Statistical Regression
Censoring
Selection
"Building Blocks" of Experimental Design
Control
Stratification
Factoring
Replication
Types of Experimental Designs
True Experiments
Randomized Controlled Trials (RCT)
Randomized Encouragement Designs (RED)
Regression Discontinuity Designs (RDD)
Quasi-Experiments
Non-equivalent Control Group Designs
Within Subjects Designs
Interrupted Time Series Designs

In the context of behavioural-based program evaluations, a true experiment is a situation in which delivery of the behavioural change mechanism can be controlled, such that some entities (for example, customers, businesses, technicians, etc.) experience it while others do not. The impact of the program is then measured as the difference between the observed behaviours when the program is present against when it is not. In an experiment, the parties that experience the program are often referred to as the treatment group, while those who do not experience the program are referred to as the control group. If it is not possible to control the delivery of the behavioural change mechanisms, then it will be necessary to evaluate the program using quasi-experimental techniques. Readers who are unfamiliar with experiments and the different research designs employed in evaluation can refer to Appendix D, which provides a focused overview of the research design principles. Examples of situations when behavioural change mechanisms are not controllable include:

- A program that may have already been implemented or is underway when the evaluator is first introduced to its evaluation. In such a case, the evaluator cannot control the delivery of the program change mechanism.
- A program where the technology is intended to induce behavioural change and the technology is sold over the counter or through the Internet directly to consumers without obtaining customer contact information. It is difficult to control who obtains such devices, and therefore, randomly assigning customers to treatment or control groups is implausible.
- Programs where the delivery of the behavioural change mechanism falls outside the program administrator’s control. For example, educational and awareness campaigns are frequently carried out in emergencies or are required by law or good administrative practices. It may not be appropriate to randomly withhold advanced notice from customers in emergencies, or from those that will experience a rate change that might cause them to incur high bills that could have been prevented with advanced notice. Such situations will challenge the evaluator and project administrator since the robustness of the experimental design that can be implemented depends entirely on the extent of control the evaluator has over the assignment of participants to the evaluation.

Identifying the degree to which the evaluator can control the delivery of the behavioural change mechanism to program recipients is a critical step in developing the research design. If the evaluator can control the delivery of the behavioural change mechanism (providing it to some parties and not others), then an experimental research design can be used in the program evaluation. Based on the determined level of control when defining the situation, the appropriate experimental design can be selected. Table 4-2 provides guidance on selecting the appropriate experimental design based on the level of control.

Table 4-2 | Appropriate Experimental Design Based on Level of Control

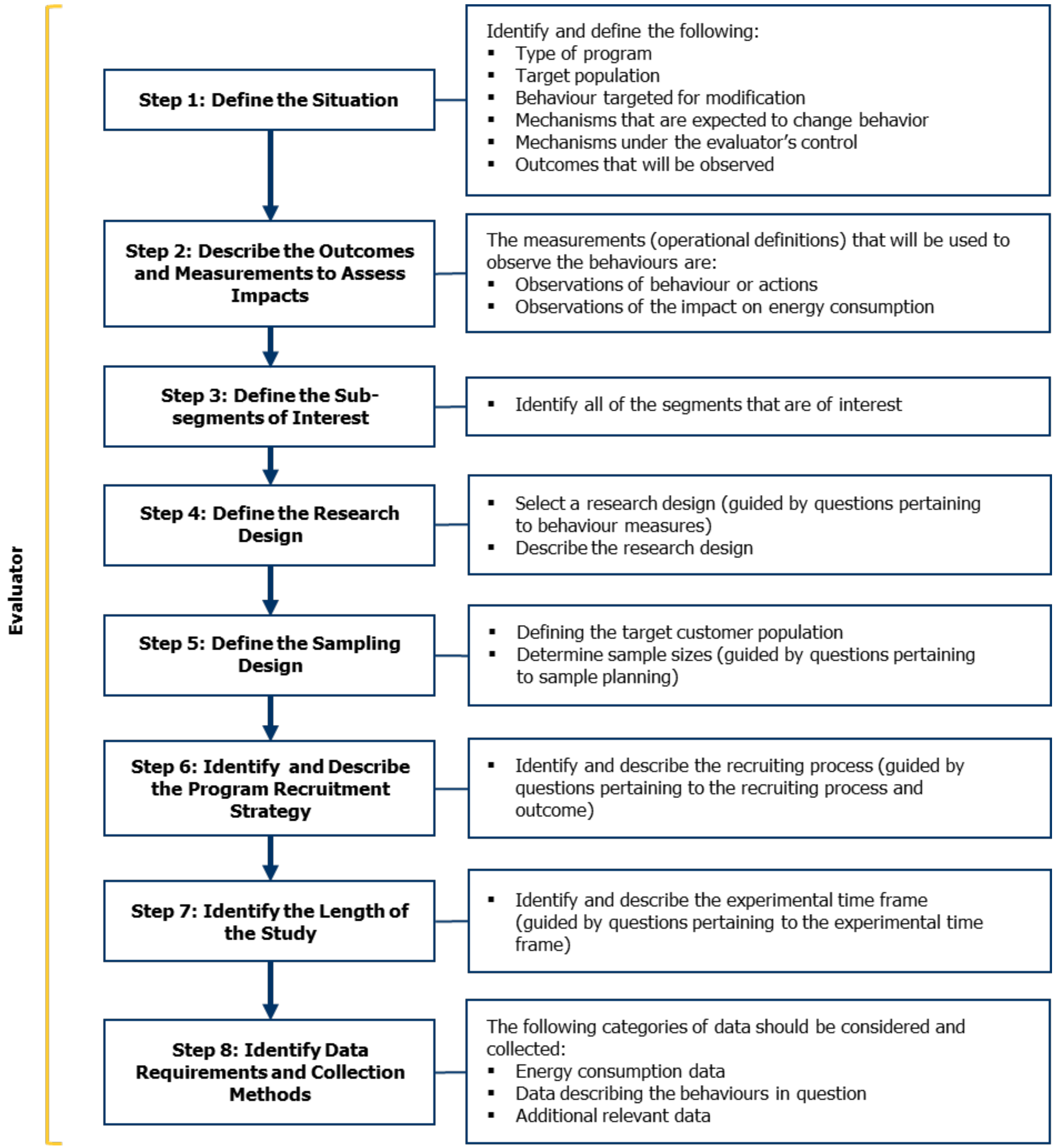
Level of Control	Appropriate Experimental Design[6]
Able to randomize presentation of treatment – mandatory assignment of subjects to evaluation and controlled conditions	Randomized Controlled Trial (RCT)
Able to **deny** treatment to volunteers – mandatory assignment of volunteers to evaluation and controlled conditions	RCT using recruit and deny tactic
Able to **delay** treatment to volunteers – mandatory assignment of volunteers to evaluation and controlled conditions	RCT using recruit and delay tactic
Able to randomly encourage subjects to accept treatment	Randomized Encouragement Design (RED)
Able to assign subjects to treatment based on qualifying interval measurement (for example, income, usage, building size, etc.)	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	Quasi-experimental designs

4.2.2 Savings Estimation

The primary stages to estimate savings of behavioural-based programs are outlined in Figure 4-6 and are described in this section.

[6] Experimental design descriptions are included in Appendix D: Principles and Types of Experimental Design.

Figure 4-6 | Steps to Estimate Savings



Step 1: Define the Situation

Identify and define the following:
- Type of program
- Target population
- Behaviour targeted for modification
- Mechanisms that are expected to change behavior
- Mechanisms under the evaluator’s control
- Outcomes that will be observed

Step 2: Describe the Outcomes and Measurements to Assess Impacts

The measurements (operational definitions) that will be used to observe the behaviours are:
- Observations of behaviour or actions
- Observations of the impact on energy consumption

Step 3: Define the Sub-segments of Interest

- Identify all of the segments that are of interest

Step 4: Define the Research Design

- Select a research design (guided by questions pertaining to behaviour measures)
- Describe the research design

Step 5: Define the Sampling Design

- Defining the target customer population
- Determine sample sizes (guided by questions pertaining to sample planning)

Step 6: Identify and Describe the Program Recruitment Strategy

- Identify and describe the recruiting process (guided by questions pertaining to the recruiting process and outcome)

Step 7: Identify the Length of the Study

- Identify and describe the experimental time frame (guided by questions pertaining to the experimental time frame)

Step 8: Identify Data Requirements and Collection Methods

The following categories of data should be considered and collected:
- Energy consumption data
- Data describing the behaviours in question
- Additional relevant data

Step 1: Define the Situation

The first stage in research design is to develop a clear understanding of the purpose of the evaluation research and the context in which it is conducted. It is generally expected that the evaluator and evaluation administrator will collaborate to effectively answer the research questions developed during the planning stage and to define the research design. The purpose of this step is to analyze the design of the program and identify an appropriate evaluation research design.

When defining the situation, the program needs to be described in sufficient detail to permit discussion of the research design alternatives with the evaluation administrator. The following items are usually included in the description:

- **Type of program.** For example, defining the program as a training support program or a neighbour comparison program. For education and awareness programs, the underlying behavioural scientific theory linking the information that is to be transmitted to the outcome should be described. For instance, the Theory of Reasoned Action diagram describing beliefs that are to be changed, or social reinforcements that are to be given.
- **The target population.** For example, in the case of a training program, identify the market actors that are targeted, such as households, businesses, school teachers, etc.
- **The behaviour(s) that is/are targeted for modification.** For example, thermostat settings, design practices, installation, operations, organizational decisions, etc.
- **The mechanism(s) that is/are expected to change behaviour.** For example, education, feedback, normative comparisons, cognitive dissonance, etc.
- **The information that is to be provided to the target population.** For example, options for reducing energy consumption, best practices for appliance sizing and installation, daily energy consumption, cost of wasting energy, etc.
- **Whether or not the evaluator can control the presentation of the behavioural change mechanism(s).** For example, whether the evaluator can decide who receives the educational material and/or when they can receive it.
- **The program outcomes that will be observed.** For example, adoption of technology, adoption of practices, sales of efficient technology, rebate requests, purchasing behaviour, energy consumption, etc.

Step 2: Describe the Outcomes and Measurements to Assess Impacts

The objective of this stage is to describe the expected behavioural outcomes from the program and the necessary measurements to assess those outcomes. To begin describing the outcomes, the evaluator can consult with the program design team, who would have defined the program objectives and associated outcomes. Additionally, some assistance from the program administrative staff will be required to identify existing data sources and data that need to be developed during program implementation to support evaluation. Several basic outcomes need to be described, as summarized in Table 4-3.

Table 4-3 | Basic Outcomes

Type of Program	Basic Outcomes
Training/capability building	• Change in participant behaviour • Adoption of best practices communicated in training • Success in marketing • Energy savings
Feedback	• Receipt/open rate of feedback • Awareness of feedback • Acceptance of feedback • Change in equipment acquisition behaviour • Change in energy consumption related behaviour • Change in other behaviours (for example, knowledge, opinions, and attitudes)
Education/awareness	• Change in beliefs and opinions related to energy consumption • Change in beliefs about what is normatively appropriate energy consumption related behaviour • Change in attitudes about energy consumption, comfort, convenience, etc. • Awareness of educational and informative messages • Awareness of channels through which messages are conveyed • Household/business energy savings

There are two categories of measurements when evaluating behavioural-based programs:

- Observations of behaviour or actions that are taken in response to the program
- Observations of the impacts of the program on energy consumption

The evaluator is to produce a comprehensive description of the outcomes that will be measured in the evaluation. All the different types of physical measurements that must be taken to assess the impacts of the behavioural program should be identified. Table 4-4 provides a summary of typical measurements that might be included.

Table 4-4 | Examples of Possible Outcome Measurements

Type of Program	Measurements
Training / capability building	• Measurements collected from tracking systems that record the progress of marketing efforts indicating who received program offers, the method of offer delivery, the volume of offers sent, the content received and response rates to these offers • Records of participation in rebate and other programs that may identify actions taken by participants in response to training • Measurements from surveys of consumers or other market actors taken before and after exposure to training • Measurements from tests given to trainees before and after exposure to training • Measurement of energy consumption before, during, and after the program for treatment and control groups
Feedback	• Measurements collected from tracking systems that record the progress of marketing efforts indicating who received program offers, the method of offer delivery, the volume of offers sent, the content received and response rates to these offers • Records of participation in rebate and other programs that may identify actions taken by participants in response to the program • When enabling devices, such as HEMS, are used – measurements of device activation rates and reasons for activation failure • Measurements from surveys of consumers or other market actors taken before and after exposure to the program • Measurements of drop-out rates and reasons for departing the program • Measurement of energy consumption before, during, and after the program
Education / awareness	• Measurements from surveys of consumers or other market actors taken before and after exposure to educational programs • Measurements from tracking systems recording the details of the educational program, including when populations were exposed to educational materials, what channels the messages were delivered through, the volume of sent messages and what content was used

Type of Program	Measurements
	- Records of response to program's measurement of energy consumption before, during, and after the program implementation for treatment and control groups

The purpose of defining the outcomes and measurements is to describe all behavioural and energy savings outcomes that are expected from a program, as well as their respective measurements in immense detail. Table 4-5, Table 4-6 and Table 4-7 provide examples of cataloging expected outcomes and their corresponding measurements for the different types of behavioural-based programs.

Table 4-5 | Examples of Program Outputs and Corresponding Measurements for Training/Capacity Building Programs

Program Outcome	Measurements
Example: HVAC Installation Contractor Training Program	
• Improved performance in carrying out best practices when calculating system size requirements and applying various technical and non-technical practices associated with installation	Behavioural measures • Comparison of work before and after training of participants in the treatment and control groups. • Written test to capture participant knowledge before and after training • Comparison of knowledge and opinions (as measured by a test) of treatment and control groups
• Energy savings resulting from an improved performance due to training	Savings measures • Comparison of average SEER of systems installed by treatment and control groups before and after training • The estimated annual, monthly, and hourly energy savings based on an average SEER difference • The estimated difference in peak kW, if any, per hour

Example: Segment Support Programs (for example, energy efficiency solutions support to municipal governments)	
• Provide technical assistance focused on increasing the uptake of energy efficiency investments in different market segments such as municipal governments, hospitals, retail shopping complexes, and water and wastewater treatment facilities.	Behavioural measures • The rate of acceptance of assistance in treatment groups • Expressed interest in assistance for control groups • Comparison of rate of adoption of various energy efficiency solutions (for example, energy efficiency plans, financial analysis, management presentations, measures adopted) for treatment and control groups
• Energy savings resulting from energy efficiency solutions	Savings measures • Comparison of annual energy consumption for treatment and control groups before and after the program

Table 4-6 | Examples of Program Outputs and Corresponding Measurements for Feedback Programs

Program Outcome	**Measurements**
Example: Normative Comparison Programs	
• Customer acceptance • Energy-related knowledge, skills and opinions • Appliance acquisition behaviour • Energy consumption related behaviour	Behavioural measures • Customer subscription rate (for opt-in delivery) and opt-out rate (for default delivery) from tracking systems • Survey responses of treatment and control groups' knowledge, skills and opinions, reported appliance acquisition behaviour and reported energy consumption related behaviour before and after the program
• Energy and demand savings resulting from providing normative comparisons	Savings measures • Observed differences in monthly or annual energy consumption and demand for treatment and control groups before and after program implementation from billing systems

Example: Other Feedback Programs (for example, smart thermostats)	
• Customer acceptance • Device commissioning • Device utilization • Energy-related knowledge, skills and opinions • Appliance acquisition behaviour • Energy consumption related behaviour • Usability • Persistence	Behavioural measures • Customer acceptance rate from tracking systems • Device commissioning rate from tracking systems • Interviews/focus groups with customer service agents • Information collected through interviews with customers regarding commissioning problems • Survey responses of treatment groups regarding satisfaction with the acquisition/installation process • Survey responses of treatment and control groups' knowledge, skills and opinions, reported appliance acquisition behaviour and reported energy consumption related behaviour before and after program implementation • Information collected through focus groups with treatment groups regarding usability and persistence
• Energy savings resulting from providing technology	Savings measures • Observed differences in monthly or annual energy consumption and demand for treatment and control groups before and after the program from billing systems
Example: Website	
• Customer acceptance • Website access • Website utilization • Opinions about website • Energy-related knowledge, skills and opinions • Energy consumption related behaviour • Customer opinions about website usability • Persistence of the above outcomes over time	Behavioural measures • Website access from tracking systems • Page views from tracking systems • Return rate from tracking systems • Focus groups with customers regarding usability • Survey responses of treatment groups regarding satisfaction with website content and performance • Survey responses of treatment and control groups' knowledge, skills and opinions, reported appliance acquisition behaviour and reported energy consumption related behaviour before and after the program

Table 4-7 | Examples of Program Outputs and Corresponding Measurements for Education/Awareness Programs

Program Outcome	Measurements
Example: Beliefs about own energy consumption	
• Beliefs and opinions related to energy consumption • Attitudes about energy consumption, comfort, convenience, etc. • Beliefs about whether a participant's energy consumption related behaviour is socially normal • Awareness of education opportunities and other related messages • Awareness of channels through which messages are delivered	Behavioural measures • Survey responses about beliefs held by participants about their energy consumption before and after exposure to the educational program for treatment and control groups
Example: Beliefs about normative energy consumption	
• Beliefs about what is appropriate energy consumption related behaviour • Perceptions of energy consumption related behaviours of others	Behavioural measures • Participant's survey responses regarding their beliefs on energy consumption related behaviour and opinions are normatively correct before and after exposure to the educational program for treatment and control groups
Example: Reported energy consumption related behaviour	
• Reported intention to take actions to reduce energy consumption • Reported appliance purchases • Reported thermostat settings • Reported use of lighting and other appliances	Behavioural measures • Survey responses regarding reported energy consumption related behaviours before and after exposure to the educational program for treatment and control groups

	Savings measures
• Energy savings resulting from providing feedback	• Observed differences in monthly or annual energy consumption and demand for treatment and control groups before and after program implementation from billing systems

Step 3: Define the Sub-segments of Interest

Behavioural-based programs frequently target various audiences. For example, trades or disciplines in the case of a training program, or customers with specific heating or cooling devices in feedback programs. Should there be a necessity to understand how a program influences different market segments, it is important to recognize these different segments during the evaluation research design process. The evaluator collaborates with the program staff to identify such relevant market segments. Examples of market segments include organization types, business types, job types, household types, usage categories and a wide variety of other potentially important classification variables.

In cases where programs are evaluated through comparison of customer behaviour in both treatment and control groups, segments should be limited to only those which can be observed prior to the assignment of participants to groups. For example, before conducting an evaluation, it is possible to determine whether an employee in an HVAC contracting firm is a sales agent or an installer. This segmentation may be valuable, as evidence shows these two roles have different responsibilities for new equipment installation.

Step 4: Define the Research Design

After the outcomes and measurements have been described, and the market sub-segments have been detailed, the evaluator then begins defining the research design. The evaluation research design can be defined by answering the following questions:

- How long is the pre-treatment period, which is the period of data collection required prior to program onset?
- Will pre-treatment data, which are measurements of interest from the period prior to program onset, be available?
- Does the appropriate data already exist for all targeted participants, or do measurements need to be taken to gather pre-treatment data?
- Is a control group(s) required for the experiment?
- Is it possible to randomly assign individuals from the target population to treatment and control groups?

When answering these questions, evaluators can refer to the framework of principles and types of experimental research design in Section 4.2.1 as a guide to selecting the experimental design that best supports the treatments, objectives, and practical realities surrounding the program. Through the guidance of the framework described in Section 4.2.1, the evaluation research design that will be used during the evaluation can be described. The evaluator is responsible for writing a detailed description of the selected research design. The description should include the following:

- An explanation of the type of research design selected (for example, RCT, RED, regression discontinuity, non-equivalent control groups, within-subjects, etc.) and the rationale for selecting the research design.
- A discussion of any operational challenges that may compromise the validity of the research design and measures that will be adopted to overcome those challenges.
- A description of the treatment groups, control groups and any segmentation (for example, by trade or industry group).
- If a random assignment cannot be achieved, the description needs to address how suitable comparison groups will be identified. Additionally, it should also address how the design provides a comparison that allows an assessment of the impact of the program on behaviour and energy consumption.

In the case of true experiments, the design can be presented in a table where the measurements are described on the column headings and segments are described on the rows.

Step 5: Define the Sampling Design

Once the appropriate experimental design has been selected, a sampling plan can be developed. The sampling design produces the sampling plan, which will guide the sampling activities. Certain critical items must be addressed in the sampling design and sampling plan, including:

- Are the results of the research intended to be extrapolated beyond the population included in the sample (for example, the results and observations from evaluating the sample of households who received a feedback technology will be extrapolated to all households eligible to receive the technology in the region served by the program)?
- Are there sub-populations (strata) for which precise measurements are required (for example, usage categories or other segments)?
- What is the absolute minimum level of change in the effect(s) of interest(s) that is meaningful from a planning perspective (for example, a 5% reduction in electricity consumption)?
- How much sampling error is permissible (for example, ±1%)?
- How much statistical confidence is required for planning purposes (for example, 90%)?

The answers to the above questions will influence the design of the samples used in the evaluation. The answers to these questions are informed by the policy considerations of the program administrator, who will use the information and results to make decisions. Once the program administrator has developed the requirements, the evaluator can then determine the sample composition and sizes needed to meet the requirements. Appendix B provides additional guidance on defining the precision and level of confidence for the sampling plan design.

The evaluator needs to define and describe the sampling plan and can use the questions provided in the information box below as guidance. The evaluator can also refer to Appendix B, where more detail is provided regarding sampling elements, such as stratification.

Questions to Guide the Development of a Sampling Plan

- Are the measurements from the experiment extrapolated to a broader population?
  - If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.
  - If no, describe the list of entities from which the sampling will be obtained.
- Are impact estimates required for the sub-populations of interest?
  - If yes, describe the sub-populations for which impact estimates are desired.
- What is the minimum threshold of difference (for example, change in energy consumption) that must be detected by the experiment?
- What is the acceptable margin of sampling error (for example, ±1%, 5% or 15%) for the impact estimates?
- What is the acceptable level of statistical confidence (for example, 90%, 95%, 99%) for impact estimates?
- Are participants randomly assigned to treatment and control conditions or varying levels of factors under study?
  - If yes, how will the random selection be addressed in the analysis and sample weighting?
  - If no, are subjects expected to enroll themselves in the treatment condition?
- If subjects will not be randomly assigned to treatment and control conditions or varying levels of factors under study:
  - Describe the process that will be used to select customers for the treatment group(s).
  - Describe the process that will be used to select customers for the control group and explain why this is the best available alternative for creating a non-equivalent control group.
- If no control group is used, the calculation of the change in the outcome variables of interest should be explained.

Step 6: Identify and Describe the Program Recruitment Strategy

There exist two types of recruitment strategies used in behavioural-based programs designed to influence energy-related behaviours. They are commonly referred to as opt-in and opt-out designs, based on the role of the target participant in deciding whether to receive treatment. The type of recruitment strategy is important, as it limits the circumstances in which the results of any evaluation research activity can be generalized. In an opt-in recruitment strategy, customers are offered the opportunity to participate in the program and are then assigned to either treatment or to control groups. Those who receive treatment are considered volunteers. In opt-out recruiting designs, treatment is administered to subjects unless they specifically withdraw from the experimental condition (usually after exposure to it).

When customers are recruited into a study voluntarily, the voluntary nature of participation can influence the outcome of the evaluation. This is particularly problematic when studying behavioural changes, as motivation is a strong behavioural determinant. It is possible to adjust the research design to overcome the problems that occur when using opt-in recruiting strategies (discussed in detail in Appendix D).

Table 4-8 provides questions and examples to guide the description of the recruitment process and the respective outcomes.

Table 4-8 | Recruiting Strategy Questions

Questions	Examples Training / Capability Programs	Examples Feedback Programs
What are the eligibility criteria for the program?	Participants must be actively employed as HVAC sales agents or installation technicians with more than 5 years of industry experience.	Households in single-family dwellings are located in climate zones X and Y.
What is the estimated number of eligible participants in the region under study?	10,000 total (sub-groups unknown)	1,000,000
How were participants recruited to the program?	Flyers were mailed to all licensed HVAC contractors in the region.	Flyers were mailed to all eligible households.
Were participants randomly assigned to treatment and control conditions?	Yes, due to limited availability, about half of interested participants were randomly admitted into the program in the first year and the remainder was asked to wait for training until the following year.	Yes, participants were randomly allocated to treatment and control groups when they signed up to participate.

Questions	Examples Training / Capability Programs	Examples Feedback Programs
If there were sampling strata, what is the number of participants recruited into each strum and group?	100 HVAC sales agents and 100 HVAC installation technicians in the treatment group, as well as 100 HVAC sales agents and 100 HVAC installation technicians in the control group.	500 customers were sampled in each of the 4 sampling strata.

Oftentimes, different recruiting processes are tested during the evaluation. An objective of such tests is the assessment of recruitment strategy alternatives and identification of the topmost cost-effective strategy to consider during program design. If various recruiting strategies are tested as part of the evaluation, the following descriptions can be used to define the recruitment strategy:

- Describe each of the recruiting options that are tested during the evaluation. This includes how potential participants are identified and contacted, what information they are receiving, are incentives offered and any other similar and pertinent information.
- Describe the research design that is being used to assess the effectiveness of alternative recruiting strategies. This includes the type of employed experimental design (For example, RCT and RED), how customers are sampled, and how many potential participants are selected for each recruiting test.
- Describe how the results of the recruiting strategy tests will be statistically analyzed.

Step 7: Identify the Length of the Study

When evaluating a behavioural intervention, it is important to consider various factors. These factors include the expected time required to perform the entirety of the intervention, the expected onset time for the effect of the program and its expected persistence after initial treatment. These considerations will determine the time required to assess the impact of the program and thereby determine the time for which the situation must be observed.

The evaluator should answer the following questions pertaining to the experiment time frame:

- Is it possible to observe the impacts of the program for a minimum of two years?
  - If it is not possible, how will the persistence of the effect be determined?
- Does pre-treatment data for the relevant variables already exist, or must time be allowed to obtain pre-treatment data?
  - If pre-treatment data does not already exist, how long must the pre-treatment period be to support the experimental objectives?
  - If pre-treatment data does not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to the sampling design will be required to employ a post-test-only design?

- What is the expected time required for participants to receive and understand the information provided to them?
- What is the expected amount of time necessary for participants to implement behavioural changes in response to the information provided?
- What is the minimum duration necessary to sustain the effect of the treatment to justify the program expenditure?
- If the duration of the experiment is shorter than the expected persistence of the program's influence, how will the determination be made as to whether the effect of the program persists long enough to be cost-effective?
- How much time is required between the completion and the approval of the research plan and when treatments are in place for experimental participants?
- How much time is required between obtaining the final data from the experimental observations and completing the analysis?

Step 8: Identify Data Requirements and Collection Methods

Lastly, to estimate savings, the evaluator needs to identify and describe the data requirements and collection methods. The data can be divided into three categories; energy consumption data, behavioural data and additional data. The following items can be addressed to identify and describe the data requirements and collection efforts for the three data categories:

- Description of variable
- Frequency of measurement
- Method of measurement
- Issues and solutions

Table 4-9 provides an example of a description of the data requirements and collection efforts for the energy consumption and behavioural data categories. This example is for a Home Energy Report (HER) program. The example is not a comprehensive list of all the variables and only includes a sample. Data relevant to the additional data category would include, for example, demographic data, such as age, gender, education and household income.

Table 4-9 | Example Description of Data Requirement and Collection Efforts for Energy Consumption Data, Behavioural Data and Additional Data

Description of Variable	Method and Frequency
Energy Consumption Data	
Electricity consumption before, during and after treatment.	Monthly electricity consumption measurement for the 12 months preceding the treatment, during the treatment, and 12 months following the completion of the treatment.
Behavioural Data: Reaction to HER content (only for treatment customers)	
Recall of HER comparison. Does the customer recall whether they are a high or low electricity user?	For a multi-year program, one survey at the end of each program year.
Acceptance of HER comparison. Does the customer believe the HER comparison?	
Credibility of HER comparison. Does the customer think the comparison is credible?	
Customer satisfaction with HER. Customer reports whether they like the HER.	
Customer actions resulting from HER. Customer reports whether they have made any changes. If changes were made, customer reports what changes were made.	
Savings resulting from HER. Customer reports how much they have saved.	
Recommending HER to others. Customer reports whether they have discussed report with others (such as family and friends).	

The descriptive items listed above need to be completed for all the measurements that will be conducted during the duration of the study. The description of the variable should include a sufficiently detailed definition of the variable for seamless knowledge transfer to third parties. The frequency of when the measurement will be taken should also be outlined in the description.

For **energy consumption**, the measurement of the variable might be once or twice (as in the case of SEER measurements), or it might be monthly, hourly, or, in the case of electricity consumption, momentarily. The method of measurement should describe the data collection process in sufficient detail. If utility billing data will be used, it is sufficient to describe the source and the intervals at which the data will be collected. If end-use metering or other measurement procedures are employed, the technology, as well as installation and data collection protocols, should be described.

Behavioural data is information describing the impact of the program on target behaviours. Examples of behavioural data appropriate for training programs include knowledge tests, skills/abilities before and after training, and observations of actions taken by participants before and after training (for example, installations or operating conditions). Behavioural data for feedback programs include the most up-to-date-reported history of appliance purchases, an inventory of energy-saving actions implemented since the start of the program, and perceptions and opinions about energy consumption.

Additional data includes a variety of data that might be useful when evaluating the impacts of the training or segment support programs. Such data may include weather data, data describing the response of the market to the program offering, and market data describing the conditions in the market before, during and after behavioural intervention.

4.2.3 Summary

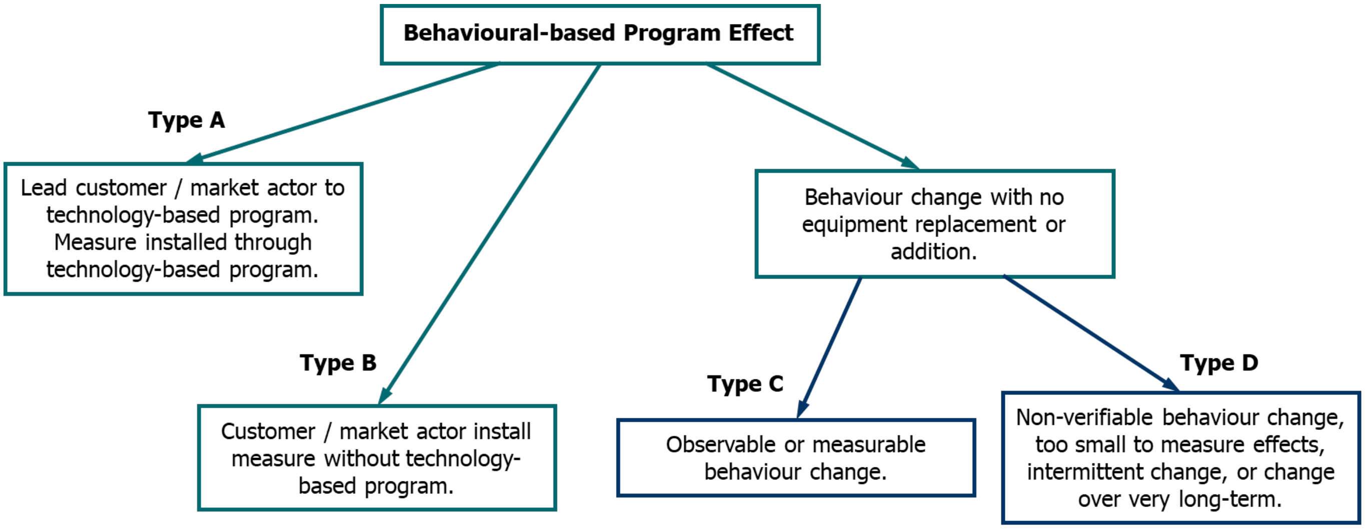
Behavioural-based energy efficiency programs achieve energy or demand savings by utilizing strategies designed to influence energy and demand consumption behaviours by consumers, operators, installers, lenders and other market actors. Behavioural-based programs consist of a diverse set of programs, which incorporate various elements, including outreach, education, competition, rewards, benchmarking and feedback. The impact evamuation of behavioural-based programs include the following steps:

- Research design
- Savings estimation:
  - Step 1: Define the situation
  - Step 2: Describe the outcomes and measurements to assess impacts
  - Step 3: Define the sub-segments of interest
  - Step 4: Define the research design
  - Step 5: Define the sampling design
  - Step 6: Identify and describe the program recruitment strategy
  - Step 7: Identify the length of the study
  - Step 8: Identify data requirements and collection methods

4.2.4 Multi-Layered Approaches

Multi-layered approaches refer to instances in which a participant implements measures and receives incentives in a specific program due to another program’s influence. These savings are not considered a spillover of the program providing the incentive unless program rules do not permit customers to participate in both programs. Information and education programs are typically designed to indirectly acquire energy or peak savings through changes in participants’ behaviour after exposure to the information. For example, the installation of energy efficiency (EE) equipment after completing a training or due to an embedded energy manager. This is mainly applicable when the implementation of an EE measure by a participant can be attributed to both a behavioural-based and a technology-based program at the same ti me. To avoid double-counting and ensure the appropriate allocation of savings, it is necessary to define the interactions between behavioural-based and technology-based programs within the same portfolio of programs and apply the appropriate evaluation design. The type of interactions between behavioural-based and technology-based programs are illustrated in Figure 4-7[7].

Figure 4-7 | The Type of Interactions Between Behavioural-based and Technology-based Programs


Behavioural-based Program Effect
Type A
Lead customer / market actor to technology-based program. Measure installed through technology-based program.
Type B
Customer / market actor install measure without technology-based program.
Behaviour change with no equipment replacement or addition.
Type C
Observable or measurable behaviour change.
Type D
Non-verifiable behaviour change, too small to measure effects, intermittent change, or change over very long-term.

The type of interactions and appropriate evaluation designs include:

[7] California Public Utilities Commission (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals.*

- **Type A:** Type A interaction occurs when a behavioural-based program (through information, education, training, advertising or other nonmonetary incentive efforts) leads customers or market actors to other technology-based programs in the portfolio. The behavioural-based program can be assigned an indirect impact evaluation to determine the program's impact on the portfolio and provide input for its evaluation process. A standard rigour level assignment requires that an impact evaluation be conducted and linked to energy and demand savings estimates. For Type A interactions, the technology-based program is subject to an impact evaluation to estimate the energy and demand savings.

  Evaluation design: The evaluation design to verify actions is the most straightforward for Type A interactions relative to the other types. Verification of behavioural program participation is sufficient given that technology-based programs are conducting their own verification and impact evaluation. The savings are attributed to the technology-based programs and the same savings are attributed to the behavioural-based program as indirect savings. The indirect savings would not be added to the portfolio level savings unless a method is used and approved by the program administrator (or regulatory entity) to ensure that these savings are not double-counted with those attributed to other programs.

- **Type B:** Type B interactions include behavioural-based programs that directly influence customer behaviour to purchase high-efficiency replacement equipment or add equipment that can save energy. A direct impact evaluation is assigned to these programs.

  Evaluation design: The evaluation design for Type B requires identifying affected customers and would have to be part of the evaluation design and the evaluation plan. The evaluation plan needs to propose the research design to identify affected customers and be approved within the evaluation planning review process. The impact evaluation estimates the energy and demand savings, which are directly attributable to the program effort being evaluated.

- **Type C:** Type C interactions include behavioural-based program-induced changes that can be observed or measured but are not tied to equipment replacement or the addition of new equipment. This could include behavioural changes, including establishing corporate or business policies regarding the adoption of energy efficiency practices and adjusting operating and maintenance schedules.

  Evaluation design: The evaluation research design needed to accomplish an enhanced rigour indirect impact evaluation for Type C is more challenging relative to the other types. A Type C evaluation plan needs to be presented in sufficient detail for its logic and potential reliability to be reviewed as part of the evaluation planning review process. Examples of Type C activities include examining business policy manuals, reviewing business programs created due to education efforts, and testing subsequent employee knowledge and reported actions.

- **Type D:** Type D interactions include behavioural-based changes that are too small, long-term or intermittent to be cost-efficiently verified through observation, field-testing or surveying with enough reliability to measure any energy and demand impacts.

  Evaluation design: For Type D, the basic rigour indirect impact evaluation needs to be applied to demonstrate the program has carried out specific activities designed to produce a behavioural change.

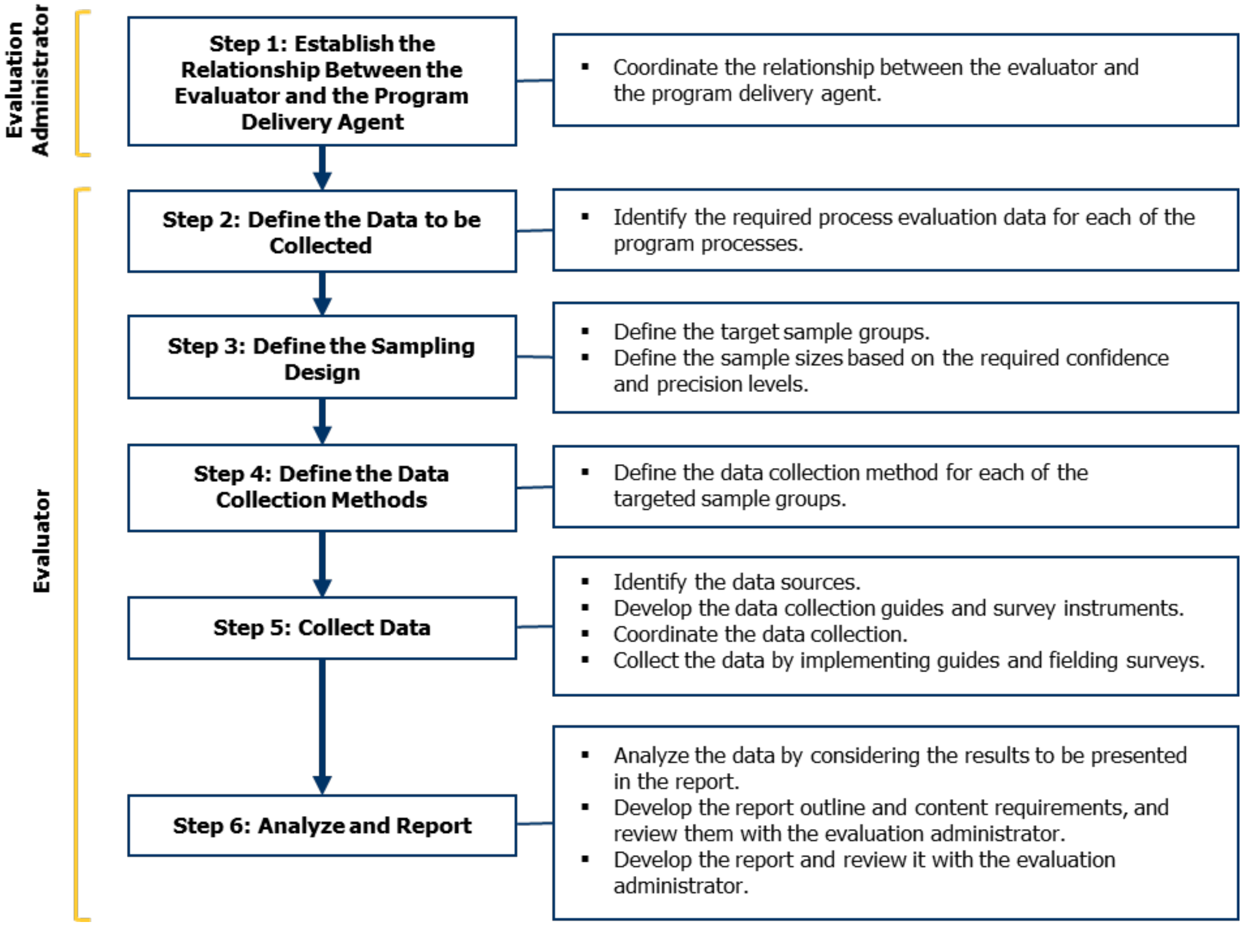
4.3. Process Evaluation

A process evaluation is a systematic assessment of a program’s design, development, delivery, and administration. Process evaluations provide practical advice through quantitative and qualitative insights to enhance a program’s design, administrative processes, and the program’s delivery service. As a result, process evaluations review the effectiveness of the services provided by the program and document the resulting operational outputs compared to the program’s objectives. More specifically, process evaluations gauge the effectiveness and appropriateness of the program’s lifecycle, as follows:

- **Program Design:** The relationship between key program elements, achievability of program objectives, and resource allocation.
- **Program Development:** The protocols and procedures that form the basic offer for implementation, training and technical assistance provided to program delivery agents.
- **Program Delivery:** The services provided by program delivery agents and the processes utilized in the field to deliver the program. The program delivery services include the services provided by third-party program delivery agents, technical reviewers, distributors, contractors, and trade allies. Program delivery also takes into account the systems used to track and monitor program outputs and the program’s expenditures over the assessment time frame. For example; the quality of the measure installation and the levels of participant satisfaction maintained throughout the program.
- **Program Administration:** The controls established for program delivery including program marketing and outreach, the procurement processes for program goods and services, and the mechanisms in place to evolve the program.

The steps involved in conducting a process evaluation are summarized in Figure 4-8 and described in more detail in the remainder of this section.

Figure 4-8 | Process Evaluation Steps


Evaluation Administrator
Evaluator
Step 1: Establish the Relationship Between the Evaluator and the Program Delivery Agent
Coordinate the relationship between the evaluator and the program delivery agent.
Step 2: Define the Data to be Collected
Identify the required process evaluation data for each of the program processes.
Step 3: Define the Sampling Design
Define the target sample groups.
Define the sample sizes based on the required confidence and precision levels.
Step 4: Define the Data Collection Methods
Define the data collection method for each of the targeted sample groups.
Step 5: Collect Data
Identify the data sources.
Develop the data collection guides and survey instruments.
Coordinate the data collection.
Collect the data by implementing guides and fielding surveys.
Step 6: Analyze and Report
Analyze the data by considering the results to be presented in the report.
Develop the report outline and content requirements, and review them with the evaluation administrator.
Develop the report and review it with the evaluation administrator.

4.3.1 Step 1: Establish Relationship between Evaluator and Program Delivery Agent

Process evaluations rely on the collaboration between program delivery agents and the program evaluator. After retaining the evaluator and the program delivery agent, the evaluation administrator coordinates the relationship between the two. It is beneficial to establish the relationship between the evaluator and the program delivery agent early in the program development lifecycle. This assures that necessary elements of the program evaluation are included, and the process evaluation is a joint effort to improve program outcomes.

4.3.2 Step 2: Define the Data to be Collected

As part of the evaluation planning stage, the evaluation administrator defined the critical research questions to be answered. Guided by the research questions and the evaluation scope, the evaluator then needs to define the data to be collected. The data to be collected can be identified by the processes that are applicable to the research questions and should align with the scope of the study. The process evaluation focuses on observable behaviours, materials leveraged, and how the program's materials were received by participants. Each process chosen for the evaluation needs to be thoroughly analyzed, however, not all processes can be included in the process evaluation due to resource limitations. The evaluated processes need to be readily distinguishable from each other. For each of the program processes, the evaluator identifies the required process evaluation data that needs to be collected. For example, the evaluation of a direct install program may include the research question "how effective is the program delivery mechanism?" To address this question, the evaluator identifies the relevant processes. These processes may include the outreach by the contractor, the installation of the technology, and the customer participation process. For each of these processes, the evaluator identifies the relevant process evaluation data that needs to be collected. For example, to answer the questions related to the outreach process carried out by the delivery agent, relevant data may include the number of customers contacted, the number of customers who participated, or demographic data (such as the number of participants per rural versus urban, or by subsector).

4.3.3 Step 3: Define the Sampling Design

When defining the sampling design, the main objective is obtaining suitable data for analysis from the appropriate population to produce results that align with the goals and objectives of the evaluation. There also needs to be a balance between other evaluations that may occur simultaneously to minimize contacting participants multiple times and prevent survey fatigue. Disregarding this factor can result in unengaged or disinterested respondents when completing a survey, and their responses will lead to low quality data.

When considering sample groups for process evaluations, it is usually beneficial to target program management and delivery staff that were involved in the program during the evaluation time frame as they would have comprehensive knowledge of the operational aspects of the program. Additionally, having a sample group of program participants may provide insight when evaluating a program's outputs and its level of effectiveness, as they are the recipients of the program outcomes.

The specific steps involved in designing the sampling plan include:

- **Define the target sample groups.** Considering the program processes to be assessed within the scope of the evaluation, the evaluator needs to define the targeted sample group (such as representatives from program administration, program delivery agents, contractors/trade allies, and participants) for data collection.
- **Define the sample sizes.** For each of the targeted sample groups, defining the sample sizes is based on the required confidence and precision levels. Section 4.1.1 provides a detailed discussion relating to defining the sample sizes. Program managers and program delivery agents often consist of teams who are led by one or two team members. These leading team members usually have extensive knowledge relating to the program and can provide data or information to answer the evaluator's questions. Therefore, it is usually sufficient to interview only these lead team members as representatives of the program managers and program delivery agents. Defining the sample sizes is often developed along with defining the data collection methods, which are described in the section below.

4.3.4 Step 4: Define the Data Collection Methods

Metrics for quantitative assessments are often tracked by program administrators and program delivery agents within tracking systems and management reports. In contrast, qualitative data is observed or collected through survey/interview techniques. During a process evaluation, there can be a mix of data collection methods, which mainly includes interviews and surveys. However, there are some additional methods listed below that may also be considered:

- **Reviewing field notes:** Field notes are brief records kept by program participants or delivery agents, typically recorded on templates or forms. These templates may be part of the program delivery model or developed by the evaluator. Examples of field notes include activity logs, diaries, inspection notes, and receipts.
- **Conducting ethnographic analyses:** Ethnographic analyses are a method of research that necessitates the evaluator's direct observation of a program activity. This may include a "ride-along", which is the process of the evaluator accompanying the service provider in the field, interacting directly with program participants, and asking program staff questions regarding their activity.
- **Conducting a Delphi analysis:** Delphi analyses involve organizing a panel of experts to explore a process or an issue. The objective is to build an agreed-upon opinion around the event or to forecast probable outcomes.
- **Conducting focus groups:** Focus groups are small group discussions, generally with the program participants and targeted market actors, to learn about their collective experience with a product or service offered by the program.

The evaluator needs to define the data collection methods for each of the targeted sample groups. When determining how to collect the data, the evaluator has to balance the cost and rigor against potential biases. The population and sample sizes are considered to establish this balance. Telephone surveys can be efficient and cost-effective for a program with, for example, five to ten participants. Conversely, for a program with, for example, 5,000 participants, conducting a census web-survey (where all participants are invited to take the survey) is more efficient and cost-effective.

4.3.5 Step 5: Collect Data

The data collection process is dependent on key tools, including interview guides and survey instruments, to collect process evaluation data. Guides are usually developed for data collection methods where the evaluator captures the responses, for example telephone interviews and facilitated workshops. Survey instruments are developed where the responses are directly captured by the data collection tool, such as web surveys and hand-out surveys. Due to the importance of interviews and web-surveys, best practices for developing interview guides and web-survey instruments are provided in the information box below.

Best Practices in Developing Interview Guides and Web-Survey Instruments

Interview guides are best developed and used in complex situations where motivations to participate in the program and behaviour influencers are likely to be multi-faceted. As a result, questions in an interview guide act as conversation prompts to guide the interview and collect the appropriate qualitative data. When developing an interview guide, it is best to keep questions open-ended, allowing the participant to elaborate and provide as much detail as possible. Interview questions often start with "why" or "how." It should also be considered whether the sample group possesses sufficient knowledge regarding the question topic to provide a relevant response and if the questions address the original research objectives. Additionally, it is best to document questions and ensure their logical flow. To maintain the interviewee's interest and ensure the interview is completed, it is recommended to keep the survey concise and limit the number of questions to focus on the key priorities identified earlier.

In contrast, a **survey instrument** is best developed and used to address a need for data collection from a large sample group, where additional detailed information regarding the program is required. These surveys can be conducted by mail, e-mail, online, or through comment cards. Surveys are usually less open-ended and respondents provide answers to predefined questions. When developing a survey instrument, it is important to consider the developed research questions, the sample group, and the selection of appropriate questions relative to the data to be collected. The following best practices can assist in developing the survey instrument:

- When utilizing multiple-choice questions, ensure that the provided list of options is exhaustive (considers all possible options) and mutually exclusive.
- Rating questions can be used when measuring a respondent's viewpoint or satisfaction of elements, including the program design, administrative processes, or the program service delivery. The scale in a rating question need to be balanced and have clearly defined scale points to avoid misinterpretation.
- Open-text questions are usually used sparingly to add details to a response. Open-text questions require more effort from the respondent and having too many could lead to survey fatigue.
- When compiling questions in a document, ensure that there is a logical flow to the questions posed. It is expected that all the relevant questions are answered by the respondents.
- To maintain the participant's interest and ensure the survey is completed, it is recommended to keep the survey concise and limit the duration of a survey between 10-15 minutes.

Once an interview guide or survey instrument has been developed, it is beneficial to review them in their entirety. Every developed question should be assessed by its projected response relative to the research question and the data it can collect. Each section of questioning needs to follow a logical sequence such that the respondent can easily follow along and complete the entire survey. For example, if the survey initially asks about program design, and then followed by program delivery, it is best not to ask about program design again.

An integral component of developing interview guides and survey instruments is having either the evaluation administrator or program administrator review them. They are subject matter experts and can provide insight regarding a question's validity and whether it will yield actionable answers when asked of the targeted sample group.

The specific steps involved in collecting process evaluation data include:

- **Identify sources of data.** As part of the process evaluation assessment, the evaluator identified the data to be collected. The evaluator then needs to identify available sources of data, along with alternative collection strategies for when data access or integrity may be questionable.
- **Develop data collection guides and survey instruments.** When developing data collection guides and instruments, it is crucial to consider both the sample groups and research questions that enable appropriate data collection for analysis.
- **Coordinate data collection.** The evaluator, with assistance from the evaluation administrator, coordinates data collection from program staff, program delivery agents, program participants, and any other identified groups, such as contractors or trade allies. Depending on the data available, it may be necessary to ensure significant time and resources are allocated to develop data collection instruments throughout the evaluation process. The evaluator coordinates data collection efforts between different evaluation tasks (for example, data collection for process evaluation, gross impacts, and NTG analyses), and ensures implementing the most efficient outreach approach to minimize customer communication touchpoints, which leads minimizing the possibility of survey fatigue.
- **Collect data by implementing guides and fielding surveys.** After developing the necessary guides and instruments, the evaluator collects the data by following and implementing the guides and instruments. Sensitive data (e.g. personal identifiable information or confidential program traits) should be identified and monitored by the evaluator during data collection efforts to ensure compliance with privacy requirements. Whether the data collected is qualitative or quantitative, the captured information must be summarized without bias.

4.3.6 Step 6: Analyze and Report

As part of the data collection stage, a significant amount of data is expected. When analyzing this data, the evaluator needs to consider how it will be summarized and presented in the process evaluation report. The detail, content, and length of the report are guided by what is most helpful for the program and what is accomplished to address the research questions. When determining what to include in the report, the evaluator needs to consider having a balance between details the evaluation administrator would look for and the reported actionable items expected by the program administrator. To achieve this balance, the evaluator needs to work with the evaluation administrator and program administrator to define the types of information required while ensuring that the information and feedback are provided on time and includes the final process assessment.

Findings of a process evaluation are usually best presented in either a graphical or tabular format since these formats can provide more information in a clear and concise manner. The text in the report can highlight key findings and link the collected data to the research methods used for data analysis. In the report, evaluators need to outline instances where the findings confirm or contradict earlier findings including specific references to any previous studies conducted.

The reporting task is described in Section 5.

4.4. Cost Effectiveness

To evaluate the cost effectiveness of programs the following IESO guides and tool are to be used[8]:

- Conservation and Demand Management Energy Efficiency Cost Effectiveness Guide
- Conservation and Demand Management Cost-Effectiveness Tool User Guide
- Integrated Cost Effectiveness Tool

The Cost Effectiveness Guide describes standard industry metrics to assess the cost effectiveness of conservation and demand management (CDM) resources. Cost effectiveness assesses whether the benefits of an investment exceed the costs. Cost effectiveness tests are comparisons of benefits and costs expressed as both the dollar value of the net benefit (or cost) and as a ratio of benefits to costs. Table 4-10 outlines each cost effectiveness test, the key question it answers and a brief summary of the approach. The Cost Effectiveness Guide describes in detail each of the tests listed in Table 4-10.

The Integrated Cost-Effectiveness Tool is an Excel-based tool intended to support IESO staff, LDC staff, and other external service providers or delivery agents to calculate resource savings, budget, and cost-effectiveness metrics for new and existing energy conservation programs in Ontario.

The guides and tool may be updated from time to time.

[8] Guides and tools are provided on this website: http://www.ieso.ca/Sector-Participants/Energy-Efficiency/Evaluation-Measurement-and-Verification

Table 4-10 | Overview of Cost Effectiveness Tests

Cost Effectiveness Tests	Key Question Answered	Summary Approach
Total Resource Cost (TRC) test	How will the total costs of energy and demand in the utility service territory be affected?	Compares the costs incurred to design and deliver programs and customers' costs with avoided electricity and other supply-side resource costs (e.g., generation, transmission, natural gas, etc.)
Societal Cost (SC) Test	Is the utility, province or nation better off as a whole?	Identical to TRC approach, but also includes the avoided cost of "externalities" (e.g., carbon emissions, health costs, etc.)
Program Administrator Cost (PAC) Test	How will utility costs be affected?	Compares the costs incurred to design and deliver programs by the program administrator with avoided electricity supply-side resource costs[9]
Ratepayer Impact Measure (RIM) Test	How will utility rates be affected?	Compares program administrator costs and utility bill reductions with avoided electricity and other supply-side resource costs
Participant Cost (PC) Test	Will the participant benefit over the measure life?	Compares costs and benefits of the customer installing the measure
Levelized Delivery Cost (LC) Metric	What is the per-unit cost to the utility?	Normalizes the costs incurred to design and deliver programs per unit saved (i.e., peak demand or energy savings)

4.5. Market Effects Evaluation

Market effects occur when there is "a change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy-efficient products, services, or practices and is causally related to market intervention(s)." In CDM program evaluation, "market" refers to the commercial activity (e.g. manufacturing, distributing, buying and selling) associated with products and services that affect energy usage. A market effects evaluation measures the net effects at a market level when one or more CDM program efforts target a market. Net market effects are those effects that are induced by CDM programs and are net of market activities induced by non-energy efficiency programs, including normal market changes. The

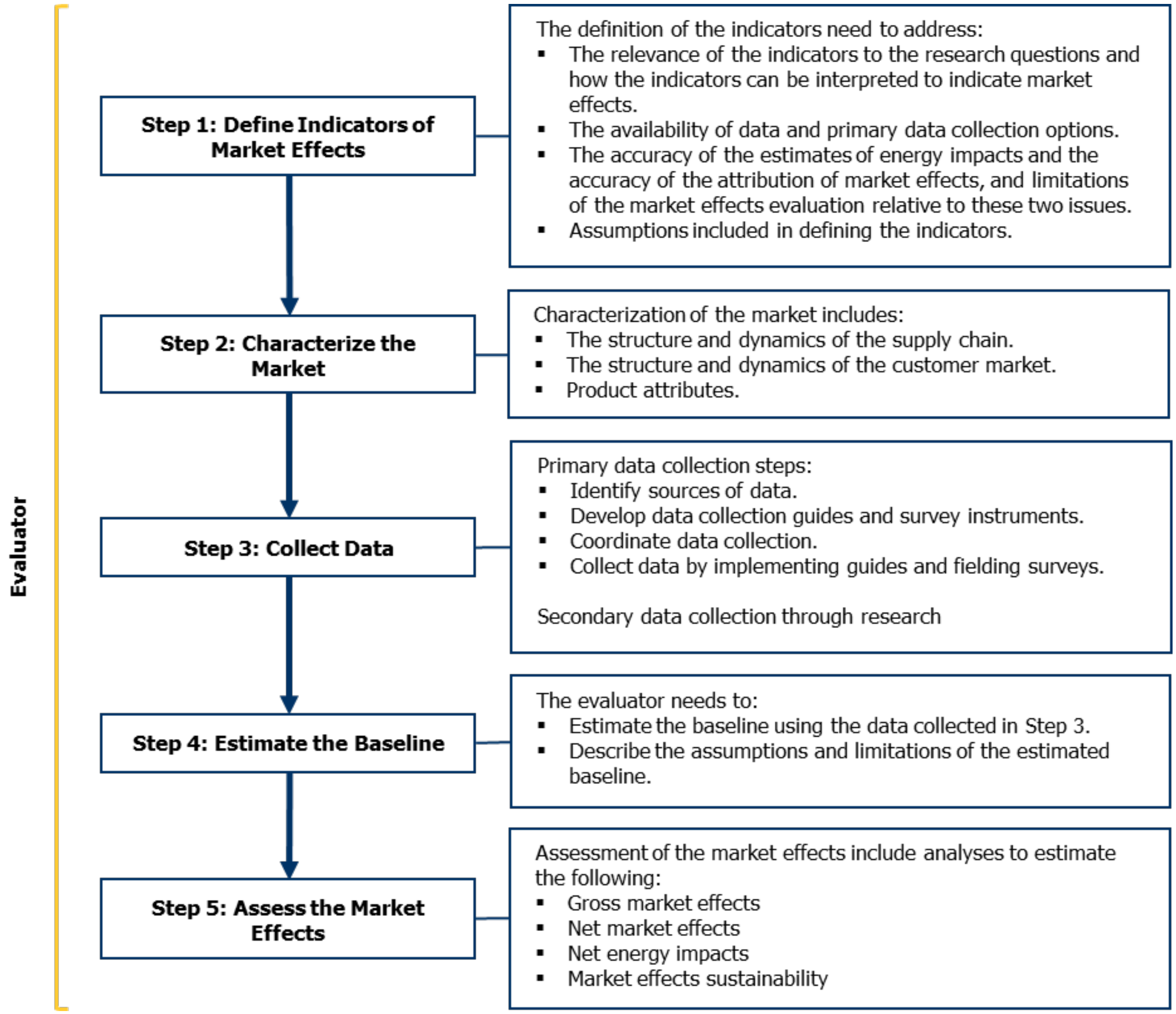
[9] The IESO, as the program administrator, would use avoided electricity supply-side resource costs. If a utility is responsible for electricity and natural gas resources, both of these benefits and costs would be included.

objective of market effects evaluation is to quantify the changes occurring in the market caused by CDM programs and to provide an estimate of the energy impacts associated with them. The market effect evaluation assesses only the current market effects and not those forecasted to occur at some future point.

A market effects evaluation usually does not apply to the measurement of individual program-level market effects or direct program savings typically used for program-level cost-effectiveness assessments and program-level adjustment decisions. Rather, the focus of the market effects evaluation is at a market level in which many different CDM programs can operate. This means that market effects usually focus on the effects of groups of programs within a market over multiple program cycles and applies when net market effects are to be estimated at a market rather than at a program level. A market effect evaluation is also appropriate when a single large and particularly effective program is expected to have broad and long-term market effects in a single market.

The scope of the market effect study and the research questions to be addressed are defined during evaluation planning, as discussed in Section 3. The steps included in the market effects evaluation are outlined in Figure 4-9 and discussed in the remainder of this section

Figure 4-9 | Market Effects Evaluation Steps



Evaluator
Step 1: Define Indicators of Market Effects
Step 2: Characterize the Market
Step 3: Collect Data
Step 4: Estimate the Baseline
Step 5: Assess the Market Effects

The definition of the indicators need to address:

- The relevance of the indicators to the research questions and how the indicators can be interpreted to indicate market effects.
- The availability of data and primary data collection options.
- The accuracy of the estimates of energy impacts and the accuracy of the attribution of market effects, and limitations of the market effects evaluation relative to these two issues.
- Assumptions included in defining the indicators.

Characterization of the market includes:

- The structure and dynamics of the supply chain.
- The structure and dynamics of the customer market.
- Product attributes.

Primary data collection steps:

- Identify sources of data.
- Develop data collection guides and survey instruments.
- Coordinate data collection.
- Collect data by implementing guides and fielding surveys.

Secondary data collection through research

The evaluator needs to:

- Estimate the baseline using the data collected in Step 3.
- Describe the assumptions and limitations of the estimated baseline.

Assessment of the market effects include analyses to estimate the following:

- Gross market effects
- Net market effects
- Net energy impacts
- Market effects sustainability

4.5.1 Step 1: Define Indicators of Market Effects

Based on the research questions, defined during the evaluation planning (see Section 3), and the program logic models (see Appendix F for an example of a logic model), the evaluator needs to define the appropriate indicators of market effects. Indicators of market effects are commonly categorized as follows:

- **Awareness and knowledge.** For example, recognition of program administrator's brand due to program activity.
- **Attitudes and beliefs.** For example, facility owners' assessment of credibility of energy efficiency information provided by different types of firms over a number of years.
- **Availability.** For example, the availability of energy efficiency technology in a market.
- **Trade ally promotional effort.** For example, extent to which participant and non-participant energy service companies have increased their marketing of energy efficient measures.
- **Incremental cost.** For example, comparing the cost of more energy efficient measures with the cost of the baseline measures.
- **Market share and sales.** For example, current purchases or sales of the technologies addressed by the energy efficiency programs.
- **Saturation and prevalence of practices.** For example, the saturation of an energy efficient technology based on the combined effects of cumulative sales / market share, removal and storage.
- **Changes in codes and standards.** For example, the influence of programs on the code and standard changes.

The indicators of market effects are used to draw conclusions about the energy changes in the market. The evaluator's definition of the indicators, need to address:

- The relevance of the indicators to the research questions and how these indicators can be interpreted to indicate market effects.
- The availability of data and primary data collection options.
- The accuracy of the estimates of energy impacts and the accuracy of the attribution of market effects, and limitations of the market effects evaluation relative to these two issues.
- Assumptions included in defining the indicators.

4.5.2 Step 2: Characterize the Market

Market characterization is defined as a "qualitative assessment of the structure and functioning of a market." The market characterization provides the evaluator with the necessary information and understanding of the market to inform the subsequent steps in the market effects evaluation. Market characterizations typically encompass the following kinds of information and analyses:

- **Structure of the supply chain.** Relationships and functions in the supply chain of key market actors including manufacturers, distributors, installers and retailers, regulators and professional associations. Additional elements of market structure include the number of firms and the level of concentration of market actors; the percentage of total supply chain revenues; and the direct customer sales accounted for by these market actors.
- **Dynamics of the supply chain.** Motivations and barriers to the development and promotion of efficient products; and services based on competitive position and/or government mandates.
- **Structure of the customer market.** Identification and size of the key customer segments and the percentage of total market revenues accounted for by those segments.
- **Dynamics of the customer market.** Motivation and barriers to the adoption of efficient products; and services based on needs, resource constraints, and established purchasing practices within the major customer segments.
- **Product attributes.** Performance and price characteristics of products; and services currently in the market; and of products and services in various stages of development. Trends in price and performance over time.

The evaluator undertakes the market characterization by considering the scope of the market effects evaluation, as defined in the evaluation planning, and guided by the information and analysis listed above. If a market characterization has been completed within the past couple of years, typically 3 to 5 years, it may not need to be redone, depending on the evaluation administrator's judgments regarding recent changes in the market.

4.5.3 Step 3: Collect Data

Data needs to be collected for the indicators of market effect to track market progress and thus determine whether market effects have occurred. Primary and secondary data are used to inform the indicators. Primary data collection involves gathering of data directly from various actors in the market of interest. Data activities that involve primary data collection vary in complexity, and the sample needs to be representative of the population of market actors. More detailed information about sampling is provided in Appendix B. Data collected through primary research methods include the methods discussed for process evaluation in Section 4.3:

- Interviews,
- Surveys,
- Delphi analyses, and
- Focus groups.

Similar to process evaluation discussed in Section 4.3, the data collection is dependent on key tools such as interview guides and survey instruments to collect market effects evaluation data. Additional information and best practices pertaining to interview guides and survey instruments are provided in Section 4.3. The specific steps involved in collecting market effects evaluation primary data is similar to the steps for process data collections as discussed in Section 4.3, and include:

- Identify sources of data.
- Develop data collection guides and survey instruments.
- Coordinate data collection.
- Collect data by implementing guides and fielding surveys.

In cases where secondary data is used, the evaluator needs to understand the manner in which the data was collected to be certain of its appropriateness for market effects estimation. Secondary data often provides a source for estimating market share for both efficient and less efficient equipment sold in a market and can be the most effective way to obtain data for non-program affected areas. Secondary data is most often obtained through research.

4.5.4 Step 4: Estimate the Baseline

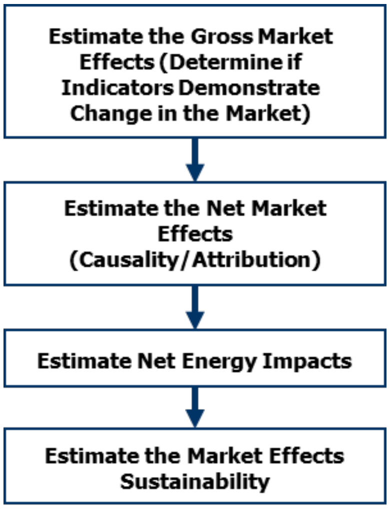
A baseline is generally used in energy efficiency program evaluation to represent both the value of the selected indicator prior to the program launch and the trend that indicator would take over time in the absence of program interventions. Baseline estimation refers to the quantitative estimation of the indicators of market effects that represent the level of market acceptance of the products and services promoted by the program under evaluation.

These indicators are generally estimated through relatively large sample surveys or through the inspection of sales data in the relatively few markets for which they are available. Development of baseline market share or other key indicators, such as the price of standard and efficient equipment, is often challenging and expensive. Evaluators need to either conduct relatively large sample surveys of end users and suppliers or negotiate with suppliers for the release of highly sensitive sales data. The evaluator needs to estimate the baseline using the data collected in the previous stepand describe the assumptions and limitations of the estimated baseline.

4.5.5 Step 5: Assess the Market Effects

The assessment of market effects includes the steps outlined in Figure 4-10. The steps are discussed in more detail below.

Figure 4-10 | Steps to Assess the Market Effects


Estimate the Gross Market Effects (Determine if Indicators Demonstrate Change in the Market)
Estimate the Net Market Effects (Causality/Attribution)
Estimate Net Energy Impacts
Estimate the Market Effects Sustainability

Once the data for the indicators of market effects have been collected for the baseline and the end date of the evaluation period, the evaluator needs to assess the change in indicators across the time period and estimate the gross market effect. For indicators such as market share and sales, it is generally reasonable to make direct comparisons between the end date of the evaluation and the baseline, since market share can be tracked directly over time. For other indicators such as awareness and knowledge, it is possible to make direct comparisons of indicators across time periods, but often the direction and intensity of change in indicators will vary. One method that has been found to be effective in this type of situation is a binomial test (see also Appendix B: Sampling Plan Design).

Causality needs to be assessed to estimate the net market effects. The goal of the analysis is to estimate the proportion of market changes that can be attributed to CDM program interventions, as opposed to those naturally occurring in the market or from non-CDM program interventions to arrive at the net market effects. Taking into consideration the indicators of market effects and the market characterization, the evaluator needs to select and apply the most feasible approaches to assess the casual links between the program activities and observer market changes. There are mainly four approaches to select from and these approaches are described in more detail in the information box.

APPROACHES TO ASSESS THE CAUSAL LINKS BETWEEN THE PROGRAM ACTIVITIES AND OBSERVED MARKET CHANGES

- Analysis of self-reported free-ridership, participant spillover, and non-participant spillover among market actors in the program domain. This approach relies on the description by local market actors of the influence of the program on end users or suppliers' decisions to characterize the extent of the program’s effect. This data is usually gathered through surveys of program participants and nonparticipants. The adoption of energy efficiency measures within the program, less free-ridership, plus participant spillover and plus non-participant spillover capture most of the net effect of the program on adoption of energy efficiency measures.
- Forecasting or retrocasting the non-intervention baseline. With this approach, evaluators develop a statistical model to estimate how the market would behave over time without the intervention of the program. A model that develops an estimate for a future date is called forecasting. A model that develops an estimate to describe pre-program conditions is called retrocasting. The forecast or retrocast estimate is compared with the actual behaviour of the market with the intervention in order to estimate net savings. For example, using prior market trends to estimate a natural adoption curve that describes how the market would behave without intervention and using it as the baseline (retrocasting).

- Cross-sectional comparisons of market conditions in the program domain to those in comparison areas. This approach uses comparisons of market share of the targeted technologies or other indicators of market effect among groups of market actors not addressed by the program as a baseline for estimating the net effects of the program in the program area.

- Structured expert judging. Structured expert judgment studies assemble panels of individuals with close working knowledge of the technology, infrastructure systems, markets, and political environments addressed by a given energy efficiency measure to estimate baseline market share and, in some cases, forecast market share with and without the program in place. Structured expert judgment processes employ a variety of specific techniques to ensure that the participating experts specify and take into account key known facts about the program, the technologies supported, and the development of other influence factors over time. The Delphi process is the most widely known method of this family of methods.

- Historical Tracing: Case Study Method. This method involves the careful reconstruction of events leading to the outcome of interest, for example, the launch of a product or the passage of legislation, to develop a 'weight of evidence' conclusion regarding the specific influence or role of the program in question on the outcome. Historical tracing relies on logical devices typically found in historical studies, journalism, and legal argument. These include:

  - Compiling, comparing, and weighing the merits of narratives of the same set of events provided by individuals with different points of view and interests in the outcome.

  - Compiling detailed chronological narratives of the events in question to validate hypotheses regarding patterns of influence.

  - Positing a number of alternative causal hypotheses and examining their consistency with the narrative fact pattern.

  - Assessing the consistency of the observed fact pattern with linkages predicted by the program logic model.

  - Researchers use information from a wide range of sources to inform historical tracing analyses. These include public and private documents, personal interviews, and surveys

To estimate net market effects, the evaluator needs to estimate the market share for the sales or counts of behavior, or other indicator(s) attributed to the program. The estimate is based on the data collected for the indicators of market effects. The net market effects need to be linked to an estimate of energy savings. The sales and counts of behavior or other indicators used to estimate market effects are linked to the energy savings. The energy savings are generally calculated by applying unit energy savings figures to the estimate of net adoption of the measures. The net adoption of the measures is determined when estimating the net market effects. The unit energy savings figures often come from the impact evaluations of resourceacquisition programs, but it can be developed using engineering-based calculations or obtained from deemed savings databases.

Market effects sustainability is the degree to which one can expect the market changes to last into the future. These assessments are prospective and involve the compilation and interpretation of information on the market effects sustainability of the various indicators of market effects. The evaluator needs to select and appropriate approach to assess the market effects sustainability, and these approaches usually include:

- Assessing data and information on the market effects sustainability of indicators obtained through choice and ranking surveys, focus groups and Delphi surveys.
- Identifying and assessing changes in market structure and operations, and how the changed market contains mechanisms to sustain them. This could include, for example, examining profitability analyses for important support businesses or business operations and how these are maintained without continued program intervention.

The result produced by the evaluator is an estimation of market effects sustainability, expressed as a statement on the likelihood of the market effects continuing without the energy efficiency program intervention or with reduced interventions.

5. Reporting

The report provides a summary of significant conclusions and presents recommendations based on the evaluation findings. When drawing conclusions, the evaluator needs to ensure their conclusions are without bias and are based on the data collected from the evaluation process as opposed to broad presumptions based solely on their experiences. Although assumptions from experience in other jurisdictions may be provided, it should only be done within the context of a comparative analysis requested in the evaluation scope of work. Given the influence evaluation conclusions and recommendations can have on an organization's priorities and budget allocations, evaluators need to ensure that they are supported by the research findings; and fall within the scope of the evaluation; and recommendations should be relevant and actionable by the organization.

In summary, the specific steps for reporting include:

- Analyze the data by considering the results to be presented in the report.
- Develop the report outline and content requirements, and review them with the evaluation administrator.
- Develop the report and review it with the evaluation administrator.

Evaluation results can be presented in a variety of ways. Most comprehensive evaluations usually includes both impact and process evaluation. Reporting the results of these evaluations typically include the information for each program as summarized in Table 5-1.

Table 5-1 | Summary of Results Included in Reports

Type of Evaluation	Summary of Results
Impact Evaluation	• Number of participants • Program realization rate (%) • Gross verified demand savings (MW) • Gross verified annual energy savings (GWh) • Gross verified lifetime energy savings (GWh) • Net to gross ratio • Net peak demand savings (MW) • Net annual energy savings (GWh) • Net lifetime energy savings (GWh) • Other key impact evaluation findings
Process Evaluation	• Research questions

Type of Evaluation	Summary of Results
	• Observations • Recommendations
Cost Effectiveness	• Program Administrator Cost (PAC): - Benefit ($m) - Cost ($m) - Net benefit ($m) - Net benefit ratio • Total resource cost (TRC): - Benefit ($m) - Cost ($m) - Net benefit ($m) - Net benefit ratio • Levelized unit energy cost (LUEC): - $/MWh - $/MW-yr • Other key cost effectiveness results

6. Glossary of Program Evaluation Terminology

8760s

Full year hourly consumption loads.

Accuracy

The correspondence between the measurements made on an indicator and the actual value of the indicator at the time of measurement.

Bias

The extent to which a measurement, sampling, or analytical method systematically underestimates or overestimates a value.

"CDM" Conservation and Demand Management

Outside of Ontario CDM is often referred to as Demand–Side Management (DSM) and so CDM and DSM are often used interchangeably.

Comparison Group

A group of individuals or organizations that have not had the opportunity to receive program benefits and that have been selected because their characteristics match those of another group of individuals or organizations that have had the opportunity to receive program benefits. The characteristics used to match the two groups should be associated with the action or behaviour that the program is trying to promote. In evaluation practice, a comparison group is often used when random selection of recipients of the program benefit and a control group is not feasible.

Control Group

A randomly selected group of individuals or organizations that have not had the opportunity to receive program benefits. A control group is measured to determine the extent to which its members have taken actions promoted by the program. These measurements are used to estimate the degree to which the promoted actions would have been taken if the program did not exist.

Cost-Benefit

Comparison of a program's outputs or outcomes with the costs. Benefit-cost is an alternate. The comparison of a cost to a benefit is often expressed as a ratio.

Cost-Effectiveness

Comparison of a program's benefits with the resources expended to produce them.

Cost-Effectiveness Evaluation

Analysis that assesses the cost of meeting a single output, objective, or goal. This analysis can be used to identify the least costly alternative to meet that output, objective, or goal. Cost-benefit analysis is aimed at identifying and comparing all relevant costs and benefits. The analysis is usually expressed in dollar terms. The two terms (cost effectiveness and cost benefit) are often interchanged in evaluation discussions.

Deemed Savings

An estimate of an energy savings or demand savings outcome for a single unit of an installed energy-efficiency or renewable-energy measure that:

- Has been developed from data sources and analytical methods that are widely considered acceptable for the measure and purpose, and
- Will be applied to situations other than that for which it was developed.

That is, the unit savings estimate is "deemed" to be acceptable for other applications. Deemed savings estimates are more often used in program planning than in evaluation. They should not be used for evaluation purposes when a program-specific evaluation can be performed. When deemed savings estimates are used, it is important to know whether its baseline is an energy-efficiency code or open-market practice. Besides the IESO's Measures and Assumptions Lists, an extensive database of deemed savings is also available in California's Database for Energy Efficiency Resources (DEER). Note that the deemed savings in DEER are tailored to California and should not be used for Ontario initiatives without thought or review.

Defensibility

The ability of evaluation results to stand up to scientific criticism. Defensibility is based on the assessment by experts of the evaluation's validity, reliability, and accuracy.

Evaluation, Measurement & Verification (EM&V)

The undertaking of studies and activities aimed at assessing and reporting the effects of an energy efficiency program on its participants and/or the market environment. Effectiveness is measured though energy efficiency and cost effectiveness.

Evaluation Administrator / Evaluation Manager

The person responsible for defining the scope for the program evaluation. This person is also the point-of-contact for EM&V contract management. This person is sometimes referred to as an evaluation manager.

Evaluator

The individual(s) or firm(s) selected to develop and implement the evaluation plan based on the scope defined by the evaluation administrator. The evaluation contractor could also be referred to as the "independent, third-party evaluator" or the "evaluator".

Free-Rider

A program participant who would have implemented the program measure or practice in the absence of the program. Free riders can be total, partial, or deferred.

Gross Verified Savings

Gross verified savings calculations are based on the difference between energy and demand use after the implementation of a program and an assumed set of baseline conditions that estimate what energy consumption and demand would have been in the absence of the program. The gross verified savings are determined by multiplying the reported savings with the realization rate.

Impact Evaluation

The application of scientific research methods to estimate how much of the observed results, intended or not, are caused by program activities and how much might have been observed in the absence of the program. This form of evaluation is employed when external factors are known to influence the program’s outcomes in order to isolate the program’s contribution to achievement of its objectives.

Indicator

An indicator is the observable evidence of accomplishments, changes made, or progress achieved. An indicator is also a particular characteristic used to measure outputs or outcomes; a performance quantifiable expression used to observe and track the status of a process.

Interactive Effects

Also referred to as cross effects, are energy effects created by an energy conservation measure but not measured within the measurement boundary.

Logic Model

A plausible and sensible diagram of the sequence of causes (resources, activities, and outputs) that produce the effects (outcomes) sought by a program.

Market Effects

A change in the structure or functioning of a market or the behaviour of participants in a market that results from one or more program efforts. Typically, the resultant market or behaviour change leads to an increase in the adoption of energy-efficient or renewable-energy products, services, or practices. Examples include an increase in the proportion of energy-efficient models displayed in an appliance store, the creation of a leak inspection and repair service by a compressed-air-system vendor, an increase in the proportion of commercial new-construction building specifications that require efficient lighting.

Measurement

A procedure for assigning a number to an observed object or event.

Measures and Assumptions List

The IESO-approved electricity-sector "deemed savings" lists is to be used for program planning and forecasting purposes. One major goal of EM&V program evaluations is to confirm or update these assumptions.

Net Verified Savings

Net verified savings recognize behavioural factors and represent benefits that are only attributable to, and the direct result of, the program in question. Program net verified savings are calculated by multiplying the gross verified savings with the net-to-gross (NTG) ratio.

Non-Response Bias

Non-response bias occurs when there is a significant difference between those who responded to a survey and those who did not due to an influencing factor preventing them from responding (for example, lack of familiarity with the survey instrument).

Normalized Savings

Savings calculated based on adjustments. The baseline energy use is adjusted to reflect "normal" operating conditions. The reporting period energy use is adjusted to reflect what would have occurred if the facility had been equipped and operated as it was in the baseline period under the same "normal" set of conditions. These normal conditions may be a long-term average, or those of any other chosen period of time, other than the reporting period.

Outcome

A term used generically with logic modeling to describe the effects that the program seeks to produce. It includes the secondary effects that result from the actions of those the program has succeeded in influencing.

Outcome Evaluation

Measurement of the extent to which a program achieves its outcome-oriented objectives.

Outcome evaluations measure outputs and outcomes (including unintended effects) to judge program effectiveness and may also assess program process to understand how outcomes are produced.

Output

A term used generically with logic modeling to describe all of the products, goods, and services offered to a program's direct customers.

Process Evaluation (or Assessment)

An evaluation of the extent to which a program is operating as its implementation intended. Process evaluations assess program activities' conformance to statutory and regulatory requirements, to program design, and to professional standards or customer expectations.

Program Administrator

The persons or organizations responsible for the design, development, and implementation of an energy efficiency, conservation, or demand response initiative. A program administrator may also be referred to as a "program manager" or a "program implementer." An LDC may also be a program administrator. Outside of the EM&V context there may be distinctions between program administrators and external program managers or other subtleties that are ignored in the EM&V context. In the EM&V context a program administrator is someone (or an entity) other than the evaluation-related staff or entities.

Program Evaluation

Program evaluations are independent systematic studies conducted periodically on an ad hoc basis to assess how well a program is working and whether the program it is achieving its intended objectives. Program evaluations are conducted by experts external to the program staff.

Program Logic Model

A diagram showing a causal chain with links that go from resource expenditure to long-term outcomes for a program.

Program Manager

The individual/group responsible for implementing a program.

Qualitative Data

Information expressed in the form of words.

Quantitative Data

Information expressed in the form of numbers. Measurement gives a procedure for assigning numbers to observations.

Quasi-prescriptive Measure

A quasi-prescriptive measure has varying resource savings estimates according to the technology or type of equipment and the context in which they are used. It contains key, measure-specific inputs to estimate energy and peak demand savings for each program participant. It provides a methodology that allows estimating resource savings for various scenarios rather than relying on a fixed savings value for all scenarios. A quasi-prescriptive approach will allow different parameters or variables to be assumed to estimate different levels of resource savings for different retrofits in different business segments.

Random Assignment

A method for assigning subjects to one or more groups by chance.

Realization Rate

At the program level, the ratio of gross verified savings to the reported savings is referred to as the realization rate.

Reported Savings

Reported savings are the energy and demand savings reported, or claimed, by applicants or program implementation vendors. The savings are determined by the applicants or implementation vendors.

Resource-Acquisition Programs

Energy efficiency programming with a focus on achieving verifiable energy and/or demand savings within the context of an existing market system.

Peak Demand

The peak demand relates to energy demanded over the course of pre-defined period of time (i.e., 1 pm-7 pm) during which the overall demand on the province's electricity grid tends to be higher, on average. The IESO defined peak demand for summer, winter and weather dependent measures. These definitions are provide in Table 6-1 and Table 6-2.

Table 6-1 | Summer and Winter Peak Demand - Average Load Reduction Over Block of Hours

Period	Time	Months
Summer (Weekdays)	1:00pm – 7:00pm[1]	June 1 – August 31
Winter (Weekdays)	6:00pm – 8:00pm	December 1 – February 28

1: Daylight savings time-adjusted.

Table 6-2 | Weather Dependent Measures Peak Demand - Weighted Average of the Monthly Maximum Load Reduction[2]

Period	Time	Months	Weighting[3]
Summer (Weekdays)	1:00pm – 7:00pm[1]	June	30%
		July	39%
		August	31%
Winter (Weekdays)	6:00pm – 8:00pm	December	65%
		January	16%
		February	19%

1: Daylight savings time-adjusted.

2: Typically implemented as "at design conditions" and/or for the top facility hour of the month.

3: Weighting is based on the proportion of Top-10 hours that occur in that month.

Prescriptive Measures

A prescriptive measure uses defined or fixed input assumptions embedded into the energy and demand savings equations. These input assumptions can include default efficiencies for a type of equipment specified or annual operating hours for the type of building selected.

Probability Sampling

A method for drawing a sample from a population such that all possible samples have a known and specified probability of being drawn.

Random Assignment

A method for assigning subjects to one or more groups by chance.

Rebound Effect

A change in energy-using behaviour that yields an increased level of service and occurs as a result of taking an energy efficiency action.

Representative Sample

A sample that has approximately the same distribution of characteristics as the population from which it was drawn.

Self-Selection Bias

Self-selection bias occurs when people volunteer to participate in a study/survey. Those who choose to participate (self-select into the study) may share a characteristic that makes them different from non-participants. In most instances, self-selection will lead to biased data, as the respondents who choose to participate will not be representative of the entire target population.

Simple Random Sample

A method for drawing a sample from a population such that all samples of a given size have equal probability of being drawn.

Spillover

Reductions in energy consumption and/or demand caused by the presence of the energy efficiency program, beyond the program-related gross verified savings of the participants. There can be participant and/or non-participant spillover.

Verified Savings

The net evaluated energy and demand savings of a program. Verified savings are used as the base for the allocation of savings to targets or for official reporting purposes.

7. Appendix A: Evaluation Scope of Work Template

Sections 3.1 and 3.2 provide guidance in developing the evaluation scope of work. An evaluation scope of work template is included in this appendix.

Evaluation Scope of Work Template

Program Overview

Program Description

- Provide a short introduction of the program offer from the perspective of the program manager. The introduction should provide a high-level description of the planned program strategy. Where appropriate, include the following descriptions:
- Goals and Objectives: A statement of the goals and objectives for the program and the rationale for the evaluation.
- Target Market: Profile each market segment targeted by the program offer. Describe the size and characteristics of each target market. The target market should match the segments defined in the Program Logic Model.
- Eligibility Criteria: Describe the protocols/procedures that will be used to qualify the program applicants or markets targeted.
- Key Program Elements: Highlight the intended program process flow. Each program element should be identified in a 1-page graphic and annotated in the text that follows. This information should be drawn directly from the program design documents.
- Program Timing: A schedule of when the key elements of the program will be in the market, including program launch date and program end date.
- Estimated Participation: Estimated participation, by measure if applicable, for the program.

Program Theory / Program Logic Model (if available)

Introduce the mechanisms by which the program will function.

Even when a program manager provides a detailed logic model, the evaluator should investigate independently the causal influence of each program element towards the realization of the intended programmatic impacts. The program manager should review the logic model to ensure it is an accurate portrayal of the program theory.

Annotate the program logic model from the top (resource allocation) to bottom (intended impacts). Of particular interest are the linkages between program outputs and observed outcomes. Where practical, each connecting line or arrow should be annotated as a researchable programmatic assumption (null hypothesis).

Previous Program Evaluations

A brief description of similar program evaluations relevant to the program within the evaluator's portfolio and in other jurisdictions, including pilots.

Evaluation Goals and Objectives

Introduce the goals and objectives of the planned evaluation and indicate the rationale for the evaluation: administrative (verified savings), experimental (measure effectiveness), qualification (program pilot), or operational (cost-effectiveness).

Overarching Concerns

Provide a list of questions posed by the evaluation administrator to the evaluator. These should be categorized and refined as necessary to adequately communicate the areas of investigation sought by those sponsoring, operating, or participating in the program offer.

Research Questions

From the overarching concerns of program stakeholders, a set of research questions should be developed by the evaluation administrator and presented in this section. The number of research questions should be limited and prioritized based on the reasonable use of resources.

Evaluation Approach

Introduce the details of the approach that follows.

Evaluation Type (repeat for each type)

Provide a description of the required evaluation types and summarize the anticipated experimental approach. The frequency of the evaluation type such as "Annual Impact Evaluation" or "Year One Process Evaluation" need to be included in the title. In the description, highlight the major deliverables needed to complete each study and special methods sought from the evaluation contractor.

Evaluation Dependencies

Discuss key collaborations essential to the successful implementation of the evaluation. Common dependencies associated with industry research include access requirements, data sharing, funding support and enabling stakeholders.

Data Collection Responsibilities

Provide a list of all the data that must be collected to support the evaluation of the program and who is responsible for its collection.

Evaluation Schedule

Provide a list of all the physical deliverables that will be part of the Evaluation. For example, evaluation plans, memos, interim reports and final reports.

8. Appendix B: Sampling Plan Design

This appendix provides additional information to assist with designing a sampling plan, and includes:

- Sampling advantages and considerations
- Precision and confidence
- Deciding on a statistical test

Sampling Advantages and Considerations

Some of the main advantages of sampling are:

- That it is less expensive than conducting a census of the whole population
- Easier data analysis and greater flexibility in the application of analytical methods
- It can lead to greater sensitivity for the study of populations and sub-populations (as required)

Below are key considerations to keep in mind while developing a sample:

- To avoid bias in the sample that will be selected, ensure that the population from which the sample will be selected is representative of the overall program. If the sample is not representative of the larger population, then it is not possible to say anything about the larger population by studying the smaller sample. Common biases found during sampling, particularly for evaluations, include self-selection bias, non-response bias, and voluntary response bias. If researchers are aware of or perceive that there is a high likelihood that such biases may impact results, steps should be taken to mitigate such biases during the sampling design stage.
- It is never certain that the sample would achieve the exact results as the population under study. The sample only provides an estimate of the program effect. There is always a degree of uncertainty embedded in the estimate. Therefore, short of taking a census, there must be a recognition that some degree of uncertainty exists in any statement of program effect.

Precision and Confidence

A critical requirement in developing a sampling design for any experiment is a clear understanding of the minimum threshold of the difference between the treated and not treated customers that are considered meaningful for those who will be using the results for program planning. As discussed below, the size of the difference that will be considered to be meaningful has profound implications for the required sample size. In general, the smaller the difference that must be detected, the larger the sample size (of treatment and control group customers) needed to detect it. If the cost of the program is known or can be estimated, it is possible to identify the minimum change in energy use that would be required to justify investment in it.

For example, suppose a 5% reduction in energy use would be required to justify investment in a given training program for the benefits to outweigh the costs. The sample sizes for treatment and control conditions should be set so that a difference of at least 5% can be reliably detected 80-95% of the time. A related issue that also influences the sample size required in an experiment is the quantity of sampling error that is tolerable from the planning point of view.

In analyzing the results obtained from a statistical experiment, it is possible to make two kinds of inferential errors arising from the fact that one is observing samples. One can incorrectly conclude that there is a difference between the treatment and control groups when there is not one due to sampling variation). This is called a Type I error. Alternatively, one can incorrectly conclude that there is not a difference when in fact there is one. This is called at Type II error. The challenge in designing experimental samples is to minimize both types of errors. This is done by choosing sample sizes that minimize the likelihood of these errors. Additional detail is provided in the information box below, Minimizing Inferential Errors.

MINIMIZING INFERENTIAL ERRORS

Type I – Statistical Significance or Confidence

It is possible to calculate the likelihood of committing a Type I error from information concerning the inherent variation in the population of interest (the variance), the required statistical precision (as described above), and the sample size. This probability, called alpha, is generally described as the level of statistical significance or confidence. It is often set to 5% so that the sample size for the experiment is such that there is no more than 5% chance (one chance in 20) of incorrectly concluding that there is a difference between the treatment and control group of a given magnitude, when there really is not one. However, as in the case of statistical precision, the selection of alpha is subjective; it depends on the experimenter’s taste for risk. It could be set to 1% or 10% or any other level with attendant consequences for confidence in the results. For training and segment support studies, it should probably be set to 5%.

Type II – Statistical Power

Type II error is the converse of Type I error – concluding that the treatment made no difference when in fact it did. For a given population variance, specified level of statistical precision and sample size, the probability of incorrectly concluding that there is not a difference when indeed there is a difference is determined by the choice of alpha (the probability of making a Type I error). All other factors equal, the lower the probability of making a Type I error, the higher the probability of making a Type II error. In other words, for a given sample size, the more confident the researcher wants to be that they are not incorrectly finding a statistically significant difference, the less sure they can be that they have missed a statistically significant difference.

The likelihood of making a Type II error can be calculated for a given experiment and generally decreases as sample size increases. The likelihood of avoiding a Type II error is generally referred to as the statistical power of the sampling design.

The statistical power used in calculating required sample sizes for experiments is subjective and has generally been set at about 90%. That is, it is set so that only one time in ten will the experimenter incorrectly conclude that there is not a difference of a specified magnitude when indeed there is one. For Capacity Building experiments, statistical power (or confidence) should probably be set at 90%.

The analysis approach used to estimate impacts can also have a significant impact on sample sizes. For example, sampling can be much more statistically efficient if the effect(s) of the treatment(s) are being measured as differences (e.g., pre-test, post-test) of ratios or as regression estimators. This is true because the variance of these parameters in populations under study is usually quite a bit smaller than the variance of the raw variables, and the smaller the inherent variance of the measurements of interest, the smaller the required sample size. As discussed below, panel regression methods with pre-test, post-test experimental designs can significantly reduce sample sizes.

Deciding on a Statistical Test

Statistical testing is generally used by researchers to describe a given population, make comparisons against a hypothetical value, or establish predictions based on known values. This section outlines statistical tests commonly used to make inferences. However, this section is not intended to be a step-by-step manual that explains how to perform these calculations, since most situations are unique in terms of inputs and desired outcomes.

As there are several types of statistical test models that can be employed during an experiment, researchers must take care to determine the most appropriate test to answer their particular research question(s). Statistical test selection can be a simple or complex exercise depending on the nature of the study. Because one or more tests may be suitable, to address a research question it is recommended that one consult a statistics professional before finalizing the required test.

To determine the most suitable test, the researcher must first determine the distribution of the population. Populations with a normal (Gaussian) distribution, or close to a normal distribution, will be more suitable to certain tests while unique techniques may make it harder to test populations with a non-normal distribution. This guideline focuses on those tests that are suitable for normally distributed populations. However, it is important to note that if the population being studied is not normally distributed, there are alternative testing methods that should be employed. Common examples of where a population may not be normally distributed include purchasers of luxury items and early adopters of new technologies.

Researchers are to determine if they anticipate one possible outcome or two possible outcomes from the test being performed. As well, the researcher must also determine the purpose for the outcome of the test. Table 8-1 below is a matrix of commonly used statistical tests for normally distributed populations. Keep in mind that the items included are only some of the tests, researchers may wish to use other test models.

Table 8-1 | Commonly Used Statistical Tests

Goal	**Possible Outcomes** **One (Measurement)**	**Two (Binominal)**
Describe a group	Mean and standard deviation	Proportion
Compare a group to a hypothetical value	One-sample t-test	Chi-square test or binomial test
Compare two unpaired groups	Unpaired t-test	Fisher’s test or chi-square test
Compare two paired groups	Paired t-test	McNemar’s test
Compare three or more unmatched groups	One-way analysis of variance	Chi-square test
Compare three or more matched groups	Repeated measure analysis of variance	Cochrane Q
Quantify association between two variables	Pearson correlations	Contingency coefficients
Predict value from another measured variable	Simple linear regression or nonlinear regression	Simple logistic regression
Predict value from several measured or binomial variables	Multiple linear regression or multiple nonlinear regression	Multiple logistic regression

9. Appendix C: Technology-Based Programs Energy Savings Calculation Methodologies

This appendix provides more information on the different energy savings calculations for technology-based programs. These methodologies are:

- Deemed savings approaches
- Custom M&V approaches

Due to the potential of advanced measurement and verification (M&V 2.0) to improve the current measurement and verification practices, an update on the current status of M&V 2.0 is provided at the end of the section.

Deemed Savings Approach

The deemed savings approach uses agreed-upon values for program-supported measures with well-known and documented savings values. Deemed savings are determined by the evaluator using prescriptive and quasi-prescriptive assumptions and standard equations for determining gross verified savings. When applying the deemed savings approach using MALs, usually no field measurement is needed to determine the savings per measure or project. Gross impacts are determined by multiplying the per measure values derived from the MALs by the verified number of installations.

Prescriptive Approach Savings Calculation

Savings are prescribed on a per-participant or per-measure basis and represent an average level of savings that would be achieved by a participant implementing the energy efficient measure. For prescriptive measures, the savings evaluation depends on:

- The type of technology
- The number of installations
- The prescribed savings estimates for the technology used

Prescriptive gross verified savings are calculated based on the number of participants and/or measures installed, multiplied by the prescribed savings per participant or measure. The gross verified savings are calculated as shown in Equation 9-1.

Equation 9-1 | Prescriptive Gross Verified Savings

$$\boldsymbol{PS_{gross} = N * s}$$

Where:

PS_{gross} = Gross Verified Savings (kWh or kW)

N = Number of tracked measures or participants

s = Prescribed savings per measure or participant as listed in MAL (kWh/unit or kW/unit)

Quasi-Prescriptive Approach Saving Calculation

Savings are determined using a prescribed methodology that uses key, project-specific inputs to estimate the savings for each participant or measure installed. For quasi-prescriptive measures, the savings evaluation depends on:

- The type of technology
- The number of installations
- Project-specific information generally collected from participants implementing the measures (for example, savings per unit capacity or per hour of operation)
- Other information needed to adjust savings estimates (scalable basis)

A common quasi-prescriptive methodology is to prescribe energy savings for a measure on a scalable basis (for example, kWh savings per unit of capacity or per hour of operation). If the relationship between the scalable basis and the savings is linear, then gross verified savings can be calculated from the number of participants or measures installed, multiplied by the average participant value of the scalable basis, multiplied by the prescribed scalable savings. The gross verified savings are calculated as shown in Equation 9-2.

Equation 9-2 | Quasi-Prescriptive Gross Verified Savings

$$\boldsymbol{PS_{gross} = N * SB_{avg} * \ s_{scale}}$$

Where:

PS_{gross} = Gross Verified Savings (kWh or kW)

N = Number of tracked measures or participants

SB_{avg} = Scalable basis (e.g., average participant equipment capacity)

s_{scale} = Prescribed savings per participant or measure (e.g., kWh per participant per scalable basis)

Other potential quasi-prescriptive approaches may, as an example, include engineering equations that utilize key participant inputs, prescribed inputs, or default values, to estimate savings estimates or use similar inputs to reference a MAL. In these instances, the gross verified savings are calculated from the sum of the savings calculated for each participant or measure installed.

Custom M&V Approaches

Custom M&V approaches are typically applied when no prescribed measures are found on the MALs for the types and conditions of measures included in a program. Custom M&V approaches require that gross verified savings be tracked and estimated on a project-by-project basis. Custom projects tend to be more complex than those using prescriptive measures (for example, building equipment retrofits where equipment load profiles are variable, etc.) and savings estimates use specific equations that can change on a project- by-project basis. Therefore, project-level M&V is essential for tracking and reporting savings and should at least be taken into consideration for all situations requiring a custom M&V.

Custom projects that require implementing a custom M&V approach include:

- Equipment retrofit only
- Operational change only
- Equipment retrofit and operational change
- Multiple energy conservation measures

Custom M&V activities typically consist of some or all of the following:

- Meter installation, calibration and maintenance
- Data gathering and screening
- Development of a computation method and acceptable estimates
- Computations with measured data
- Reporting, quality assurance, and third-party verification of reports

Depending on the measure type, as well as the uncertainty and magnitude of the reported savings estimates, the evaluator conducts one of two levels of rigour for each sampled project to calculate the gross verified savings – basic (verification only) or enhanced (measurement and verification). The basic and enhanced levels of rigour, which are based on the widely recognized International Performance Measurement and Verification Protocol (IPMVP), are:

- Basic Rigour – Simple Engineering Model with On-Site Measurement
- Enhanced Rigour – Retrofit Isolation Engineering Models with On-Site Measurement
- Enhanced Rigour – Billing Analysis with On-Site Verification Only
- Enhanced Rigour – Whole Building Simulation with On-Site Verification Only

In order to successfully implement a custom M&V approach, the evaluator needs to collect the following information:

- Type(s) of equipment being installed
- Type(s) of equipment being replaced
- Customer address or location
- Engineering analyses and/or computer simulations
- Other information needed to determine savings for custom projects

While conducting custom M&V activities, the evaluator needs to ensure proper documentation and reporting on project-level assessments are provided. The M&V report needs to contain sections and complete descriptions of the processes used and results for all required elements in the M&V plan.

Advanced M&V: M&V 2.0

Background

Advanced measurement and verification (M&V 2.0) has the potential to modernize energy efficiency markets, create innovative delivery mechanisms and set new evaluation standards and policies that will improve the current measurement and verification practices. The purpose of this section is to provide updates on the current status of M&V 2.0 and discuss the benefits and limitations of using M&V 2.0 for conducting EM&V. Given that the research in this field is constantly evolving, there are several definitions of M&V 2.0 in the literature. However, the following definition from Lawrence Berkeley National Laboratory and Rocky Mountain Institute comprehensively describes M&V 2.0[10]: “M&V 2.0 refers to the increasing granularity of available energy consumption data, and the enabling of automated M&V methods that continuously analyze the data and provide early, accurate and valuable insights to various stakeholders about energy savings estimates”. This advancement in data collection and processing has the potential to enhance EM&V approaches and eventually reduce the time and cost typically associated with evaluation.

Current State of M&V 2.0

The implementation of M&V 2.0 for EM&V is still in the development stage, and further efforts are required to establish universal standards and policies to adopt M&V 2.0 for evaluation purposes. Several pilot studies have implemented M&V 2.0 for program evaluations, and the insights from these studies are summarized below.

Methodology

M&V 2.0 applications are generally characterized by:

[10] Lawrence Berkeley National Laboratory (2017). The Status and Promise of Advanced M&V: An Overview of “M&V 2.0” Methods, Tools and Applications. Website: https://eta.lbl.gov/sites/all/files/publications/lbnl-1007125.pdf

- Real-time data collection using advanced metering infrastructure and the availability of more granular data for analysis.
- Automated analysis of large volumes of data.

M&V 2.0 software tools or techniques employ the IPMVP Option C – Whole Facility Measurement – approach for data analysis. IPMVP Option C involves the whole facility, utility or sub-meter data analysis procedure to verify the performance of energy efficiency projects. The metered data can either be monthly bill data or more granular in the form of daily, hourly, or sub-hourly readings from advanced meters. Utilizing this metered data, a statistical regression model is developed to quantify the energy savings. Hence the basic approach for measurement and verification is unchanged for the advanced M&V. However, the use of more granular data and automated analysis allows for the following advantages when compared to traditional M&V:

- Completing M&V in a shorter timeframe, as data is available with near-real-time access. Completing the M&V can be done as early as three (3) months after project implementation, depending on the data source and retrofit type.
- Ability to realize savings at a lower threshold. Savings as low as 5% can be quantified using hourly data rather than 10% savings if monthly data is used.
- Quantify savings seasonally by the time of day and/or day of the week.

The granular data collected using advanced meters can be processed using free or proprietary software tools, open-source code or any custom code. An example of an open-source code that can be used for M&V 2.0 is CalTRACK 2018. There are two statistical models used by the software tools to calculate the savings; (1) the change point model or (2) the Time of Week and the Temperature (TOWT) model. These models are based on linear regression of energy use with respect to outdoor air temperature.

Using M&V 2.0 to Conduct EM&V

The advancement in data collection and automated processing creates an opportunity to conduct a real-time evaluation of energy efficiency programs to gain immediate insight into program performance and make the necessary adjustments to programs as required. A comparison of the common elements of traditional and real-time evaluation approaches is presented in Table 9-1 below. Traditional M&V methods used for evaluation purposes allow a relatively small sample of sites to represent the overall program population. Unbiased sampling and rigorous site analysis can provide acceptable estimates of energy savings with a certain statistical precision (for example, 90/10 confidence and precision). With M&V 2.0, analysis is not limited to a narrow sample of sites that undergo rigorous M&V efforts. Evaluation efforts can be cost-effectively scaled up to include the entire population in the analysis, potentially removing any sampling error. A model that tracks the required data can be deployed to predict energy savings and automatically output the result in an ongoing manner (while the program is implemented concurrently). Although M&V 2.0 has the potential to remove sampling error, it raises a concern regarding the accuracy and bias of the model that is being used to process the real-time data.

Further efforts are required to predict if M&V 2.0 accurately models buildings' load in order to assess individual measure impacts. Additionally, M&V 2.0 approaches automatically analyze energy consumption data to estimate measures' savings and might fail at separating the influence of other variables that can influence the energy usage, such as varying operating loads and implementing other retrofits at the same time.

Table 9-1 | Traditional versus Real-time Evaluation Approach

Element	**Traditional Evaluation Approach**	**Real-Time (M&V 2.0) Evaluation Approach**
Timing of Sampling	1-2 years after implementation is complete	Granular data can be collected instantaneously after measure implementation. The entire sampling population can be evaluated.
Metrics Tracked	Single post realization rate, NTG values	Rolling realization rate and net savings value
Timing of Reporting	Often at the end of the evaluation. Can be around two years after implementation	Can be concurrent with program implementation
Communications with implementers	Feedback to the implementers may not be received when the contractor is actively implementing the program	Real-time feedback can be provided to the program implementer
Data Analysis	Completed after program implementation (Often a year after)	Hourly and sub-hourly data available and with high accuracy. Savings as low as 5 % can be verified using utility bills.

Research Efforts

The use of M&V 2.0 for evaluating energy efficiency programs is still in the development stages, and further research is required before universals standards and protocols can be developed, specifically:

- More pilot studies or real-life programs that compare advanced M&V tools to traditional measurement and verification/evaluation need to be completed to demonstrate real-life examples.

- Ability to test M&V tools and software results against a set of consistent performance criteria or standards. Guidelines need to be developed to validate the outputs from models used in M&V 2.0 to estimate energy savings.
- Expansion of methods or tools to handle baselines other than existing conditions. Current models or tools only use the actual baseline metered data. They lack the capacity to account for an adjusted baseline (for example, federal standards, or replace-on-failure/burnout measure, where the pre-existingequipment is not the appropriate baseline), which is vital for evaluation purposes.
- Availability of more granular data results in utilities having to deal with large sets of data, and hence more resources are needed to maintain data privacy and cybersecurity. Moreover, access to this data by third-party evaluators or implementing contractors need to be scrutinized to maintain data security.

10. Appendix D: Principles and Types of Experimental Design

This appendix presents the principles and types of experimental design and provides key considerations to keep in mind while designing an experiment.

Principles of Experimental Design

Three conditions must be met to conclusively prove that a behavioural-based program has caused a change in behaviour (for example, the use of best practices in the design and installation of HVAC systems):

- The behavioural intervention has to precede the behavioural change in time.
- The behavioural intervention must be correlated with the behavioural change – that is, when the intervention is present, the behavioural change occurs, and when it is not present, the behavioural change does not occur.
- No other plausible explanations can be found for the behavioural change other than the intervention.

An experiment is an actively controlled testing situation designed to fulfill these conditions. In an experiment, the researcher controls the circumstances so that the outcome (for example, a behavioural change) cannot occur before the causal mechanism is presented, the objects on which the intervention is supposed to operate are observed with and without the treatment, and efforts are made to ensure that other plausible explanations for any changes in the objects of study have been eliminated.

The simplest kind of experiment involves observing behaviour before and after exposure to a treatment (for example, a training program). This is known as a pre-test – post-test (or pre-post test) design. This kind of design is seldom employed because of weaknesses, which are included in the text box below: Factors to Consider - Threats to Internal Validity. However, it is useful as a framework for discussing the sources of inferential error that can arise when certain critical elements of experimental design (for example, randomization of exposure to experimental treatments) are ignored.

FACTORS TO CONSIDER – THREATS TO INTERNAL VALIDITY

During a pre-post test experiment, a number of things can happen that can result in changes in an outcome variable of interest (e.g., specified size of an AC unit) that are not a direct consequence of the treatment (e.g., training). The change in outcome variable of interest may look for all intents and purposes exactly like an effect that might have arisen from the treatment, but not be caused by it. For example, in a simple comparison of annual kWh before and after exposure to a given training process, there are a number of possible alternative explanations for differences that might be observed besides the effect of the training mechanism, including the following:

- **History** – when a difference in behaviour is observed between two points in time, it is quite possible that the difference has been caused by some factor other than the experimental treatment variable. Weather is an example of a variable that might cause a difference in the application of an HVAC installation procedure, since air flow testing cannot be conducted when the ambient temperature is less than 20°C. So, depending on the timing of the experiment, the effects of weather might mask the effect of the treatment or lead to a belief that training had an effect when it did not. But weather is only one of many historical factors that could change and produce observed differences in behaviour variables between two points in time, either masking effects that are attributable to the intervention or producing effects that look like the effects of the intervention but are not.
- **Maturation** – when a difference in behaviour is observed at two points in time, the subject of the observation has gotten older and it is possible that some influence regarding the aging process has caused the change in the behaviour that is observed, and not the treatment. Maturation can influence behaviour in different and subtle ways. For example, sales and installation technicians are naturally gaining experience during and after the time they receive training. Over the whole population of interest, this aging process in the population may produce an increase or decrease in the use of various installation practices or the resulting energy consumption of their installations that could mask an otherwise observable effect of training or produce an effect that looks like some behavioural change that might have resulted from training, but did not. It is possible that the observed difference before and after training is nothing more than the effect of increased experience that would have occurred with or without the training.
- **Testing** – when a difference in behaviour is observed at two points in time, it is possible that the testing process itself has altered the situation. When humans are involved in experiments, they sometimes react to the measurement process in ways that produce the appearance of a change in behaviour resulting from treatment. An example of such a testing effect is what is known as a Hawthorne effect – named for a famous operations research experiment in which worker productivity increased significantly when better lighting was installed not because of the lighting

improvement, but because the subjects knew they were being observed. Testing effects can arise any time humans are aware they are being observed; and it is unusual for experiments with humans to be undertaken without them being aware of it. They are particularly likely to occur with repeated measures in which it is possible for subjects to learn the correct answers during the testing process.

- **Instrumentation** – when a difference in behaviour is observed at two points in time, it is possible that the calibration of the instruments used to measure the behaviour has changed – producing the appearance of a behaviour change that is nothing more than slippage in the calibration of the measuring instrument. Calibration problems can occur with all kinds of instruments. For example, if mechanical meters are changed to advanced meters during the course of an experiment, the improvement in the accuracy of the new meters will create the appearance of a change in behaviour (for the worse). Calibration problems are even more likely to occur with survey instruments and other self-administered behavioural measures. Minor changes in instrument design between time periods of observation can produce apparent (reported) differences between observations taken at different points in time that are solely due to respondents' interpretation of survey semantics or to the insertion of questions that alter the interpretation of questions seen later in the survey instrument.
- **Statistical Regression** – when a difference in behaviour is observed at two points in time, it may be that measurements taken in a second time period are different and closer to the statistical mean of the overall population than the initial, pre-treatment, measurement. This difference can create the suspicion that an effect occurred as a result of the treatment or it can cause the effect to be masked. While statistical regression can affect any sort of pre-post measurement it is not likely to seriously influence measurements of behaviour change related to training.
- **Censoring** – censoring is similar to maturation except the observed effect of the experimental condition arises from the fact that some subset of a group of observations is not observable at the second time period (the post-test) for reasons unrelated to the experimental condition. For example, in an experiment involving training, it is common for a certain percentage of trainees to move or withdraw from the training between initial assignment to treatment conditions and observation of the behaviour of interest after exposure to the treatment. This causes the measurement of the outcome variable to become censored in the post-test period for a subset of the customers. If the group that has withdrawn from the experiment is different from the remaining group on factors related to the outcome measurement of the study (for example, younger and less experienced technicians are more likely to be laid off during a downturn), this difference may produce the appearance of a change in behaviour when nothing more than censoring has occurred.

- **Selection** – this occurs when groups for which a comparison is being made (experimental vs. control) are significantly different before the treatment group is exposed to the experimental variable. In this case, there is no basis to infer that the treatment was solely responsible for the differences observed after exposure to the treatment. The most effective way of guaranteeing the assumption that the groups are similar is to randomly assign subjects to treatment and control groups. However, as will become apparent below, because it will often be impossible to randomly assign consumers to treatment and experimental groups in training experiments, selection is potentially a very important source of inferential error that must be controlled in experiments involving capacity building.

The weaknesses of the pre-post test occur because conditions other than the treatment can cause changes in behavioural outcome measures (for example, installation practices or annual energy consumption) when the effect is measured by comparing observations of a single group at two points in time (for example, before and after exposure to training or support).

It is possible to eliminate these problems by changing the design of the experiment so that instead of comparing the reactions of a single group of subjects (for example, trainees, consumers or organizations) at two points in time, the impacts of the experimental variable are observed by comparing the behaviours of two different groups of subjects – one group exposed to the treatment and the other not exposed. If the groups are similar, they will experience the same history, mature in the same way, react to testing and instrumentation in the same manner, and experience the same censoring. In other words, all of the possible challenges mentioned above will affect both groups in the same way. The only difference between the groups will be the treatment and it therefore can be considered to be solely responsible for the observed difference in behaviour. In doing so, the threats to the experimental validity described above will be eliminated.

Certainly, the assumption that both groups are similarly questionable. The drawback to inferring cause from differences between groups is that the groups may not have been the same, to begin with. If they were not, then any observed difference between them could simply reflect the pre-existing difference.

If left uncontrolled, threats to the internal validity of experiments are plausible alternative explanations for why a difference might be observed at two points in time (before and after exposure to an experimental condition) for a single group, and for why a difference between two groups exposed to a given experimental condition might occur. Establishing experimental procedures that ensure internal validity is a critical requirement in experimentation. Experiments that are not internally valid (for example, methodologically flawed) are generally not useful because they do not conclusively show that the experimental variable is the sole cause of a change in the outcome variable. They can lead to more damaging outcomes if the results confirm some prior expectation of the result, and therefore, are readily accepted without additional verification.

There are four basic “building blocks” of experimental design. They are control, stratification, factoring and replication. Taken together, these building blocks form a solid basis for constructing experiments designed to assess the extent to which a policy intervention has altered behaviour in a desired manner. They are discussed in the information text box below.

“Building Blocks” of Experimental Design Control

Control is central to the design of experiments. By taking control of the timing and exposure of subjects to experimental factors thought to change behaviour, it is possible to ensure that the experimental factor occurs before the onset of the desired behaviour. Aside from the possibility that some other causal mechanism occurs at precisely the same time as the experimental factor, controlling the administration of causal factors makes the inference about the primacy of the experimental factor more or less unequivocal.

Factors that are thought to cause changes in behaviour can be controlled in a variety of ways to observe their effects. Oftentimes, causal factors are treated as binary variables – they are either present or they are not. Sometimes they can take on a spectrum of values that may have different consequences for behaviour (for example, one might imagine training programs targeted at the same audience lasting different periods or being presented in different formats). As a result, it is possible to imagine experiments that range from very simple comparisons between the behaviours exhibited by just two groups, to experiments which contain numerous levels of exposure to an experimental factor.

A critical aspect of control in any experiment is the process used to assign customers to treatment and control groups or groups exposed to different levels of the treatment variable. When groups are compared to observe the effect of a treatment, the most fundamental assumption is that the groups are sufficiently similar at the outset of the experiment. Thus, any difference after exposure to the experimental factors can be deemed to have resulted from the factor, not a pre-existing difference. By controlling the assignment of experimental subjects to treatment and control groups (or different treatment levels), one can ensure that the groups assigned to experimental conditions are statistically identical before the experimental factor (treatment) is presented. Typically, this is done by randomly assigning subjects to comparison groups (for example, treatment and control groups or levels of treatment). This occurs because the random variable, by definition, is extremely unlikely to be correlated with any other variable.

Stratification

In evaluating the impacts of a behavioural intervention on energy consumption related behaviour, it is often useful to observe the effects of the experimental treatment for different sub-groups or market segments. For example, in studying the effects of training, it might be useful to observe the magnitude of the effect of the training for different trades (for example, sales technicians and installation technicians). Breaking up experimental groups (treatment and control groups) into sub-groups based on criteria that are observable in advance of an experiment is called stratification. The table below describes a simple experiment involving stratification on trade.

	Training	No Training
Sales Staff	n1	n3
Installers	n2	n4

In addition to providing useful information about the effects of experimental treatments within the sub- populations of interest (for example, sales staff and installers), stratification can be useful for reducing the amount of statistical noise that is present when one is attempting to observe a behavioural change (particularly energy consumption) between treatment and control groups. This is so because it is possible to reduce the variation in the measurements of the treatment and control group measures by observing the behavioural change within the sub-groups – ignoring the differences between the sub-groups.

Factoring

Sometimes behavioural interventions consist of treatments that contain more than one factor. For example, it is often the case that behavioural interventions intended to change energy consumption contain a technology component (for example, a field computer or device that simplifies the application of a given installation protocol) and an information component (for example, training designed to encourage the application of best practices). In assessing the impacts of such a combined treatment, it is necessary to structure the experiment in such a way as to allow for the estimation of:

- The interaction between technology and the training in changing the behaviour of the subjects under study. An interaction is a situation in which the presence of one factor multiplies the effect of the other. For example, an interaction between technology and training would be present if the effect of these two factors taken together was greater than the effect that would occur if their individual effects were just added together.
- The main effects of the treatment variables (for example, technology and training). The main effect of a treatment is the effect that occurs solely as a result of exposure to the treatment variable alone – separate from any impact that might occur as a result of combining that treatment with some other factor.

Typically, an experiment involving factoring is described as a matrix with the row and column variables containing the different levels of the treatment variables. The table below describes a simple factoring experiment in which two treatment variables, with two levels, are examined.

	Technology	No Technology
Training	n1	n3
No Training	n2	n4

In the experiment, subjects would be randomly assigned to one of four groups (n1 to n4) in sufficient numbers to be able to estimate the differences in the outcome behaviours of interest among the various groups.

The difference between stratification and factoring is that stratification is simply the creation of test groups that are different in meaningful ways at the outset of the experiment. Factoring involves the exposure of experimental subjects to different levels of treatment variables that have been nested to allow the estimation of treatment effects within levels.

It is possible to combine stratification and factoring to create very complex experiments that can isolate the effects of experimental variables for different sub-populations. The appeal to create such complicated experiments involving many factors and strata should be approached cautiously because of the inherent difficulties encountered in carrying out such complex experiments.

Replication

Perhaps the single most important tool for evaluating the impacts of behavioural interventions is replication. Replication occurs when the conditions involved in an experiment are repeated to confirm that a result which has been reported can be repeated by a different investigator, in a different setting, at a different time and under different circumstances. If the reported effect can indeed be repeated, there is a reason to be confident that the reported result is robust and did not arise by accident or because of something the investigator did that was not reported in the results of the study.

While replication is seldom described as a consideration individual investigators should study while designing evaluations, it is a very powerful tool that should be used to assess the veracity of research findings at the program level. In evaluations of behavioural interventions, investigators should be encouraged to structure their studies in such a way as to produce replications. It is particularly useful in situations where multiple experiments can be carried out in different geographical locations (for example, among the various LDCs implementing programs) sequentially or simultaneously. Evaluators carrying out behavioural experiments across multiple LDCs should be encouraged to design their experiments as replications of a single administration.

True Experiments

True experiments are research designs in which the evaluator has control over the exposure of experimental subjects to treatments. There are three types of true experiments – Randomized Controlled Trials (RCT), Randomized Encouragement Designs (RED) and Regression Discontinuity Designs (RDD). These research designs provide the most robust tests of the impacts of behavioural interventions on energy use related behaviour. They are discussed below.

Randomized Controlled Trials (RCT)

The RCT is an evaluation research design in which experimental subjects are randomly assigned to treatment and control groups. The results observed for the groups are compared to discover whether the treatment has caused a behavioural change. The process of random assignment causes the resulting groups to be statistically identical on all characteristics before exposure to the treatment within a known level of statistical confidence, given the sample sizes being employed. This is true because each observation assigned to both groups has the same probability of being assigned to each group (for example, 1/n; where n is the number of total subjects being assigned.) The mathematical consequence of this assignment constraint is that the treatment and control groups will be more or less statistically identical after the assignment process is complete. That is, the groups will contain about the same percentage of males and females, have the same average age, come from the same geographical locations, have about the same amount of prior years of experience – and so on, for virtually all the variables one can imagine – whether we can observe these variables or not.

Naturally, because sampling is involved, the above statement is true to the extent that relatively large samples are involved but only to within a certain level of statistical confidence. Indeed, anything can happen in the real world – which means that even with a truly random assignment with large samples, it is possible to create treatment and control groups that are not statistically identical. It is a common practice to ensure the groups that will be studied in an RCT are indeed more or less identical, at least on the outcome variable before they are administered the treatment. It is also advisable to obtain and include pre-test measurements for both the treatment and control groups on the outcome measures of interest to control for any pre-treatment differences that may occur on the outcome variable of interest.

RCT designs are often referred to as the "gold standard" of research designs to be applied to observing behavioural change. Reasons underlie this designation are:

- **Validity** – an RCT controls for most of the above described threats to internal validity – most importantly for selection bias or the possibility that the groups under study were somehow different before the experimental factor was presented.
- **Simplicity** – analyses of results obtained from RCT designs are simple and straightforward and do not rely heavily on assumptions about the specification of estimation equations or error structures. They are often as simple as a difference in differences calculations. Consequently, the estimated impacts derived from studies employing RCTs do not depend heavily on the skill or artfulness of the analyst.

- **Repeatability** – because these designs are relatively simple, it is possible to accurately recreate the conditions under which observations were taken, thereby making replication easy.

Despite these obvious advantages, there are several aspects of RCT designs that require caution in the application. First, the assignment of subjects to experimental treatments does not guarantee that the groups that are eventually observed in an experiment are equivalent. There are two easy ways in which the initial random assignment may be invalidated during an experiment. They are:

- **Volunteer Bias** – randomly assigning subjects to treatment and control groups in which treatment group members must agree to participate after assignment can result in treatment and control groups that are very different. This is the essence of selection, so care must be taken to ensure that significant numbers of randomly assigned subjects do not migrate out of the study between the time they are randomly assigned and the time the results of the treatment are observed. If subjects must volunteer for the treatment or acquiesce to it, then random assignment to treatment and control groups should occur after they have volunteered or agreed to be in the study.
- **Rejection** – human subjects virtually always have the right to withdraw from a treatment to which they have been experimentally assigned. They may withdraw for reasons that are unrelated to the experimental treatment, or they may withdraw because of the treatment. In either case, out-migration from the treatment and control groups may invalidate the effect of the initial random assignment, and care must be taken to ensure that observations for out-migrants are properly handled. If the number of customers who reject the treatment becomes large (for example, more than 1 or 2 percentage points) then it is necessary to analyze the results of the experiment as though it was a RED design.

When regulatory policies or concern about customer experience prohibit the arbitrary assignment of subjects to experimental conditions, it may still be possible to randomly assign customers to treatment conditions by using one of the following research tactics:

- **Recruit and deny** – experimental subjects are recruited to an experiment with the understanding that participation is not guaranteed (for example, is contingent on winning a lottery). In such a situation, subjects are told that the experimental treatment is in limited supply and that they will be placed in a lottery to decide whether they will receive it. The lottery winners are chosen at random, and winners are admitted to the treatment group, while losers are assigned to the control group. Losers may be offered a consolation prize to reduce their disappointment in not being chosen for the lottery. As long as the transaction cost involved in participating in the lottery are not too high, this strategy can overcome objections that stakeholders may have to randomly assign subjects to test conditions. This approach is particularly useful when the experimental treatment (for example, an attractive new technology) is in limited supply so that it can be argued that the fairest way to distribute the benefit is to distribute it randomly among interested parties.

- **Recruit and delay** – like the recruit and deny design experimental subjects are recruited to an experiment with the understanding that participation in the first year is contingent on winning a lottery. The lottery winners are chosen at random, and winners are admitted to the treatment group in the first year. Losers are assigned to a control group that is scheduled to receive the treatment in the second year. This approach can be implemented without causing significant customer dissatisfaction. However, because the control group must also receive the treatment in the second year, it will result in a higher cost for equipment and support than the recruit and deny approach.

Randomized Encouragement Designs (RED)

Sometimes regulatory or administrative considerations require that all subjects who are eligible to receive some behavioural intervention must receive it if they desire it. For example, an administrative policy might dictate that all qualified HVAC technicians have access to training that would result in their receiving a certificate that can provide a competitive advantage or may be required to provide certain contracting services. In such a situation, it is virtually impossible to deny some contractors access to the supposed behavioural intervention to create a legitimate control group.

It is possible to create a legitimate randomized experiment when all parties in the market must be eligible for treatment by employing what is known as a Randomized Encouragement Design (RED). In a RED design, the treatment (for example, training program) is made available to everyone who requests it. However, while all contractors are eligible for training, a subset of the eligible contractors is randomly chosen to receive significantly more encouragement for seeking the training than the control group, which is not encouraged. If the demand for the training is relatively low (in the absence of encouragement) it may be possible to significantly increase the rate of exposure to the training among volunteers in the encouraged group by more intensively marketing the training program to them. The encouragement might include more intensive efforts to contact and recruit contractors, providing economic incentives for participation, or reducing transaction costs associated with subscribing to the treatment.

The impact of the treatment is estimated by comparing the outcome variable of interest for the randomly selected encouraged group with the same outcome variable for the randomly selected group that was not encouraged. This comparison is referred to as an intention to treat analysis, as it focuses on the measurement of the difference in the behaviour between those who were intended to be treated and those who were not. Because encouragement was randomly assigned, any difference between the encouraged and not encouraged groups must necessarily have resulted from the fact that the encouraged group contains more parties who received the treatment. Since the acceptance rate in the encouraged group is known, it is possible to inflate the observed difference between the outcome of interest in the encouraged and not encouraged group to obtain a reliable estimate of the average impact of the treatment on those who received it.

The analysis of the impact of the encouragement and treatment is straightforward algebra, and the results are easily explained. So, one is tempted to conclude that the RED design is the best option for overcoming the difficulties that are often cited with the application of RCT designs in evaluations related to energy consumption behaviour. Unfortunately, this is not the case. As in the case of the RCT design, certain cautions must be observed when implementing a RED design.

First, the RED design rests on the assumption that the only factor that is influenced by the encouragement applied to the encouraged group is the acceptance of the treatment. While it is difficult to imagine circumstances in which encouragement to participate in a training program or receive organizational support would result in other actions than changed behaviour or energy consumption, it is logically possible that encouragement stimulates some other actions that either enhance or attenuate the observed effect of the treatment. This possibility should be considered in deciding whether to employ a RED design.

A second and more important caution in applying RED designs arises out of the likely increase in the sample sizes required to detect effects using a RED design. In a RED, the measurement of the impact of the treatment on behaviour is diluted because some (in many cases most) of the parties who were encouraged to be treated did not accept the treatment. So, only a small portion of the subjects who are encouraged to be treated may accept it. Nevertheless, they are counted as intended to be treated. The larger the fraction of the group that was intended to be treated that does not receive the treatment, the more muted the measurement of the treatment effect will be, and vice versa. For example, if 5% of the population normally accepts the treatment without encouragement and 20% of the population accepts the treatment with encouragement, then it can be said that the encouragement has significantly increased the rate of acceptance of the treatment. However, the impact of the treatment on the outcome measures in the encouraged group will be based on the responses of only 20% of subjects who received the treatment. If the actual behavioural change for individuals receiving the treatment is 1 unit, then the difference that will exist between the encouraged group and the not encouraged group will be only 0.2 units. This mathematical fact imposes powerful limits on the usefulness of RED designs. Depending on the magnitude of the targeted behavioural change and the effectiveness of encouragement, the RED design may require much larger sample sizes in the treatment groups than the conventional RCT. In cases where the effect of the treatment on behaviour and the acceptance rate for the treatment are in the single digits, the sample sizes required to detect the resulting difference between the behaviour in the encouraged and not encouraged groups may be so large as to be practically impossible to observe.

In most cases, with training programs that involve, at most, hundreds of subjects, the usefulness of RED designs will depend heavily on the ability of evaluators to develop effective encouragement. Even then, these designs should only be used when relatively large impacts on behaviour and energy consumption are expected.

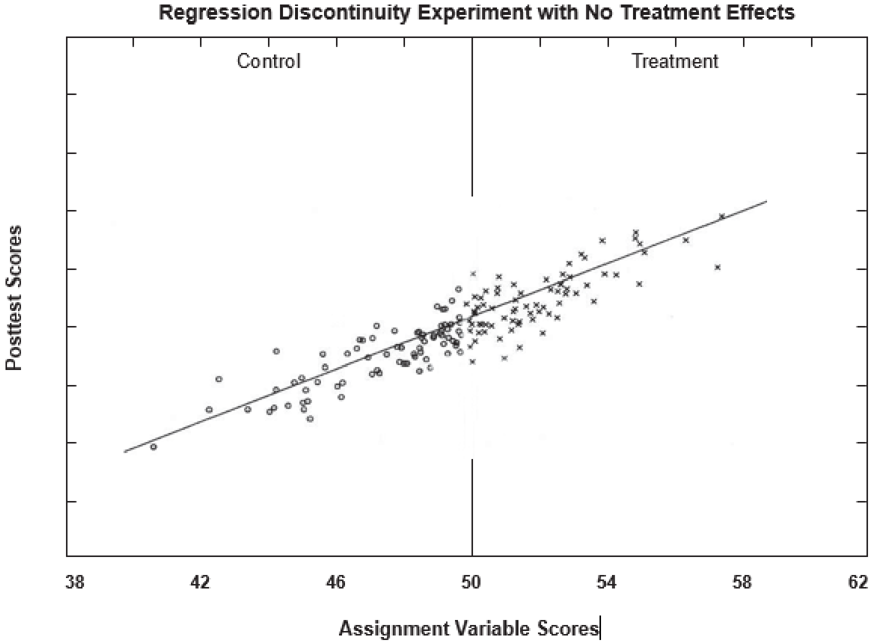
Regression Discontinuity Designs (RDD)

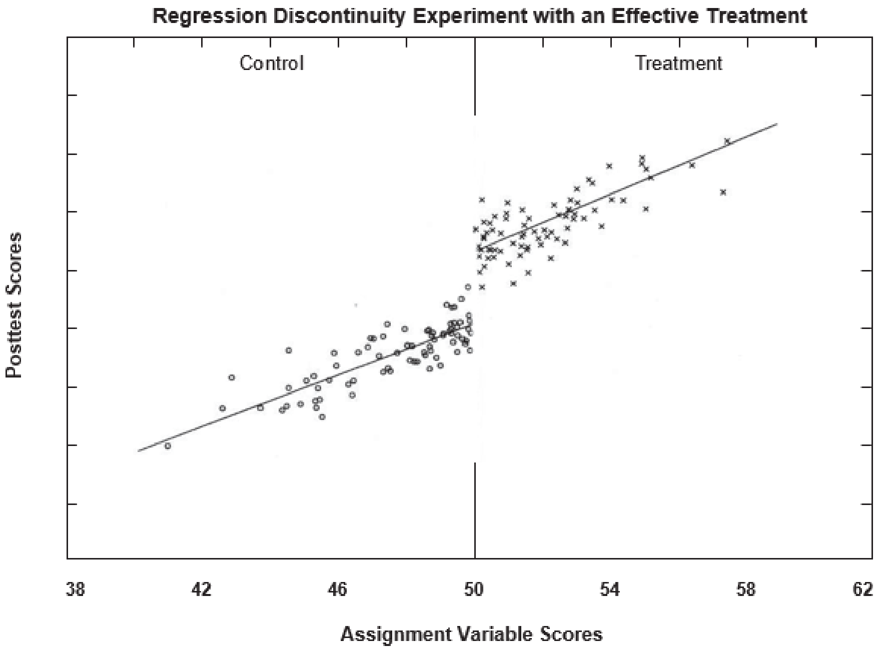
In the two true experimental designs discussed above (RCT and RED), subjects are randomly assigned to experimental groups – thereby establishing their statistical similarity. Under certain circumstances, the assignment of subjects to treatments can be non-random, where provided subjects are assigned to treatment and control groups precisely based on their score on an interval level variable such as age, years of experience, number of annual installations completed, etc. Such an experiment is called a Regression Discontinuity Design (RDD). In an RDD, everyone above or below some point (the discontinuity) on the selected interval scale is assigned to the treatment group, and everyone else is assigned to the control group.

It is possible to specify a regression equation describing the relationship between the assignment variable and the outcome variable of interest in the experiment. It might be that the outcome measure increases with the value of the assignment variable, decreases with it, or does not vary systematically with the outcome variable at all. It does not matter. It can be shown that the RCT is just a special case of the RDD, where an assignment variable is a random number. For example, everyone above a certain point on the random number distribution is assigned to the treatment group and everyone else to the control group.

The impact of the treatment variable in an RDD is observed by examining the regression function at the point at which the assignment was determined. The figure below displays an example of a regression discontinuity analysis. The top panel of the figure displays the relationship between the assignment variable and the outcome variable for the experiment when no effect is present. The assignment in this example takes place at the scale value 50. In the top panel, the regression line continues unperturbed at the assignment value, as indicated by the vertical line in the center of the plot. There is no discontinuity indicating that there is no difference between the treatment and the control groups.

The bottom panel shows what the regression line might look like if the treatment caused a change in the outcome variable of interest. In such a situation, there is a visible discontinuity at the point on the assignment scale at the value of 50. The difference in the post-test score values at the intersection of the two regression lines depicted in the bottom panel is the effect of the treatment. This effect is illustrated in the figure on the next page, by the difference in the horizontal axis between the projections of the two intersection points on the vertical discontinuity indicator





The RDD is an extremely powerful tool that can be used when subjects must be assigned to treatment conditions based on some pre-existing qualifications. It controls all possible alternative explanations for the observed program effect. However, certain important caveats must be met to justify using this design:

- Assignment to the treatment must be strictly determined by the assignment variable. Even the slightest deviation from this requirement will undermine its validity.
- Care must be taken to remove any crossovers among experiment subjects from the analysis. For example, sometimes parties will migrate into the treatment group from the control group and vice versa.
- Care must be taken to ensure that the functional form of the regression is correctly specified. If the relationship in the estimated regression is specified as linear, but the underlying, predicate relationship is not, the regression discontinuity analysis may incorrectly interpret the point of inflection on the non-linear function as a discontinuity. This will result in a serious estimation error.
- Likewise, if the treatment interacts with the assignment variable so that the slope of the regression line changes at the assignment variable due to the treatment effect (causing a jackknife shaped function), and the function is not properly specified as such. This will cause a serious error in which the effect of the experimental treatment will be underestimated. Protecting against this possibility requires estimating non-parametric (nonlinear) regression functions, which imposes additional complexity.

Quasi-Experiments

It is not always possible to control the assignment of observations to treatment and control conditions. Often, evaluators are given the task of evaluating the impacts of a behavioural program after key marketing and enrollment decisions have been made. It is also impossible to use true experiments when treatment condition of interest is compulsory (everyone is required to be exposed to the treatment), or when observations have the ability to select whether or not they are subjected to the experimental condition.

When the assignment to the treatment condition is not under the control of the experimenter, the design of experiments is much more complicated than it is with true experiments. When observations are randomly assigned to treatment and control conditions or assigned based on a pre-existing interval level variable, as is the case with the true experiments, all plausible alternative explanations (for example, history, maturation, etc.) for an observed effect are logically and mathematically eliminated. When this is not so, it is necessary to structure the experiment/analysis in such a way to observe whether these alternative explanations are plausible, measure their magnitude, and if possible, control for them analytically. This is the domain of quasi-experiments.

It should be clear that the decision to abandon random assignment can have profound consequences for the internal validity of the experimental design. It places a much heavier burden on the researcher to show that the study’s findings are not the result of some unknown and uncontrolled difference between the treatment and synthesized control groups. It can be the first step down a slippery slope that leads to an endless and irresolvable debate about the veracity of the study’s findings.

Several types of quasi-experimental designs are particularly important in behavioural experiments involving training. They vary according to their robustness (the extent to which they can achieve the credibility of a random experiment) and difficulty in their execution. They are:

- Non-equivalent control groups designs
- Within subjects designs
- Interrupted time series designs

Non-equivalent Control Groups

In true experiments, subjects are assigned to treatment and control groups in such a way that they are either known to be statistically identical before exposure to the treatment factor (as in the case of the RCT and RED designs) or are different in a way that is perfectly measured and thus capable of being statistically controlled. It is not always possible to implement true experiments for reasons already discussed, and for cost and practical reasons, it may be necessary to select control groups after the subjects to be treated have been selected. These are called non-equivalent control group designs. They are called non-equivalent control group designs because the estimates of the impacts of treatment factors from such designs rest on a comparison of treated subjects with subjects who are identified in such a way that we can never be certain that they are truly equivalent to the treatment group subjects. The results obtained from non-equivalent control group designs are analyzed in the same manner as they are with true experiments.

The objective of a non-equivalent control group design is to identify a control group of subjects that is as similar as possible to the treatment group based on pre-existing information we have about parties who are eligible for the treatment. Non-equivalent control groups are created by selecting control group members from the same population (for example, firms, business types, markets, regions, cities, trades, etc.) from which the treatment group came based on their similarity to members in the treatment group.

This is done by a process called matching. Matching is an old concept, and dozens of slightly different matching procedures have been tested over the past several decades. Matching is a highly controversial procedure for developing control groups because it is impossible to guarantee that a matching effort, no matter how sophisticated, has successfully created a control group that is similar to the treatment group in all important respects.

Recent professional practice favours the use of what is called propensity score matching – a procedure that attempts to match control observations with treatment observations based on an estimate of the probability that subjects were selected for (or selected themselves into) the treatment group. This technique requires estimation of the probability of selection into the treatment group using a logit regression model containing as many known predictors of treatment group participation as can be found.

In simple terms, a logit model is a type of regression model designed to predict the probability that an event happens (for example, signing up for training) based on information about readily observable independent variables that may be correlated with selection into the treated group (for example, firm size, years of experience, expressed interest in training, etc.). Once the parameters in the logit model have been estimated, members of the treatment group and other subjects who are not part of the treatment group are assigned propensity scores based on their characteristics and the model parameters. Treatment group subjects and others are then matched according to the values of those scores. Once matching has been completed, the results from the treatment and control groups in the experiment are analyzed in the same manner in which the results from true experimental designs are analyzed.

Matching methods by themselves are to be used with caution because they are prone to the introduction of bias that cannot be anticipated or measured. However, compelling the results based on experience, intuition, or other indicators of a treatment effect, an experiment involving non-equivalent control groups does not provide incontrovertible evidence that the observed effect is attributable solely to the treatment. That said, this may be all that is possible under some circumstances.

Within Subjects

All of the preceding experimental designs rest on the comparison of the behaviour exhibited by groups of subjects who have been exposed to treatment with the behaviour exhibited by groups that have not been exposed to a treatment (control groups). The difference between the behaviours exhibited by the two groups (exposed and not exposed) reflects the effect of the experimental treatment.

The principal threat to the validity of such designs is the possibility that the groups were different in some way that produced the appearance of a treatment effect when one did not exist. In the true experiments, this threat to validity is eliminated by controlling the assignment to treatment and control groups in such a way as to ensure that the comparison groups are statistically identical or different in ways that are known with certainty. However, it is not possible to control for this possibility when non-equivalent control groups are used as the standard of comparison. That is, it is always possible that non-equivalent control groups are different from the treatment groups in some important way before the onset of the experimental treatment. This problem is inherent in the comparison of treatment and control groups to infer the effect of the experimental treatment.

Under some circumstances, it is possible to avoid this problem. The solution rests in comparing what happens to experimental subjects in the presence of and in the absence of treatment. That is, it rests on observing the effect of the treatment by comparing the behaviours exhibited by experimental subjects before the treatment is presented and after, or when it is at high levels vs. low levels. In this way, the subjects in the experiment serve as their own control group. This experimental design is called a Within Subjects design.

The defining characteristic of a within subjects design is that each experimental subject is exposed to all levels of the experimental factors under study as well as the absence of the experimental factor (for example, the control condition). Under the appropriate conditions, this is a very powerful quasi-experimental design, since it eliminates the possibility of selection effects because it eliminates the control group.

Interrupted Time Series

Another quasi-experimental design that is appropriate to studies of the impact of behavioural interventions on energy consumption related behaviour is the interrupted time series design. An interrupted time series design consists of repeated measures of the behaviour of interest before and after a treatment has been administered. This design is particularly useful when variables related to usage or other frequently measured behaviours are under study – thereby creating the opportunity to observe the time series of measurements.

The basic idea behind interrupted time series designs is that if the onset time of the treatment is precisely known, it should be possible to observe and quantify a perturbation in the time trend of the outcome variable (energy use related behaviour) after the onset of the treatment. In other words, there should be a measurable change in the functional relationship between the treatment and the outcome variable after the treatment is started. In a sense, this is analogous to regression discontinuity, where time is the selection indicator. This design depends on several important considerations:

- The onset time of the treatment can be definitively established (for example, it is known that treatment commenced abruptly at a certain time

- The effect of the treatment must be large enough to rise above the ambient noise level in the outcome measurement (time series data often contain cycles and random fluctuations that make it difficult to detect subtle effects of time trend influences)
- If the treatment is expected to have gradually impacted the outcome of interest, the time series before and after the treatment must be long enough to reflect the change in the intercept or slope of the outcome variable after the treatment has occurred
- The number of observations in the series must be large enough to employ conventional corrections for autocorrelation if the statistical analysis is required (as it almost always is)

Like all comparisons that rest entirely on observing the difference in behaviour before and after exposure to treatment, the interrupted time series designs are subject to several weaknesses that can undermine the validity of the inference that observed change has been caused by the experimental treatment. Most important among these weaknesses is the possibility that the observed change in the intercept or slope in the time series may have been caused by something other than the treatment (for example, an exogenous but contemporaneous factor with historical antecedents). It is also possible that some aspect of the testing process that is coincident with the delivery of the experimental factor is responsible for the observed change (for example, a Hawthorne effect). To control for such intervening explanations, a variety of quasi-experimental control techniques can be employed, including the use of non-equivalent control groups as described above, adding non-equivalent dependent variables (for example, other variables that are expected to be impacted by the same historical forces as the dependent variable but not the treatment factor), and manipulating the presentation of the treatment factor (adding and removing it) to observe the impact on the outcome variable. The latter is only appropriate when the effect of the treatment factor is expected to be transient. In the parlance of statistics, these designs are a type of within subjects or repeated measures design.

11. Appendix E: Substantiation Form

This appendix includes an example of a substantiation form, which is the Measures and Assumptions Substantiation Form used by the IESO. Evaluators are encouraged to use the template, or at least consider it as a guideline upon submission.

Name of Measure

Measure Description

Energy Efficient Equipment Description

Enter base EE description.

Base Equipment Description

Enter base case equipment description.

Code, Standards and Regulations

Enter any codes, standards and resulations that may be applicable to the measure.

Resource Savings Assumptions

Measure Assumptions

Base Measure Demand Assumptions

Base Measure	kW
	--- [1]

Energy Efficient Measure Demand Assumptions

Conservation Measure	kW
	--- [2]

[1] Reference 1

[2] Reference 2

Assumed Hours of Operation

Lighting hours are based on facility types as follows:

Facility Type	**Assumed Annual Operating Hrs**
Lighting - Food Retail	6074
Lighting – Hospital	5182
Lighting - Large Hotel (Corridor/Lobby)	7884
Lighting - Large Non-Food Retail	4089
Lighting - Large Office	3610
Lighting - Nursing Home	4308
Lighting - Other Commercial Buildings	2857
Lighting - Other Non-Food Retail	4089
Lighting – Restaurant	5110
Lighting – Schools	2596
Lighting - University Colleges	3255
Lighting - Warehouse Wholesale	3759

Energy and Demand Savings[3]

Energy Savings

Annual Energy Consumption (kWh/yr) [base] = Base Measure Wattage x Operating Hours

Annual Energy Consumption (kWh/yr) [conservation]

= Conservation Measure Wattage x Operating Hours

Annual Energy Savings (kWh/yr) = Annual Energy Consumption (kWh/yr) [base]

- Annual Energy Consumption (kWh/yr) [conservation]

Lifetime Energy Savings (kWh) = Annual Energy Savings (kWh/yr) x EUL (yr)

[3] Reference 3

Annual energy Savings (kWh) are estimated based on facility type. Refer to the MAL excel worksheet for the annual energy savings of each conservation measure for different facility types.

Connected Demand Savings

Demand Savings (kW) = Base Measure Wattage – Conservation Measure Wattage

Measure Name	Demand Savings (kW)

Summer Peak Demand Savings

Summer Peak demand savings ΔkW_{peak} are calculated by multiplying the Annual Energy Savings ΔkWh with the Summer Peak Demand Factor from the Energy Load Profile. Refer to CE Tool for the formatted load shapes and peak demand factor end use:

$$\Delta kWpeak = \Delta kWh * \text{Summer Peak Demand Factor}$$

End Use Load Profile

Indicate the name of the end use load profile for this measure plus the EM&V factor (only the 4 values, not including the 8 buckets).

EM&V Peak Definition

Summer Peak Demand (kW/kWh)	Winter Peak Demand (kW/kWh)	Alternative Summer Peak Demand (kW/kWh)	Alternative Winter Peak Demand (kW/kWh)
---	---	---	---

Effective Useful Life (EUL)[4]

Measure	EUL

Incremental Costs[5]

Measure	Life cycle incremental cost
	---

[4] Reference 4

[5] Reference 5

Other Resource Savings

Enter Other Resource Savings here.

Revision History

Revision #	Description/Comment	Date Revised
---	---	---

12. Appendix F: Example Logic Model

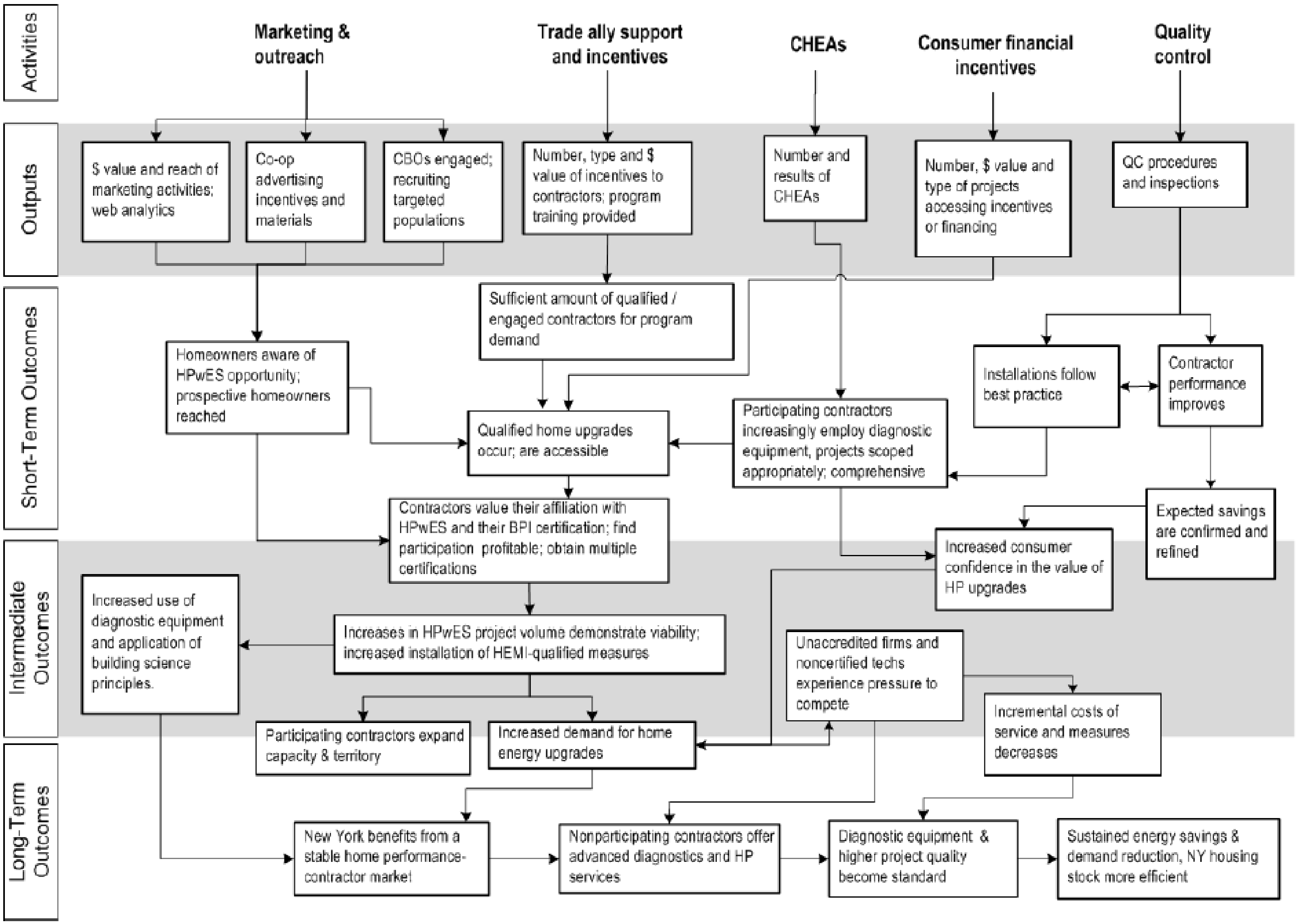
Logic models address program activities, outputs and outcomes. An example of a program logic model is the logic model developed by the New York State Energy Research and Development Authority (NYSERDA) for the Home Performance with ENERGY STAR program. The process flow diagram of the logic model is provided on the next page, and a detailed description of the program logic model is provided in NYSERDA's final report[16].

Definition of acronyms used in the logic model:

CBO:	Constituency-based organization
CHEAs:	Comprehensive home energy assessments
QC:	Quality control
HPwES:	High performance with ENERGY STAR
BPI:	Building Performance Institute
HP:	High performance
HEMI:	High efficiency measure incentive

[16] NYSERDA (2014). *Home Performance with ENERGY STAR Logic Model – Final Report.* Website: https://www.nyserda.ny.gov/-/media/Files/Publications/PPSER/Program-Evaluation/2014ContractorReports/2014-PLM-Home-Performance-Energy-Star.pdf

Figure 12-1 | Logic Diagram


Activities
Outputs
Short-Term Outcomes
Intermediate Outcomes
Long-Term Outcomes
Marketing & outreach
Trade ally support and incentives
CHEAs
Consumer financial incentives
Quality control
$ value and reach of marketing activities; web analytics
Co-op advertising incentives and materials
CBOs engaged; recruiting targeted populations
Number, type and $ value of incentives to contractors; program training provided
Number and results of CHEAs
Number, $ value and type of projects accessing incentives or financing
QC procedures and inspections
Sufficient amount of qualified / engaged contractors for program demand
Homeowners aware of HPwES opportunity; prospective homeowners reached
Qualified home upgrades occur; are accessible
Participating contractors increasingly employ diagnostic equipment, projects scoped appropriately; comprehensive
Installations follow best practice
Contractor performance improves
Expected savings are confirmed and refined
Contractors value their affiliation with HPwES and their BPI certification; find participation profitable; obtain multiple certifications
Increased consumer confidence in the value of HP upgrades
Increased use of diagnostic equipment and application of building science principles.
Increases in HPwES project volume demonstrate viability; increased installation of HEMI-qualified measures
Unaccredited firms and noncertified techs experience pressure to compete
Incremental costs of service and measures decreases
Participating contractors expand capacity & territory
Increased demand for home energy upgrades
New York benefits from a stable home performance-contractor market
Nonparticipating contractors offer advanced diagnostics and HP services
Diagnostic equipment & higher project quality become standard
Sustained energy savings & demand reduction, NY housing stock more efficient

Independent Electricity System Operator
1600-120 Adelaide Street West
Toronto, Ontario M5H 1T1

Phone: 905.403.6900
Toll-free: 1.888.448.7777
E-mail: customer.relations@ieso.ca

ieso.ca

@IESO_Tweets
facebook.com/OntarioIESO
linkedin.com/company/IESO

ieso
Connecting Today.
Powering Tomorrow.