Evaluation, Measurement and Verification Protocols and Requirements V3.0

Interim Framework 2019-2020

April 1, 2019

ieso
Connecting Today.
Powering Tomorrow.

Preface

The Evaluation, Measurement and Verification (EM&V) Protocols and Requirements V3.0 consists of three volumes:

Volume I : EM&V Protocols and Requirements

Part 1 – Developing, Procuring and Reporting on Evaluations (Audience: Evaluation Administrators)

Part 2 – Conducting Evaluations (Audience: Evaluation Contractors)

Volume II : Protocols for Evaluating Behavioral Programs

Volume III : CVR Impact Evaluation Protocols

Acknowledgements

The Independent Electricity System Operator (IESO) would like to recognize the Efficiency Valuation Organization (EVO) who developed the International Performance Measurement & Verification Protocol (IPMVP). Their work serves as a valuable reference and foundation upon which the "EM&V Protocols and Requirements V3.0" are developed.

Readers wishing more information on program evaluation methods can access the library of materials available from the US Department of Energy and the Office of Energy Efficiency and Renewable Energy at: https://www.energy.gov/eere/analysis/program-evaluation

Table of Contents

Tables and Figures

Figures

Tables

Abbreviations

CDM	Conservation Demand Management
CMVP	Certified Measurement and Verification Professional
CVRMSE	Coefficient of Variation of the Root Mean Squared Error
DEP	Draft Evaluation Plan
ECM	Energy Conservation Measure
EM&V	Evaluation, Measurement and Verification
EUL	Effective Useful Life
FEP	Final Evaluation Plan
IESO	Independent Electricity System Operator
IPMVP	International Performance Measurement and Verification Protocol
LDC	Local Distribution Company
M&V	Measurement and Verification
MAL	Measures and Assumptions List
NEB	Non-Energy Benefit
NTG	Net-to-Gross
NTGR	Net-to-Gross Ratio
O&M	Operating and Maintenance
OEB	Ontario Energy Board
RFP	Request for Proposal
TOU	Time of Use
TRC	Total Resource Cost
VORL	Vendor of Record List

Introduction

Document Introduction

Thank you for your interest in the 2019 – 2020 Interim Framework Evaluations, Measurement and Verification (EM&V) Protocols and Requirements V3.0 (the Protocols).

EM&V is critical in establishing Conservation and Demand Management (CDM) as a credible and reliable "first choice" resource in meeting future electricity supply needs of Ontario. EM&V provides information to decision-makers, system planners and program administrators for use in developing long term demand/supply plans, to maximize program performance, and to determine whether energy savings and demand reduction targets are being met.

The EM&V Protocols and Requirements V3.0 helps program and evaluation administrators create and manage objective, high quality, independent, and useful conservation program evaluations. It provides an administrative protocol; governing the "who," "how," "what," and "when" of EM&V. In addition to what has been described above, the "why" is to ensure that the Province and all market players can depend on CDM as a resource. Supporting technical guides, aimed primarily at independent Evaluation Contractors, cover off the remaining "how" elements of completing a high quality evaluation.

Intended Audience

There are two main audiences for this document:

- **PART 1** is intended primarily for Evaluation Administrators who are charged with managing the program evaluation process
- **PART 2** is intended primarily for Evaluation Contractors, though the information is valuable to Program Administrators as well.

The document is also a resource for program design, as it is important to have a general understanding of evaluation methodologies so that programs are designed in a manner that allows for impacts to be measured and evaluated.

Background

Across North America, increased attention is being devoted to program evaluations. Today, more than ever, increased scrutiny of government spending and rising energy prices require a prudent review of program investment. As such, linking program resource expenditures with program results has become a necessity.

In general, program evaluations include market assessments, process evaluations, retrospective outcome/impact assessments and cost-benefit evaluations. These types of evaluation studies help

program administrators/program managers determine what adjustments are needed in the program offer to enhance programmatic achievements relative to the committed resources.

Program evaluations are in-depth studies of program performance and customer needs. The benefits of conducting an evaluation are numerous, including:

1. Helping Evaluation Administrators and Program Managers estimate how well the program is achieving its intended objectives;
2. Helping administrators and managers improve their efforts; and
3. Quantifying results and communicating the value of program efforts amidst a multitude of regional, regulatory, and legislative priorities

Intended Use

The EM&V Protocols and Requirements V3.0 are intended for use by CDM market players in the Province of Ontario who have an interest in CDM Program Design, Delivery and Evaluation. The protocols provide guidance for a robust evaluation, listing guidelines and general instructions. They identify the practice required to evaluate, measure and verify energy savings and demand reductions associated with CDM activities in Ontario. They are not intended for training, nor as an assurance of flawless evaluations. Still, by following these protocols, the appropriate regulatory agencies and administrative agencies can have confidence that each evaluation served is identifiable and comparable to the others using similar processes.

The different types of evaluations require data-collection and analysis methodologies with which some Evaluation Administrators will have little familiarity. It will not be necessary to have in-depth working knowledge of the many methods available. It is highly advisable to have some familiarity with basic evaluation techniques so that selecting and monitoring an Evaluation Contractor is possible, since they will recommend and implement specialized analytical methods.

While the value of program evaluation is well established, the questions of who should do what, how (rigour level and consistency) it should be done, and when (rapid versus after-the-fact feedback as well as recurring studies) are far less well defined. EM&V protocols are intended to address the following key issues:

- The need for separation between the department responsible for program delivery and the department responsible to assess program performance to realize credible and effective evaluation.
- The proper allocation of EM&V costs; typically higher for more project-based evaluations or pilots and typically lower for larger, ongoing programs.
- The proper attribution of savings, when results from multiple evaluations have to be credibly tabulated into a collective total by following common rules and processes.
- The appropriate use of ex ante input assumptions (e.g. the Measures and Assumptions Lists) during program planning, monitoring and evaluation.
- Procedures to identify and prevent duplication of evaluation efforts.
- The realization of "economies of scale" by evaluating similar initiatives and efficiency projects together, such that fewer individual and potentially inconsistent sets of results emerge at the end of a program cycle.

- How the five major streams of evaluation work may be combined or separated in various ways for efficiency and quality:
  - Outcome (summative; ex post; conducted to verify cognitive and behavioural changes).
  - Impact (summative; ex post; can include M&V engineering conducted for the purpose of developing new or improved ex ante savings estimates).
  - Process Assessment (develop conclusions about program performance; includes audits; can include behavioural research for the purpose of developing new or improved ex ante savings estimates).
  - Market Study (market characterization that can contribute to evaluating the impact of codes and standards, time-of-use rates, market transformation elements of efficiency programs and may also contribute to the development of ex ante savings estimates).
  - Cost Effectiveness (economic analysis that compares the benefits of an investment with the costs).
- How to incorporate the temporal element of moving from "Resource Acquisition" to "Market Transformation", using "Capability Building".
- To ensure a consistent approach to hiring and managing Evaluation Contractors across the Province.

The entire EM&V effort is used to develop a reliable net savings estimate—those savings attributable to or resulting from program-sponsored efforts as distinguished from savings that would have occurred anyway, be that from individual behavioural choice, public acknowledgement, or from naturally occurring market adoption.

Presentation of Information

This document takes a process-driven approach in presenting the information. The information is presented as a series of steps an Evaluation Administrator would take in managing the evaluation process, from designing evaluations, to hiring Evaluation Contractors, to reporting evaluation results. Of course in the real world, the process is not purely linear – many steps are interrelated and, to some, degree the process is iterative.

Structure of the Document

The document is divided into two sections:

PART 1: DEVELOPING, PROCURING AND REPORTING ON EVALUATIONS

Part 1 guides an Evaluation Administrator through the first 12 steps in the overall EM&V process: from documenting a program's market strategy, hiring an evaluation contractor and managing and publishing the evaluation results.

PART 2: CONDUCTING AN EVALUATION

Part 2 is intended primarily for Evaluation Contractors, but it is also a useful reference for Program Administrators, providing them with a high level understanding of the technical processes required to

carry out the evaluation. Part 2 contains 10 Technical Guides.

Evaluation Administrators need a high-level understanding of the work the Evaluation Contractor is undertaking, therefore it is recommended that Evaluation Administrators also become familiar with the techniques and methods outlined in Part 2.

EM&V Protocols and Requirements V2.0 (2015-2020) vs. EM&V Protocols and Requirements V3.0 (2019-2020)

This document replaces the previous version of the EM&V Protocols and Requirements V2.0 (2015-2020), with an enhanced version that provides additional guidance and clarification on how to undertake an evaluation for energy efficiency and behavioral programs and the addition of conservation voltage protocols.

Part 1:

Developing, Procuring And Reporting Evaluations

Audience: Evaluation Administrators

The *2019-2020 Interim Framework EM&V Protocols and Requirements* helps Program Administrators and Evaluation Administrators create and manage objective, high quality, independent, and useful conservation program evaluations. This Protocol was developed for all staff who plan, commission, and manage program evaluation services across the province.

In the most general sense, Evaluation Administrators are persons or organizations responsible for evaluating energy efficiency, conservation, or demand response initiatives. In the EM&V context, Evaluation Administrators are those who are specifically responsible for designing and implementing the Evaluation Measurement and Verification Plan (EM&V Plan) of energy efficiency, conservation, and demand response initiatives.

Part 1 guides the Evaluation Administrator through the initial steps that lead to conducting the agreed on evaluations by an Evaluation Contractor. The Evaluation Administrator will employ industry best practices for procuring an Evaluation Contractor and working with the selected Contractor to develop and implement the EM&V Plan. Evaluation Administrators are responsible for developing an EM&V plan for a particular program or portfolio. They are also the point-of-contact for EM&V Evaluation Contractors. Evaluation Administrators are sometimes referred to as Evaluation Managers. In general terms, these steps involve the following activities:

- **Hiring an independent, qualified Evaluation Contractor** – this involves inviting qualified vendors to bid on the project and selecting an appropriate contractor from among the bidders.
- **Coordinating Evaluation Contractor's activities** – this involves working with the Evaluation Contractor to determine the detailed research methods that will be used.
- **Managing the evaluation process** – this requires a combination of skills including: balancing resources, overseeing the flow of data and information between persons involved in the evaluation, ensuring quality control with regard to the work being conducted, and ensuring project timelines are satisfied.

The Management Board of Cabinet's Procurement Directive requires that the following principles guide the procurement process:

- **Vendor Access, Transparency, and Fairness:** The procurement process should be conducted in a fair and transparent manner, providing equal treatment to all vendors. Conflicts of interest, both real and perceived, must be avoided. Particular vendors should not be relied on continuously, or routinely be granted contracts, for a particular kind of work.
- **Value for Money:** Goods and services must be procured only after consideration of the business requirements, alternatives, timing, supply strategy, and procurement method.
- **Responsible Management:** The procurement of goods and services must be responsibly and effectively managed through appropriate organizational structures, systems, policies, processes, and procedures.
- **Geographic Neutrality and Reciprocal Non-Discrimination:** Entities subject to Ontario's Trade Agreements must be geographically neutral with respect to vendor access to government business.

- **Documenting the program strategy and offer** – this requires an understanding of the program's logic model.
- **Properly scoping the program evaluation** – this involves selecting elements of the program logic model to be evaluated and drafting the research questions.
- **Identifying analytical approaches to address research questions** – this requires exploration of factors potentially influencing the program and identifying key metrics for each program element to be studied.
- **Specifying evaluation deliverables** – this involves deciding on the frequency and timing of planned evaluations, specifying the primary analytical methods the Evaluation Contractor is expected to use, and creating a detailed timeline of project deliverables.
- **Creating the Draft Evaluation Plan** – the draft evaluation plan forms the basis for the scope of work that is set out in the Request for Proposals (RFP) process which is used to hire an Evaluation Contractor.
- **Assessing the reasonableness of the Evaluation Contractor's findings and conclusions** – this involves linking conclusions to findings and providing context for findings.
- **Publishing the evaluation report** – this includes explaining the evaluation results and providing recommendations on enhancing the program.

The draft EM&V plan defines the Evaluation Contractor's scope of work. When procuring an Evaluation Contractor, the Evaluation Administrator must balance product quality, reliability, and pricing. The following factors will come into play when selecting an Evaluation Contractor:

- Selected areas of study
- Choice of analytical methods
- Availability of staffing
- Timing of evaluation tasks
- Data collection and analysis requirements
- Competitiveness of the offer

Evaluation Administrators and Program Administrators should expect the Evaluation Contractor to propose a variety of approaches for carrying out the work. Given the nature of research, an EM&V plan developed by an Evaluation Administrator is always a draft, with specific research activities developed after work begins and uncertainties managed to achieve the desired levels of precision and accuracy based on the facts revealed.

Step 1: Document Market Strategy and Program Offer

Key Points / Highlights

Documenting Market Strategy and a Program's Offer involves the following tasks:

1a. Specify Market Needs

1b. Identify Program Strategy

1c. Tabulate Impact Forecasts

1d. Highlight Program Benefit-Cost Ratios

Task 1a: Specify Market Needs

To plan a program's evaluation one needs a good understanding of the program. As such, the program description should include discussion of relevant market conditions and the needs of targeted stakeholders.

Given that the purpose of a program is to cause change in the market, the program description should point out key market hurdles and barriers. The descriptions should include a table that identifies and distinguishes between:

Market Hurdles – these are temporary obstacles that discourage the adoption of desired behaviours. A well-designed program can, in the short term at least, directly influence market hurdles such that changes to behaviour can occur. For consumers in the business sector, an example of a market hurdle is the payback period or return-on-investment thresholds for investing in energy-efficient equipment. For individual consumers, a market hurdle could be the price of energy efficient appliances. With such hurdles, a financial incentive could help the consumer overcome this one-time investment hurdle.

Market Barriers – these are on-going obstacles that prevent adoption of desired behaviours. A well-designed program can also directly influence market barriers, but it typically takes longer for change to occur with market barriers than with market hurdles. For schools, for example, a market barrier might be a lack of trained maintenance staff. If that's the case, a useful program design strategy might be to offer technical training for maintenance staff on energy savings strategies and practices.

The Evaluation Administrator and Program Administrator are both responsible for properly classifying targeted market opportunities as either market hurdles or market barriers.

A program's design reflects an underlying theory about how and why the program activities will achieve the desired results. In particular, the underlying theory illustrates how program activities will help participants overcome one or more market barriers or hurdles, thereby leading to the adoption of energy efficiency or conservation measures.

Task 1b: Identify Program Strategy

Traditionally, programs were classified as having an underlying strategy that is either:

Resource acquisition – these programs address market hurdles and are characterized as involving the direct purchase of GWh or MW. (Programs based on this strategy are referred to as Resource Acquisition Programs). Or,

Market transformation – these programs address market barriers and are characterized as involving activities where GWh or MW savings are the logical extension of market-based outcomes. (Programs based on this strategy are referred to as Market Transformation Programs).

The Evaluation Administrator must identify whether the program strategy is resource acquisition or market transformation in nature.

Program strategies have evolved and now some programs are hybrids, meaning they include incentives aimed at overcoming market hurdles and producing short-term energy savings directly and they overcome market barriers, leaving market conditions that are favourable for continued realization of program impacts. Where a program involves a hybrid strategy, the Evaluation Administrator must identify which activities are associated with market transformation and which are intended for resource acquisition.

Regardless of the program type, Program Administrators should forecast the demand impact from the program. This information is required in order to address system reliability. Although the system peak demand savings of all programs offered will be assessed, outcome evaluations may also examine the other benefits. To ensure demand savings can be calculated using a variety of demand definitions, hourly load impacts should be produced to allow for flexibility. More details about calculating demand savings are in **Technical Guide 6: Demand Savings Calculation Guidelines.**

Supportive and technical guidelines on the following are included in Part 2 of this document:

- **Technical Guide 5:** Gross Energy Savings Guidelines
- **Technical Guide 6:** Demand Savings Calculation Guidelines
- **Technical Guide 7:** Market Effects Guidelines

Task 1c: Summarize Budget Allocation

The program description should include a summary of the spending on program activities. In short, Program Evaluations focus on the largest program expenditures or on where the largest program impact is forecasted. A simple table showing the budget allocation per class of activity is necessary to address the Program Manager's level of commitment to the program strategies chosen.

Tabulate Impact Forecasts

For example, lighting programs make broad assumptions about existing measures: their age, use, and condition. These assumptions change the forecasted energy profile of the lighting measures being removed. Similar assumptions alter the forecasted energy profiles of replacement equipment. The mean difference between these two forecasted energy profiles represents the forecasted energy and demand savings expected for the program.

Task 1d: Highlight Program Benefit-Cost Ratios

Program cost-effectiveness has broad implications for program planning, design, and implementation. The Program Administrator should develop reasonable forecasts of program costs and savings, making sure that cost-effectiveness screenings fairly represent the anticipated ratio of program costs to benefits. Where verified benefit streams or real costs differ significantly from those forecasted, the Evaluation Contractor should note critical variances and offer conclusions about their impact on program theory. As well, the Evaluation Contractor should make recommendations regarding ways of resolving large differences. Moving forward, the Program Manager will be expected to use this information to narrowing these variances.

When preparing program cost-effectiveness one should consult the cost-effectiveness policy and procedures explained in **Technical Guide 2: Cost-Effectiveness Guidelines**

To help optimize implementation effectiveness, cost-effectiveness studies may be done with regard to specific program activities or with regard to particular measures. These studies can be valuable at the early stages of a program offer, or after program processes have been significantly altered.

Summary of Actions

- Classify targeted market opportunities as either market hurdles or market barriers
- Identify whether program strategy is resource acquisition or market transformation in nature
- If hybrid strategy involved, identify activities associated with market transformation and resource acquisition
- Include summary of spending on program activities
- Indicate anticipated level of demand and energy savings expected
- Report cost of conserved energy (and cost of demand savings, if demand savings program)

Step 2: Anticipate Program Causes and Effects

Key Points / Highlights

Anticipating Program Causes and Effects involves the following tasks:

2a. Summarize Resources Available for the Program

2b. Categorize Planned Program Activities

2c. Specify Expected Return on Program Investments

2d. Highlight Potential Outcomes Resulting from the Program Offer

2e. Specify the Desired Impacts from the Program Offer

2f. Illustrate and Annotate Program Logic

2g. Verify Savings Attribution Pathway

Task 2a: Summarize Resources Available for the Program

Programs allocate resources in an effort to cause energy and demand savings. In the explanation of the theory on which a program is based, program administrators must specify the resources available (namely, the monies and time allocated to the program) to achieve the desired effects.

Capital is money allocated to fund specific program activities and administrative expenses, including money allocated directly to program service provisioning.

While capital covers the majority of program funding, other contributions, such as in-kind contributions for infrastructure and staff, may add significantly to the program offer without affecting the budget allocation. Where in-kind contributions are relevant to achieving energy and demand savings one should identify them as key resources available to the program.

Non-capital funding – Examples of important non-capital sources of program funding:

Infrastructure (in-kind) – business and information systems a sponsoring organization may provide to operate the program, such as the organization's procurement services center, its billing information systems, or its training facilities.

Human (in-kind) – expertise and support staff offered by an organization to help deliver a program without a direct budget allocation, such as utility account representatives, training center staff, marketing professionals, etc.

Strategic Relationships – ties or relationships sponsoring organizations may have with vendors that may provide time or expertise without significant added cost.

Task 2b: Categorize Planned Program Activities

Program activities are generally categorized based on the nature of their intervention into the marketplace. Here are the main categories into which program activities usually fall:

Financial Assistance – this is the payment of cash to encourage customers to engage in desired behaviour. Financial assistance may include direct financial incentives, rebates, or in-store discounts. Other financial assistance in the form of financing, guarantees, or price buy-downs may also be used.

Technical Assistance – these are when services are offered to buyers of energy efficiency measures or channel partners. This assistance may be consulting services, training courses, or access to help lines. The goal of technical assistance is to facilitate the introduction, installation, or maintenance of energy efficient technologies within the market.

Informational and Educational Materials – this is basically materials focused on communicating technical information, or information about technology options, end-use applications, or emergent practices. The materials can be bill inserts, information brochures, client testimonials, booklets, radio spots, exhibition booths, websites, smartphone apps, etc. The form of media is less important than the message included: namely, technical information rather than promotional material.

Promotional Materials – materials aimed at encouraging program uptake using media to highlight a program's presence within a market. Often promotional materials are not considered part of the planned offer program. However, Evaluation Administrators and Program Managers should insist on including them in order to show how promotional activities contribute to changing market attitudes that then lead to changes in behavior and energy demand.

Task 2c: Specify Expected Return on Program Investments

Monies paid for goods and services that result in Program Outputs are program expenditures. Program Outputs are the most direct returns that can be measured from program expenditures. Program Managers must highlight on a Program Logic Model the Program Outputs that:

- lead to outcomes along the "Critical Savings Attribution Pathway" and
- involve the expenditure of a significant amount of program resources that have been expended (regardless of their contribution to energy and demand savings). (See **Task 2f: Illustrate and Annotate Program Logic** for examples)

Where possible, Program Managers should specify an average cost per unit of Program Output.

Program Outputs are basically the tangible results achieved by a program. Program Outputs are monitored based on metrics the Program Manager establishes, such as: participants served, the number of end-use measures installed, the number of workshops held, pass/fail rates from training programs, etc. It should be noted that though program outputs are critical to a program's success, they are only intermediaries that demonstrate resource allocation and contract compliance.

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

Task 2d: Highlight Potential Outcomes Resulting from the Program Offer

Program Outputs should lead to some anticipated market change. These changes are themselves outcomes that Program Managers must include in the Program Logic Model. The cognitive, structural, and behavioral outcomes necessary to achieve demand and energy savings must be distinguished in the Program Logic Model, along with other market changes that bring about the desired program impacts.

Types of Outcomes

Cognitive Outcomes: Changes in attitude of people and organizations as a result of a program. Such changes can be reflected in learning, knowledge or understanding, perception, outlook, ambition, desire, etc. They are changes in mental abilities or perceptions that influence people and cause them to change their behaviour in a desired way.

Structural Outcomes: Changes in the target market's ability to observe and/or adopt behavioural outcomes as a result of a program. These changes can be reflected in things like enhancement of skills, technological innovation, changes in market structure, increased fiscal support and other market-based changes that support the short-term, intermediate, and long-term abilities of market actors.

Behavioural Outcomes: Changes in behaviour as a result of structural or cognitive outcomes achieved by a program. These changes can be reflected in purchasing decisions, stocking practices, technology utilization, energy consumption, load shifting, etc. When assessing whether there have been behavioural changes as a result of a program one must be sure to filter out changes that might have occurred as a result of external influences.

Task2e: Specify the Desired Impacts from the Program Offer

For funded conservations programs the desired impact is usually demand and energy savings. However, governmental and sustainability initiatives may be part of a particular program, in which case societal impacts may come into play, such as job creation, emission credits, and so on.

The Evaluation Administrator must document the program demand and energy impacts, specifying the hours of demand reduction and the annualized energy savings. The Evaluation Administrator may also include other societal impacts (job creation, non-energy benefits etc.) in the evaluation, but the Evaluation Administrator must quantify the impacts using standards applicable in the particular industry. For example, an Ontario utility may wish to calculate emission credits associated with electricity demand reduction. To do so, the utility should apply the standards and protocols set out in the International Program Measurement and Verification Protocols (IPMVP) on emissions credits, which may require measurements before and after a retrofit. Furthermore, claiming and selling/assigning any emission credits to other organizations is subject to a complex and changing legal framework. As such, Evaluation Administrators must understand the protocols applicable to all impacts claimed.

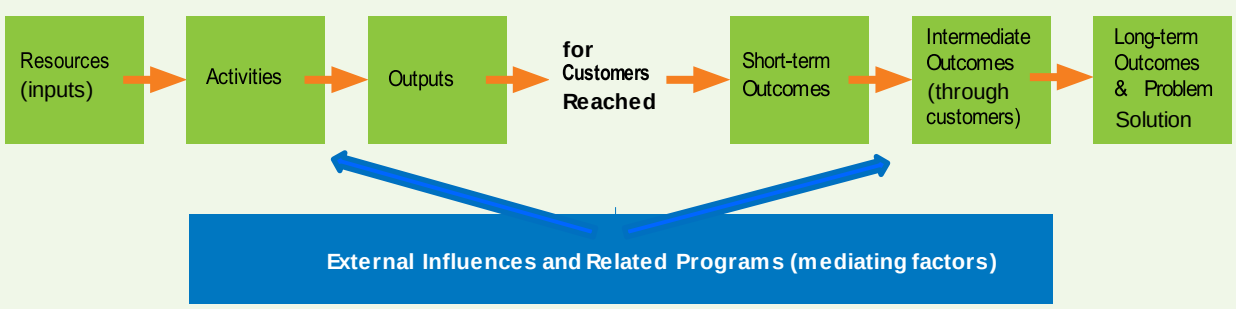
Task 2f: Illustrate and Annotate Program Logic

As noted, a program logic model is an illustration of program logic as a causal chain from resource expenditure to the long-term impacts of the program.

Figure 1.0: The Basic Elements of a Logic Model shows the basic elements of a program logic model. Crafting a good logic model requires that Evaluation Administrators and Program Managers think about what the program is attempting to achieve and what the causal chains are to achieve the desired outcomes.

The arrows linking program activities to outputs, outputs to outcomes, and outcomes to impacts represent the intended cause and effect relationships underlying the program. As such, these linkages must be explained in the EM&V Plans.

Figure 1.0: The Basic Elements of a Logic Model[2]



Task 2g: Verify Savings Attribution Pathway

The Evaluation Administrator should document the intended impacts of the program (reduced energy demand and savings.) and unintended impacts that may occur as a result of the program. For example, a residential demand response/load control initiative could provide a mechanism, for example, *a programmable thermostat,*) that if program participants use at times that are outside those expected (*an unintended impact*) as a result of their increased awareness (a cognitive outcome), and change their heating and air conditioning consumption patterns (*behavioural outcome*).. The primary reductions in peak demand that result from the thermostat (*intended impact*) are central to the initiative. The Evaluation Administrator should clearly identify at least one (if not more than one) pathway (referred to as an attribution pathway) leading from program resource expenditures directly to energy and demand savings.

Attribution Pathway: A relationship from one or more program-sponsored activities to outcomes and impacts being asserted by the Program Administrator or Evaluation Administrator. The pathway is a set of logical connections between resource expenditures and specific impacts so that cause and effect can be attributed to the program offer.

By identifying an attribution pathway, the connection between program intentions and verified program energy and demand savings, including unintended savings impacts, can easily be seen.

By exploring alternative hypotheses about how outcomes evolved one can identify questions about the potential effects of market externals that can be researched. The development of a logic model helps evaluators understand all of the possible ways the program outcomes might ripple through the targeted population. Ripple effects occur, for example, when people mimic desired actions without involvement in the program or as a result of previous participation in the program. Once such additional outcomes are identified, evaluators will know to ask questions about why they occurred. Without such investigation potential outcomes may go unnoticed and both direct and indirect outcomes that could add to program impacts may be missed.

Evaluation Administrators are encouraged to look for, and document, alternate pathways for demand and energy savings. Including these pathways within the logic model provides a means for claiming energy and demand savings that result from unintended, yet highly desirable, market behaviours.

Summary of Actions

- Specify resources (time and money) available to achieve desired effects
- Highlight on Program Logic Model: Program Outputs that lead to outcomes along Critical Savings Attribution Pathway
- Highlight Program Outputs involving significant expenditure of program resources
- Distinguish types of outcomes resulting from program
- Document program demand and energy impacts
- Document intended impacts of program
- Look for and document any alternative pathways for demand and energy savings

Step 3: Properly Scope Program Evaluation

Key Points / Highlights

Properly Scoping Program Evaluations involves the following tasks:

3a. Select Elements from the Program Logic Model to be Assessed

3b. Specify Types of Evaluation to be Completed

3c. Clarify Intended Use of Evaluation Findings

3d. Draft Research Questions

Task 3a: Select Elements from the Program Logic Model to be Assessed for the Evaluation

Budgets for evaluations are generally constrained. Therefore, staff will have to make choices regarding the scope of the evaluation.

Evaluation Administrators and Program Managers must choose which elements will be evaluated. The selection of elements should be based on the logic model created under Step 2. Depending on the size and magnitude of the evaluation, all or some elements in the Attribution Pathway (see **Figure 1.0**) can be included in the evaluation.

Task 3b: Specify Types of Evaluations to be Completed

When all the elements that will be included in the evaluation were selected, the evaluation objectives associated with the elements should be specified. In developing a statement of work for an Evaluation Contractor, the evaluation administrators should determine the types of evaluations that should be requested to ensure that the evaluation objectives can be met.

Types of evaluations include:

- **Outcome Evaluation** – this is conducted to verify cognitive and behavioural changes believed necessary for the realization of program objectives (outcome evaluations are summative and *ex post*).
- **Impact Evaluation** – this is conducted to measure the change in energy consumption or demand caused by the program (Impact Evaluations are summative and *ex post)*. Such evaluations can also include M&V engineering processes used for developing new or improved *ex ante* evaluation estimated savings.
- **Process Assessment Evaluation** – this is conducted to explain the program impact and/or identify lessons learned to inform future program strategies (in other words, to develop conclusions about program performance). Such assessments can include conducting behavioural research for the purpose of developing new or improved ex ante evaluation estimated savings.)

- **Market Study Evaluation** – the study of market characterization is conducted because it can contribute to evaluating the impact of codes and standards, TOU rates, and so on it can act as a benchmark for market transformation elements of efficiency programs and may contribute to the development of *ex ante* savings estimates.
- **Cost Effectiveness Evaluation** – a cost effectiveness evaluation includes "standard" cost effectiveness tests as provided in **Technical Guide 2: Cost-Effectiveness Guidelines**. Where the Evaluation Administrator or Evaluation Contractor deems it appropriate, it may also involve exploring the cost-effectiveness of individual measures, program elements, and/or implementation procedures.

Keep in mind that the analytical methods used in each type of evaluation will depend on the type of program evaluated. For example, program administrators will use a different analytical method for a demand response program impact evaluation and will report different information for such an evaluation than for an evaluation of an energy efficiency program.

When conducting evaluations, one must develop a robust analytical approach that yields statistically significant findings. Part Two of this guide provides guidance on the assessment of conservation programs. The manner in which a program is offered must be considered in the assessment. Therefore, all EM&V plans must provide a strategy that will result in evaluated savings estimates associated with the program.

When applicable, Evaluation Administrators must work with the Evaluation Contractor and apply the methods recommended in the Part Two. Programs must follow the guidance developed in the following sections:

- **Technical Guide 3: Process Evaluation Guidelines** are for all instances where a process assessment is sought or where concerns over operational efficiency have been expressed.
- **Technical Guide 4: Project Level Energy Savings Guidelines** are for single site implementation programs, such as those used for custom industrial process optimization.
- **Technical Guide 5: Gross Energy Savings Guidelines** are for most mass market energy efficiency programs and conservation initiatives.
- **Technical Guide 7: Market Effects Evaluation Guidelines** are for programs thought to change conditions, processes, or practices.
- **Technical Guide 8: Net-to-Gross Adjustment Guidelines** are for all savings claims and primarily used for energy efficiency programs.

Task 3c: Clarify Intended Use of Evaluation Findings

The Evaluation Contractor and Evaluation Administrator must understand how the evaluation will be used beyond the determination of verified savings estimates and must document these intended uses within the EM&V plan. For example, a program design team may commission a research study to assist in designing a program to estimate measure level effectiveness. The intended use of the evaluation findings will influence the evaluation plan and the manner in which the data is presented.

Task 3d: Draft Research Questions

Once evaluation objectives are established program administrators must convert them into general and specific research questions that then become the focus of the evaluation effort.

Program administrators should derive the general questions from the evaluation objectives. Each general question implies specific research questions that are capable of being answered through data collection and analysis.

Clear research questions help build consensus among evaluation stakeholders and offer guidance on the areas of investigation, which increases the likelihood of coming up with valuable evaluation findings, insightful conclusions, and useful program recommendations. Properly stated research questions:

(a) flow directly from the evaluation objectives

(b) are specific and solicit significant finding

(c) can yield answers that are actionable and

(d) are answerable within the constraints of the evaluation budget and other resources.

Keep in mind that for each research question there are distinct experimental considerations, such as the sample size and parameters, relevant comparison group, data collection methods, and so on. As a result, few research projects effectively answer more than a handful of research questions. The narrowing of research questions is a fundamental activity within EM&V planning and is necessary for a manageable evaluation. Evaluation Administrators should narrow the inquiry to less than a dozen, well-crafted research questions.

Summary of Actions

- Choose the elements to be evaluated
- Ensure evaluated savings estimates are provided rather than deemed savings estimates
- Convert evaluation objectives to general and specific research questions

Step 4: Identify Analytical Approaches to Address Research Questions

Key Points / Highlights

Identifying Analytical Approaches to Address Research Questions involves the following tasks:

- **4a.** Construct Chain of Logic Connecting Resource Expenditure to Program Impact
- **4b.** Explore Factors that May Influence Program
- **4c.** Document Market Conditions and Research Constraints
- **4d.** Specify the Populations of Interest and Sampling Strategy
- **4e.** Identify Key Metrics for Each Program Element to be Studied

Task 4a: Construct Chain of Logic Connecting Resource Expenditure to Program Impact

Evaluation administrators must convert the research questions developed in Step 3 into experimental inquiries to estimate demand and energy savings. In general, each research question will require verification of outputs and outcomes and quantification of impacts.

Converting the research question must be done by testing a series of research hypotheses along the "attribution pathway" (see **Step 2**) associated with each research question under investigation.

An example of a hypothesis often used in our industry is that a particular financial incentive caused the participant to adopt the particular energy efficiency measure. Like all hypotheses, that hypothesis may or may not be supported by evidence. Given that it is commonly accepted that some program participants would have adopted the particular measure without the incentive, it is clear that common hypothesis is not always supported. Still, the hypotheses may be supported more often than not. So, the attribution pathway is still valid, but only for a proportion of the participants

Evaluation Administrators and Program Administrators must not stop at an overly simple inquiry; instead they must validate the theory underpinning a program based on a continuous set of hypotheses along the attribution pathway. For the theory to remain valid, the hypotheses must be explicitly stated in the evaluation plan and tested using valid analytical methods.

Task 4b: Explore Factors that may Influence Program

Considering the unintended impact of external factors helps evaluators isolate and report on program cause and effect. Formalizing the consideration of unintended impacts of a program is necessary to attribute impacts to specific program offers and to allocate savings.

Examining external, non-program factors that might influence an expected outcome can reveal non-program relationships and suggest alternative hypotheses about how outcomes occur. The process of examining the underlying theory, making the logical relationships explicit between the program components, and considering external influences can suggest the need for changes to a program's design or the evaluation plan.

Task 4c: Document Market Conditions and Research Constraints

Deciding on the resources to dedicate to program evaluation involves simultaneous consideration of:

(1) the importance of the program decisions to which the evaluation will contribute (i.e. achieving CDM targets)

(2) the resources needed to satisfy the evaluation's objectives and,

(3) the resources the program can afford.

Where external influences prohibit the study of critical elements on which the program is based, the constraints prohibiting the analysis should be explicitly stated within the program evaluation plan. The rationale for doing so is not only for simple transparency; rather the reason is grounded in the fact that evaluation staff or contractors will likely see the importance of various elements of program theory the protocols require explicit disclosure of constraints to an area of relevant investigation.

Evaluation Administrator should narrow the areas of investigation before the evaluation contractor begins their work. Doing so after-the-fact can jeopardize the evaluators' autonomy to explore program cause and effect.

Task 4d: Specify the Populations of Interest and Sampling Strategy

Quantitative research aims to determine the relationship between one or more independent variables (for example, installation of program measures) and a dependent variable (for example, GWh savings) within a target group (for example, low-income households).

Evaluation Administrators may use either a descriptive or experimental study approach to determine the relationship between independent or dependent variables.

A **descriptive study** establishes only the association between variables, such as, the propensity for energy savings among program participants. An **experimental study**, on the other hand, establishes causality between installed energy efficiency measures and observed demand reductions.

In practice, true experiments are difficult to establish for CDM initiatives. So, the industry has adopted quasi-experimental approaches that accept market characterization and measure effectiveness testing that is commonly used to support or confirm findings from other evaluation efforts. Methods such as tabulating descriptive measurements and finding the statistical significance of a relationship between variables are usually not thought of as research designs, but in fact, the process of going from the results of these analytical procedures to answer evaluation questions involves hypothesis testing and, therefore, undergoes a similar process to research design.

If, however, a Program Evaluator needs to determine the proportion of a quantified outcome that can be attributed to the particular program instead of to external influences (that is, the Evaluation Administrator needs to conduct an impact evaluation), then one must use a credible research method. The method should allow to estimate what actions participants would have taken (outcomes) had the program not existed. The difference between what participants would have done and what they actually did, is the amount of the observed outcome that can be attributed to the program.

Evaluation research designs that allow Evaluation Administrators to make claims of effect are called "experimental" or "quasi-experimental" designs.

In the experimental method Evaluation Administrators must fully define the study and comparison populations. They must describe how to determine each sample group, the numbers included in the study, and the resulting precision expected. Unless an exception is granted (and exceptions are typically only granted for market effects), the confidence in the quantitative findings must be at least 90%.

Evaluated savings (as opposed to deemed savings estimates) must be provided, unless unique circumstance prohibit comparison group selection (for example, if evaluating a large industrial energy efficiency program and similar conditions or processes are unlikely to exist for comparison, or where there are no or limited comparison groups such as in new construction). In such cases, refer to the **Technical Guide 4: Project-Level Energy Savings Guidelines** or the **Technical Guide 9: Guideline for Statistical Sampling and Analysis.**

Task 4e: Identify Key Metrics for Each Program Element to be Studied

The research questions developed earlier will help prioritize the areas of study around essential program elements

Evaluation administrators must identify the sources of data for each question, along with alternative strategies for collecting data where data access or integrity may be suspect. Where there is a lack of data to calculate indicators for each program indicator, one must revisit the research questions.

In a separate table organize the program elements against the program theories. For each program element being studied identify the potential data source and collection method.

Summary of Actions

- Convert research questions in experimental inquires to estimate demand and energy savings
- Specify irrelevant assumptions to be excluded from investigation
- Ensure evaluated savings are provided
- Create table highlighting key metrics and linking them to relevant theories underlying the program.

Step 5: Specify Evaluation Deliverables

Key Points / Highlights

Specifying Evaluation Deliverables involves the following tasks:

5a. Draft EM&V Project Gantt Chart(s)

5b. Consider Cross-Cutting Approaches

5c. Identify Study and Comparison Groups

5d. Highlight Analytical Methods Expected

5e. Explore Data Collection Opportunities and Constraints

5f. Change in Hourly (8760s) Load Shapes Explored

5g. Formalize the Draft Evaluation Plan

Task 5a: Draft EM&V Project Gantt-type Chart(s)

As a result of working through the previous steps, the evaluation requirements have been defined. Using established project management techniques, the Evaluation Administrator must manage the delivery of requirements.

Evaluation deliverables must be depicted in a project chart (for example, a Gantt chart or something similar) showing the timing of each component of the EM&V project and resources related to each component. Evaluation administrators should show the types of evaluations to be completed over the course of the portfolio/program offer. Keep in mind that to show this, the chart may have to include timeframes beyond the program expiration date. For example, for weather-sensitive loads, the chart may have to show timelines that extend to 18 months or more, as utility data may need to be captured for one full year, with an additional six months required to analyze and report the final program year savings.

The Evaluation Administrator must decide on the frequency, duration, and timing of planned evaluations, as well as the types of evaluations that will be completed. The types of evaluations within the scope of the 2019-2020 program offerings are those described on page 29 **(Draft Evaluation Plan Template 2019-2020).**

Types of studies defined in **Task 3b: Specify Types of Evaluations to be Completed** should be represented as milestones on the project chart. The evaluation administrator must include details regarding each type of evaluation in the project chart, including the start and end dates of major deliverables related to the particular evaluations. Time should be allocated to each major deliverable within the scope of each evaluation including, among other things, the following evaluation activities:

- **Finalizing the Evaluation Plan** – The Evaluation Contractor who will conduct the actual evaluation may need to refine the Draft Evaluation Plan presented to them. When putting together the Final Evaluation Plan, be sure to leverage the Evaluation Contractor's experience and knowledge to ensure that the scope and resources dedicated to the evaluation are optimal and realistic.

- **Developing Data Collection Instruments** – Data collection instruments include surveys, field work, focus groups, etc. The Evaluation Contractor with assistance from the Evaluation Administrator must coordinate data collection from program implementers, utilities and program staff. Keep in mind that, depending on the data available, it may be necessary to allocate significant time and resources for developing data collection instruments throughout the evaluation process.
- **Collecting Field Data** – In-field data collection involves data about the relationship between the Program Administrator and its customers. Note that because such information can be considered sensitive (i.e. use of personal information), the Evaluation Administrator must monitor in-field data collection efforts. Field data can be quantitative (collected from metering studies, mystery shoppers, on-site inspections, etc) and/or qualitative (collected from focus groups, panel studies, process reviews, etc) Whether the data is qualitative or quantitative the collected information must be summarized without bias.
- **Presenting the Findings** – The dates at which the summary of findings will be presented to the Evaluation Administrator must be included on the project chart. These dates are often a couple of weeks after surveys, or at pre-defined periods before the preparation of the draft evaluation report. Evaluation Administrators must ensure the Evaluation Contractor presents a summary of its findings and supporting data in a timely, constructive manner.
- **Delivering the Draft Evaluation Report** – The project chart should specify the date when the first draft of the evaluation report is to be delivered. When setting this deadline it is critical to allow sufficient time for the program administrator and other interested stakeholders to internally review findings and results emerging from the draft evaluation reports.
- **Delivering the Final Evaluation Report** – The delivery date of the final evaluation report must be specified in the project chart.

Task 5b: Consider Cross-Cutting Approaches

Conducting multiple analyses or evaluations simultaneously is known as cross-cutting. Applying a cross-cutting approach can help optimize evaluations. For example, when one adjusts an end-use measure and that adjustment causes changes to another end-use measure, the resulting change is referred to as a cross effect. A cross-cutting approach can be used to analyze cross effects. Where the Evaluation Administrator thinks using a cross-cutting approach would be useful, the EM&V scope of work should explicitly state that the approach should be used.

Because different scenarios could theoretically result in either overstatement or understatement of program savings, the Evaluation Administrator must disclose how cross-cutting techniques will be used to optimize evaluation cost-effectiveness while adding to the reliability of evaluation findings.

Task 5c: Identify Study and Comparison Groups

Examples of When Cross-Cutting Is Useful

A lighting program may involve replacement of incandescent lamps with compact fluorescent lamps that can provide the same lumen output with greater efficiency. Installation of the compact fluorescent lamps also means that less heat would be emitted by the light source, which could have a positive effect on cooling loads (adding to efficiency gains when a space requires cooling) or a negative effect on heating loads (reducing efficiency gains in a residential single family home) if the installations occur in conditioned spaces. To account for these cross effects, cross-cutting analytical approaches 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 should be considered when evaluating demand response initiatives within the same sector. Evaluation contractors will often look only at the direct influence of one program on another where a participant in one program is screened for participation in another. By failing to use a cross-cutting approach in such a case, the Evaluation Administrator risks understating savings.

Comparison Groups

A brief description of the anticipated study group and comparison groups must be stated for each analytical approach that will be used in the evaluation. The Evaluation Administrator must explicitly state in the evaluation plan the need for a comparison group. Furthermore, the Protocol specifies the methods by which comparison groups are selected – the selection process should be conducted by the Evaluation Contractor. Where possible, the comparison group(s) should be representative of the study group. The EM&V plan must consider comparability between the study and comparison groups in a manner which result in statistical significant findings.

Task 5d: Highlight Analytical Methods Expected

The Evaluation Administrator develops a list of analytical methods to best achieve the defined objectives in the EM&V Plan. The Evaluation Administrator must specify the primary analytical methods the Evaluation Contractor is expected to use. For example, the Evaluation Administrator may specify that estimated program savings should be based on billing analysis rather than engineering models.

Furthermore, the EM&V plan must include information regarding the savings attribution model. Attribution models are used to define the process an evaluation will follow to determine whether energy and demand savings are due to program influence.

Task 5e: Explore Data Collection Opportunities and Constraints

The Evaluation Administrator must make clear to the prospective Evaluation Contractors what data will be available for analysis and the timing of data acquisition. And the Evaluation Administrator should ask Evaluation Contractors to propose 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 must highlight these constraints in the scope of work provided to the Evaluation Contractor.

If data acquisition constraints exist, they must not be allowed to affect evaluation practices and the integrity of an evaluation. Most Evaluation Contractors have encountered data constraints and have experience with similar analyses from which they likely can recommend alternatives for data collection.

Where the data constraints are expected to be persistent, the Evaluation Administrator must indicate the steps that are to be taken to ensure EM&V best practices are upheld. Timelines within which data constraints are required to be resolved must be set out in the EM&V plan and time should be built into future evaluation cycles, or at least discussed with the Evaluation Contractor, to ensure the constraints are resolved.

Task 5f: Explore Changes in Hourly (8760s) Load Shapes

With the introduction of smart meters to Ontario's residential sector, some LDCs have usage and demand data that can be analyzed as a part of the evaluation of load shapes. Evaluation contractors that have experience with load shape analysis can provide insight into how interval data can be used for program evaluation.

Given Ontario's electricity reliability standards, using interval data for load shape analysis may be much more illustrative of the achieved impacts than traditional annual estimates of demand and/or energy savings. As a result, when estimating demand and energy impacts, where appropriate, priority may be given to using interval data.

Task 5g: Formalize the Draft Evaluation Plan

Evaluation Administrators must create a Draft Evaluation Plan. The Draft Evaluation Plan, must conform to the specifications established in **Step 7: Evaluation Plan Development Guidelines**.

Summary of Actions

- Create project chart showing timing of each component of EM&V project and resources related to each component
- Decide on the frequently, duration, and timing of planned evaluations
- If cross-cutting techniques are used, disclose how they will optimize evaluation cost-effectiveness
- Specify the primary analytical methods Evaluation Contractor is expected to use
- Provide information about savings attribution model used
- If there are constraints related to data collection, highlight them in Evaluation Contractors scope of work

Step 6: Evaluation Classification Protocols

Key Points / Highlights

When undertaking a program evaluation, the following types of evaluation should be taken into consideration

- **6a.** Impact Evaluations
- **6b.** Process Evaluations
- **6c.** Market Effects Evaluations
- **6d.** Cost-Effectiveness Evaluations
- **6e.** Outcome Evaluations

Introduction

In the Draft Evaluation Plan, the Evaluation Administrator must specify the types of evaluations to be completed. Impact, Process, Market Effects and Cost-Effectiveness Evaluations are the most discussed evaluations for energy efficiency programs. Another type of evaluation is an Outcome Evaluation. Outcome Evaluations are often useful when there is a need to establish the cause of observed effects. Therefore Outcome Evaluations can be highly relevant to the research

Task 6a: Impact Evaluations

Impact Evaluations are assessments of both intended and unintended effects that can be attributed to a program, policy, or project. Impact evaluations are the most rigorous of all evaluations since the attribution chain must be established from program outputs through observed outcomes to the realization of tangible impacts. Such evaluations are most appropriately applied to those measures that have a direct causal impact, like the installation of insulation on building heating and cooling efficiency.

For an impact evaluation, the contribution of external factors toward the realization of desired impacts should be limited to factors that are reasonable and can be accounted for within the analysis. In the prior example of building insulation, the external factors are weather and the set point for the interior temperature. For weather effects we generally normalize to some long-term weather trend or establish a reference weather year. For participant behaviours we hypothesize and test whether the program under study substantively influences the behaviours of the target market (participants).

In general, an impact evaluation addresses the following question: *What are the verified quantifiable effects (impacts) attributable to the program?* For CDM initiatives, the primary impacts are energy (GWh) savings and demand (MW) reductions.

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 benefits (NEBs)?
- What is the direct impact of individual program activities on non-energy benefits?
- What is the magnitude of observed effects? What proportion of those effects can be attributed to the program?
- What key factors are responsible for the verified savings?
- 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?

Task 6B: Process Evaluations

Process Evaluations are assessments of program policies, procedures and practices, along with a review of organizational controls that contributed to their realization. Unlike management consulting mandates, which tend to be forward-looking, process evaluations are retrospective in nature.

Process Evaluations review practices that were implemented over the period under review, outlining the strengths and weaknesses of program processes and seeking opportunities for improved operational efficiencies.

Process Evaluations verify program expenditures, review the efficacy of the services provided by the program and document the resulting operational outputs to program objectives.

Examples of research questions used in Process Evaluations:

- Are program designs and supporting organizational controls adequate?
- Is the program producing the outputs intended?
- 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?

The Evaluation Administrator should work with the Program Administrator to re-state specific program concerns into researchable questions to be investigated by Evaluation Contractors. The following general questions are good examples to be reframed for a program Process Evaluation:

- Are program objectives set too high? Too low? What market actors are being served and through what delivery channels?
- Is it easy for customers to join or participate in the program? What motivates them to participate?
- Are the available tools and services supporting program delivery? Are the tools used properly by program delivery agents?
- Are customers participating at expected levels? Are some customer groups participating more than others? Why?
- Which tools and services are being used? By what groups? Are customers satisfied with the program?
- Are the resources assigned to the various program components adequate to achieve the desired objectives?
- Is the program leveraging available funds effectively? How could additional resources be applied? Are detailed program expenditure records maintained?
- How can the program better engage non-participants and hard-to-reach populations? What recommendations do participants and non-participants have for the program?
- Would administrative improvements better support the provisions of program services?

Task 6c: Market Effects Evaluations

Market effects evaluations assess the changes, due to program, policy, and projects, in both short-term and long-term structural elements of the market place, as well as the cognitive processes and behaviours of key market actors that lead directly to energy savings and demand reductions.

For resource acquisition programs, market effects evaluations serve to measure the net effect of programs by accounting for key major net-to-gross effects: spillover and free ridership. Market effects evaluation also seeks to attribute transformational impacts on the market resulting from application of codes and standards, legislation, innovation, and capability-building initiatives.

Evaluation Administrators should include market effects evaluations when Program Administrators suggest intended changes to target markets, or when they espouse a long-term approach with proposed exit strategies, or suggest that actors' behaviours will persist beyond the scope of the intervention.

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?

Task 6d: Cost-Effectiveness Evaluations

Cost-effectiveness evaluations measure the stream of benefits against the costs to achieve those benefits. In general, cost-effectiveness evaluations are implemented at the program level by leveraging industry-established tests. The details of the tests required in Ontario can be found in **Technical Guide 2: Cost-Effectiveness Guidelines**. Cost-effectiveness evaluations may also target measures, program delivery agents, and specific program activities.

Examples of research questions used in cost-effectiveness evaluations:

- How much did the verified energy savings and demand reductions cost to achieve?
- What benefits resulted from individual program activities relative to their costs?
- Was the program cost-effective? Does this program pass the cost-effective hurdles established for the Province of Ontario?
- Which delivery channels are working best to achieve program objectives?

Task 6e: Outcome Evaluations

Outcome evaluations are similar to market effects evaluations except that output evaluations do not link program expenditures to program impacts. Outcome evaluations are used to document causal linkages between program outputs and program outcomes or, to test elements of complex program theory.

Outcome evaluations are used to establish the efficacy of market transformational initiatives, policy directives, social programs and other interventions within a complex environment where direct impacts may be difficult to isolate from influences beyond those resulting from program-sponsored activities.

Examples of research questions used in outcome evaluations, often the first step in an impact assessment looking at indirect or unintended program impacts:

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

Summary of Actions

- Determine what elements need to be assessed to quantify program impacts
- Identify the type of evaluation used to assess the program impacts
- Verify whether the examples of research questions pertain to the program evaluation

Step 7: Evaluation Plan Development Guidelines

Key Points / Highlights

Evaluation Administrators should consider the following tasks when developing an Evaluation Plan:

7a. EM&V Plan Content and Structure

7b. Final Evaluation Plan (FEP)

7c. Key Evaluation Consideration

The Evaluation Administrator authors the evaluation planning documents. The first step is development of a *Draft Evaluation Plan*. An evaluation plan results from the steps presented in above.

The Evaluation Administrator uses the logic model to select areas of study and to choose the types of evaluations sought.

Program Managers and Evaluation Administrators use their knowledge of program objectives, delivery mechanisms, and motivations to properly scope the evaluations needed. Evaluation planning includes allocating program resources to monitoring, measurement, verification and evaluation.

Types of Evaluations and Assessments Typically Included in Draft Evaluation Plans

- **Impact Evaluations** – these look at behavioural outcomes and their likelihood to generate the intended program impact (typically demand reductions and energy savings). They may also look at both positive and negative unintended impacts. To the extent that unintended impacts have a substantive impact on program outcomes, they should be evaluated.
- **Process Evaluations** – these are used to explore the methods, activities, and expenditures used to generate program outputs. They evaluate things like the effectiveness of promotional campaigns, informational materials, educational seminars, training, financial assistance, technical assistance, etc.
- **Market Effects Evaluations** – these are used to estimate the contribution of program outcomes to market trends. They may also be used to evaluate the converse: how trends in the market place (for example, electricity pricing, rate schedules, legislation, and so on) impact program outputs.
- **Cost-effectiveness assessments** – these are used to quantify and analyze the benefit and cost streams (for example, cost-benefit ratios). These are generally conducted after the impact and process evaluations have been completed.
- **Outcome Evaluations** – these are used to explore how behaviours arise from program-sponsored activities. They seek to explain behavioural choices in the context of desired attitudes and added abilities resulting from program outputs.

Each of these types of evaluations is discussed in detail in **Step 6: Evaluation Classification Protocols**.

Task 7a: EM&V Plan Content and Structure

An example of a Draft Evaluation Plan Template is provided (p. 29). Unless there is a specific reason for using some other format, using it as such is recommended because it facilitates easy review of plans and approvals from Program Managers and executive management across different programs.

Task 7b: Final Evaluation Plan (FEP)

A *Final Evaluation Plan* builds on the *Draft Evaluation Plan*. The Evaluation Contractor works with the Evaluation Administrator to formalize all elements and objectives of the evaluation. The Evaluation Contractor submits the *FEP* to the Evaluation Administrator for final approval. The FEP is detailed enough to ensure the approved evaluation activities yield a high level of confidence in the reported energy savings, demand reductions and program cost effectiveness.

Task 7c: Key Evaluation Considerations

When planning evaluations, Program Administrators and Evaluation Administrators should consider how the evaluation serves as a management tool. The evaluation provides savings estimates that demonstrate program impact and cost-effectiveness, which may be used for regulatory purposes. Evaluation findings are used to improve both short-term and long-term impacts, allowing mid-course corrections to enhance program achievement. To realize these benefits it is important to keep in mind that evaluations are not meant as mere audits of program performance.

To help ensure the usefulness of evaluations, keep the following in mind:

- **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* is prepared as part of the program design and an evaluation budget is assigned at that stage. On completion of the program design, the evaluation plan is implemented to ensure data is collected and reported in a timely manner, allowing for incremental feedback to guide Program Managers.
- **Program Goal Setting** – If the program (or portfolio) goal is to save electricity during peak hours, the evaluation goal is to accurately document how much electricity demand is reduced during the peak hours (gross savings), how much of these savings can be attributed to the program (net savings).
- **Preparing for Program Launch** – Ideally, the draft evaluation plan should be prepared before the program is launched. If it cannot be developed before program launch, it should be drafted as soon as possible following program launch. Baseline data should be collected before, or soon after, program launch so that market effects resulting from the program offer are documented.

- **Defining the Evaluation Objectives** – Evaluations focus on the linkage between program outputs and the resulting program outcomes. The evaluation should provide guidance to the Program Administrator on ways to enhance program efficacy. To this end, Program Administrators and regulators need to be assured that the evaluations conducted will deliver the type and quality of information needed.
- **Program Implementation** – Some baseline data collection and all program reporting continues throughout program implementation. The incremental data informs and updates program metrics. The Evaluation Administrator should analyze and present performance metrics to Program Managers as findings from Evaluation Contractors. Keep in mind that evaluation activities often continue after the program year is completed.

Evaluations must also be properly scoped. Addressing issues that are not program priorities or issues, or employing unnecessarily complex methods, can waste valuable resources. When faced with limited evaluation resources prioritizing the key activities will ensure the evaluation objective have been met without straining resources.

Summary of Actions

- Scope of Evaluation deliverables
- Create a draft Evaluation Plan
- Work with Evaluation Contractor to complete the Final Evaluation Plan

Draft Evaluation Plan Template

Program Overview

Program Description

Provide a short introduction of the program offer from the perspective of the program manager. It 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 Program Logic Model.
- **Eligibility Criteria:** Describe the protocols/procedures that will be used to qualify program applicants or markets targeted.
- **Key Program Elements:** Highlight the intended program process flow. Each program element should be identified in the 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 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 evaluation administrator should investigate independently the causal influence of each program element towards the realization of intended programmatic impacts. The program manager should review the logic model and ensure it is an accurate portrayal of the program theory.

Annotate the program logic model from 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, 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 program stakeholders to the evaluation administrator. 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 here. The number of research questions should be limited and prioritized based on reasonable use of resources.

Draft Evaluation Plan Template

Evaluation Approach

Introduce the details of the approach that follows.

Evaluation Type (repeat for each type)

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

☐ **[Frequency] Impact Evaluation.**	Impact evaluation description.
☐ **[Frequency] Process Evaluation.**	Process evaluation description.
☐ **[Frequency] Market Effects Evaluation.**	Impact evaluation description.
☐ **[Frequency] Cost Effectiveness Evaluation.**	Cost-effectiveness evaluation description.
☐ **[Frequency] Outcome Evaluation.**	Outcome evaluation description

Study Focus. Associate the planned approach to applicable research questions. Indicate how the planned evaluation activities contribute to or answer the questions at hand. This is often done in the form of a null hypothesis.

Data Collection Plan. Describe the processes deemed appropriate to collect, validate, and audit the data used in the evaluation.

Analysis Methods. Describe the specific analytical methods sought for the evaluation. For example, one may wish to normalize weather to a specific year versus a long-term normal average daily temperature.

Limitations/Caveats. Describe limitations and restrictions associated with intended approach; thereby, providing evaluation contractors and implementers the ability to improve upon the planned evaluation.

Study Outputs. Identify the specific outputs expected by the evaluation administrator of the evaluation contractor. This description may include a report template, presentation requirements, delivery media, ownership of resulting datasets, etc.

Evaluation Dependencies

Discuss key collaborations essential to the successful implementation of the evaluation. The following are common dependencies associated with industry research, more may be added as appropriate for the planned evaluations.

☐ **Enabling Stakeholders**	Identify and discuss as is appropriate.
☐ **Access Requirements**	Identify and discuss as is appropriate.
☐ **Data Sharing**	Identify and discuss as is appropriate.
☐ **Funding Support**	Identify and discuss as is appropriate.

The evaluation activities undertaken as part of the program evaluation should be carried out using the guidelines specified in the 2019-2020 Interim Framework EM&V Protocols and Requirements v3.0.

Draft Evaluation Plan Template

Special Provisions

Clarify any typical considerations associated with the planned evaluations.
Where necessary and helpful, attach materials necessary to fairly represent the work envisioned.

Data Collection Responsibilities

A listing of all the data that must be collected to support the evaluation of the program and who is responsible to collect it.

Evaluation Schedule

A listing of all the physical deliverables that will be part of the Evaluation, e.g., evaluation plans, memos, interim reports, final reports.

Evaluation Deliverable	Date
Draft Evaluation Plan	
Final Evaluation Plan	
Other Deliverable #1	
Other Deliverable #2	
Other Deliverable #N	
Draft Final Evaluation Report	
Final Evaluation Report	

Step 8: Hire an Independent, Qualified Evaluation Contractor

Key Points / Highlights

Hiring an Independent, Qualified and Authoritative Evaluation Contractor involves the following tasks:

8a. Provide for EM&V Contractor Autonomy

8b. Request Independent Verification of Program Outputs

8c. Select an appropriate Methodology

Task 8a: Provide for EM&V Contractor Autonomy

An independent evaluation requires that unbiased parties with no real or perceived conflicts of interest conduct the planned evaluations. Evaluations conducted by Program Managers themselves are not considered sufficiently "independent" to verify program savings.

An organization can sponsor both a program and its evaluation but in that case, the sponsoring organization must procure a third-party evaluator to implement EM&V plan, drafted by the Evaluation Administrator. In very narrowly prescribed situations, an organization sponsoring a program may appoint an internal review board or specialized program evaluation staff to assess a program offer and implement the approved EM&V plan. In such cases the sponsoring organization must be able to demonstrate autonomy between the groups implementing the program and the groups evaluating the program.

In all cases, whether it is the Evaluation Contractor or an internal review board must be free to report their findings without consequence or retribution.

Task 8b: Request & Ensure Independent Verified Results

The intended impacts of the programs will always be reduced energy demand and savings. The Draft Evaluation Plan must explore unintended impacts that may result from the intervention. The requirement that the Evaluation Contractor must be exploring both the positive and negative impacts expected from the program must be part of the evaluation scope of work and must set out in the contract with the Evaluation Contractor.

The Evaluation Contractor must be free to present to the appropriate regulatory authority or administrative agency its findings, results, and conclusions without limitation. Under no circumstance may valid findings of fact, substantive conclusions, verified impacts, or program recommendations be censored. Where the sponsoring organization (the Evaluation Administrator or Program Manager) and the Evaluation Contractor disagree about a point, the disagreement should be outlined in footnotes in the EM&V report. The footnotes should clearly outline the opposing arguments, including attribution to the person raising the concern.

Task 8c: Select an Appropriate Methodology

Depending on the program evaluation different methodologies can be proposed by the Evaluation Contractor. It is the job of the Evaluation Administrator to ensure that the appropriate methodology is selected. Notably, methodologies can vary depending on the data available to conduct the analysis.

Clear and specific capacity/demand reduction targets, as well as energy savings targets, have been established for CDM initiatives. And, thanks to the installation of smart metering technologies, data related to energy use exists.

While hourly load shapes add rigor to EM&V practices, the Evaluation Administrator must not dismiss the basic principles of program impact assessment. Savings calculations require a gross-to-net savings adjustment, either by generally accepted net-to-gross calculations or through net-savings calculations based on experimental or quasi-experimental models.

This task establishes a preference for advanced analytics involving smart-meter data as a key method for the verification of demand reduction and energy savings associated with CDM initiatives.

Summary of Actions

- Outline (in a footnote) any disagreements between the Evaluation Contractor's findings and conclusions and those of the sponsoring organization.
- Have an independent review of program monitoring practices carried out.

Step 9: Vendor Selection Process Guidelines

Key Points / Highlights

A competitive procurement process allows the Evaluation Administrator to choose from a number of proposals, which helps the Evaluation Administrator to balance many factors in an effort to meet the evaluation priorities.

9a. Evaluation Contractor Selection Process

9b. Budget Consideration

There are a number of reasons why EM&V services should be procured through a competitive process. Second, the contracted values generally associated with EM&V services often exceed the monetary thresholds that trigger competitive procurements within the public sector. Secondly, since varied approaches can often be taken for the provisions of EM&V services, by using a competitive process the Evaluation Administrator may have several options from which to choose. Lastly, a competitive solicitation ensures multi-jurisdictional vendor support for Ontario's EM&V service requirements.

Public procurements in Ontario are expected to comply with the December 2014 Procurement Directive issued by the Management Board of Cabinet. The overall objective of this Directive is to ensure acquisition of goods and services are conducted in the most economical and efficient manner.

A benefit of relying on a competitive procurement process is that the Evaluation Administrator generally will be able to choose from a number of proposals, which helps the Evaluation Administrator balance many factors in an effort to best meet emergent priorities. Vendors often submit proposals that set forth methods that tackle issues and tasks in unanticipated, clever, and meaningful ways; providing a learning opportunity for Evaluation Administrators and Program Managers.

The Draft Evaluation Plan (found in **Step 7: Evaluation Plan Development Guidelines**) forms the basis of the request for consulting services.

Task 9a: Evaluation Contractor Selection Process

Once a valid RFP process (as described in the section above) has been held, a winning bidder must be selected. It is important that an objective selection process be followed and that appropriate documentation of the selection process is recorded and filed.

The simplest way to avoid bias or the perception of bias in the selection process is to employ an Evaluation Contractor Selection Committee. Generally it is best to form a cross functional team representing the varying interest in the evaluation results.

Task 9b: Budget Considerations

When issuing an RFP for evaluation services to vendors, information on the program's budget for services will not be included.

There are general guidelines on the appropriate amount to spend on evaluation relative to the size of a program. As detailed in the Protocols, the typical range is 4% to 6%. Small pilot studies where very detailed information will help inform and reduce risk for a potential broader roll-out strategy could justify spending the same amount

as the program itself. In fact, pilots could be considered a form of evaluation. On the other end of the spectrum, a program that has been running consistently for several years and that has no new or unusual activity happening in it may require only a basic level of field verification and audit and so it should not require a significant expenditure. The cost to achieve a successful evaluation is also affected by whether multiple evaluation categories are required (outcome, impact, process, market, cost-effectiveness) or just a selected one.

The second reason not to include budget expectations in an RFP is because Evaluation Contractors will propose alternate methods and approaches to achieve the same end result. And, since there is more than one appropriate and acceptable way to accomplish most energy program evaluation tasks, alternate methods may have different cost implications. It is best to allow the proponents to detail their position as to why the combination of quality and cost they propose should outrank their competitors.

A third reason is that evaluation methodologies and best practices are also evolving. So, at any time, proposals may present a new way to measure performance results. A core purpose of the competitive process is to spur this type of innovation and creative thought process. We want RFP respondents to continually strive to provide the best value proposition.

Lastly, it will be rare that the absolute best quality approach will get selected or even proposed. Energy program evaluation is always a compromise between best practice and available resources. Managing this balancing act 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.

Summary of Actions

- Public procurements in Ontario are expected to comply with the December 2014 Procurement Directive issued by the Management Board of Cabinet.

Step 10: Coordinate EM&V Activities and Report Findings

Key Points / Highlights

Coordinating EM&V Activities and Reporting Findings involves the following tasks:

10a. Detail Research Methodologies Employed

10b. Present Evaluation Findings

10c. Assess Reasonableness of Conclusions and Recommendations drawn from Evaluation Findings

Task 10a: Detail Research Methodologies Employed

Evaluation reports must include a detailed statement of the analytical methods used. Such reports should include a description of the evaluation objectives, a list of the research questions addressed, the approach taken to answer the research questions, the experimental model(s) employed and the analytical methods used in the presentation of findings. This will require data collection instruments (i.e. survey work) to be appended to the report.

The descriptions used in the evaluation report must be detailed enough to allow other evaluation professionals to repeat the procedures used by the Evaluation Contractor and to facilitate audits administered by the appropriate regulatory bodies and/or administrative agencies.

Task 10b: Present Evaluation Findings

The findings of an evaluation report should be presented clearly in either graphical or tabular format. Text must highlight key findings and link the data to the research methods used to analyze data. The Evaluation Contractor must outline in the report instances where the findings confirm or contradict earlier findings, including specific reference to the previous study.

Evaluation Administrators and Program Managers may find regular monthly program reporting and EM&V findings overlap during early stages of a program offer. Such redundancy can be helpful in verifying that critical program outputs and outcomes have been achieved. As programs mature and EM&V efforts focus on downstream behavioural outcomes and program impacts, the frequency of the reports maybe reduced, depending on the regulatory requirements of the jurisdiction.

Task 10c: Assess Reasonableness of Conclusions and Recommendations Drawn from Evaluation Findings

Evaluation Contractors must reference their conclusions to the key findings upon which the conclusions are based. Furthermore, the conclusions must be based on the data actually collected from the evaluation process versus broad inferences based solely on their experience in other jurisdictions. It should be noted that while inferences from experience in other jurisdictions may be provided, the inferences must be provided within the context of a comparative analysis explicitly requested in the evaluation scope of work.

Because conclusions and recommendations made by the Evaluation Administrator and the Evaluation Contractor often drive policy decisions, it is important that conclusions be drawn from actual findings and that the context be clearly stated. In other words, given the effect of evaluation conclusions and recommendations on organizational priorities and budget allocations, Evaluation Administrators and Evaluation Contractors must ensure the conclusions and recommendations formulated can be supported by the research findings and fall within the scope of the funded evaluation.

Summary of Actions

- Provide a detailed statement of the analytical methods used
- Clearly present the evaluation findings graphically or in tabular format
- Ensure conclusions are referenced in key evaluation findings
- Ensure context for evaluation findings is stated

Step 11: Publication of Evaluation Reports

Key Points / Highlights

Publication of Evaluation Reports involves the following knowledge:

11a. Address Timelines and Veracity of Savings Claims

11b. Address Comparability of Results

11c. Address Use of Utility Billing and Meter Data

11d. Address Defensibility of Gross-to-Net Calculations

11e. Presentation of Evaluation Results

Task 11a: Address Timeliness and Veracity of Savings Claims

The appropriate regulatory authority and administrative agencies establish annual energy savings and demand reductions.

The information the participant provides regarding claimed saving is used to determine portfolio savings estimates. It is important to conduct the evaluation in a timely and efficient manner so that the results can be used by the varying audiences for program enhancements, program design and forecasting etc.

The savings target reconciliation as established by the appropriate regulatory authority and administrative agencies is final. As such, Evaluation Administrators and Evaluation Contractors are encouraged to administer EM&V as outlined within these protocols.

Task 11b: Address Comparability of Results

Demand reductions and energy savings are considered verified estimates of program impacts. Since point estimates of energy and demand savings may vary in both precision and levels of confidence, the statistical reliability of the reported impacts are considered when comparing impact assessments.

The Evaluation Administrator should prefer a 5% confidence interval around point estimates and ensure a .95 level of confidence for claimed impacts. Where necessary experimentally, exceptions may be used by the Evaluation Contractor. It is helpful if options, including the cost implications, for 5%/0.95 and 10%/0.90 confidence are provided for in Draft Evaluation Plan requests and responses so that Evaluation Administrators can assess the benefit-cost of increased accuracy in the context of their total evaluation budget.

Task 11c: Address Use of Utility Billing and Meter Data

Evaluation Administrators are strongly encouraged to seek the most robust and direct measurement of energy savings and demand reductions available. Site-specific hourly load shape analysis is the preferred method for calculating achieved results.

Studies using pre/post billing and meter data comparisons are given added weight over studies using prescriptive and quasi-prescriptive estimates of savings based on measure savings assumptions. Evaluated retrofits, for example, must be both measured and verified.

Whole premise measurements should use revenue-grade meters to ensure the most precise estimate of energy use and demand requirements. Where retrofits are isolated and individually metered, meter precision must be addressed when stating the achieved energy savings or demand impacts. If information regarding metered results for both a pre-retrofit and post-retrofit period is lacking, the Evaluation Contractor may use a calibrated simulation. Use of a calibrated simulation should be a method of last resort, but it may be used when evaluating new construction, constant load lighting, re-commissioning projects, and industrial process initiatives. Please refer to **Technical Guide 5: Gross Energy Savings Guidelines** and **Technical Guide 4: Project-Level Energy Savings Guidelines**. Note that use of the International Performance Measurement and Verification Protocol (IPMVP) is an integral part of project-level savings assessments.

Task 11d: Address Defensibility of Gross-to-Net Calculations

Gross saving estimates are not applied to program targets because gross savings estimates do not account for what would have normally occurred absent of program incentives or energy efficiency upgrades. As a result, net savings are used. Given this, it is essential that the calculations used to establish net savings are defensible.

Technical Guide 8: Net-to-Gross Adjustment Guidelines is provided as a reference, but does not replace the expert judgment of the Evaluation Administrator and Evaluation Contractor.

Both the Program Administrator and the Evaluation Administrator must address the calculation of net savings in the development of an EM&V plan. Furthermore, the Evaluation Contractor must be provided with the latitude to adjust gross savings estimates. Where possible, evaluated savings should be normalized to long-term weather and socio-economic trends so that year-over-year savings estimates can be compared.

Summary of Actions

- Ensure claimed savings are accurate
- Ensure comparability of study groups
- Choose appropriate cost/confidence level
- Verify type of meter data used
- Specify meter precision information
- Explain how net savings figures were arrived at
- Consider normalizing savings to applicable long-term trends

Task 11e: Presentation of Results

Evaluation results can be presented in a variety of ways. Evaluation Administrator should apply the preferred method to present results. However, at a high-level comprehensive evaluation report should contain the following information:

Summary of Impact Evaluation Results

For cross-cutting evaluations, include additional columns for each initiative and a total column

Program Metric	Program 1	Program 2	Total
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

Summary of Process Evaluation Results

Key Process Evaluation findings

Research Question	Observations	Recommendations

Cost Effectiveness Results

Cost Test		Program 1	Program 2
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

Conclusion and Recommendations

Step 12: Guideline for Managing Program Evaluation Contractors

Key Points / Highlights

The responsibilities of an Evaluation Administrator for managing program evaluation contractors include:

12a. Optimizing Resource Utilization

12b. Project Coordination

12c. Providing Data

12d. Quality Assurance

After the evaluation has been planned and an Evaluation Contractor assigned, program evaluation tasks must be implemented and managed. The Evaluation Administrator serves as a liaison with the Evaluation Contractors, coordinating a number of tasks over the course of the EM&V efforts. While the contracted evaluator completes the bulk of the work, the Evaluation Administrator has the following responsibilities:

Task 12a: Optimizing Resource Utilization

Evaluation Administrators must balance resource commitments within and between multiple projects. Plotting all evaluation activities on a single research calendar helps to identify opportunities to integrate data collection strategies and analysis method, even where the activities cross programs, portfolios, or evaluation disciplines. The proper use of resources avoids sampling fatigue among study populations, maximizes the available funds, and provides valued output.

Task 12b: Project Coordination

Work, schedules and deliverables must be reviewed daily. The management of evaluations requires the organization of meetings, the establishment of goals, management of stakeholder participation, coordination of evaluation activities among team members, integration of study findings and publishing of results.

Task 12c: Providing Data

Evaluation requires an exchange of information between planners, implementers, program participants, trade allies, comparison groups, involved organizations, and agencies. Data tracking and warehousing requires an infrastructure for this exchange. Data quality must be ensured before an analysis will meet the reliability standards established by the industry. While this work may be sourced to specialty contractors, the transformation of raw data into consumable and valued information requires significant oversight. As part of the data collection process, the Evaluation Administrator and the Evaluation Contractor should also be familiar with the Freedom of Information and Protection of Privacy Act and privacy laws in general. In particular, a data management plan should be developed for the collection, storage, disclosure and disposal of any personal information as part of the evaluation process.

Task 12d: Quality Assurance

Administrative agencies and regulatory authorities rely on the quality of the planned evaluations. The Evaluation Administrator is responsible for ensuring quality work has been completed before the results are published and presented to key decision makers. Quality assurance requirements have been established with the Protocols, as well as in the Technical Guidelines. The Evaluation Administrator must ensure information in each published evaluation report, summary of findings, or memo, adheres to the established standards.

Summary of Actions

- Ensure the program evaluation contractors are provided with sufficient resources in accordance with the contract
- Ensure the results of the evaluation adheres to the established standards

Part 2:

Conducting an Evaluation

Audience: Evaluation Contractor

Introduction to Part 2

The primary audience for Part 2 is Evaluation Contractors.

This Part is comprised of Technical Guides that relate to different technical processes and techniques that Evaluation Contractors use in conducting evaluations. Because the Technical Guides in this Part cover different topics, each can be read on its own. The Technical Guides provide information on:

- Technical Guide 1: Using Measures and Assumptions Lists
- Technical Guide 2: Program Cost-effectiveness Reporting
- Technical Guide 3: Conducting Process Evaluations
- Technical Guide 4: Determining Project-level Energy Savings
- Technical Guide 5: Determining Gross Energy Savings
- Technical Guide 6: Calculating Demand Savings
- Technical Guide 7: Determining Market Effects
- Technical Guide 8: Net-to-Gross Adjustments
- Technical Guide 9: Statistical Sampling and Analysis
- Technical Guide 10: Behaviour-Based Evaluation Protocols

Also Useful to Program Administrators

The work carried out by the Evaluation Contractor involves data collection and analyses that can be relatively technical. To ensure the Evaluation Administrator is able to effectively manage the process and gauge the quality of the work the Evaluation Contractor is doing, the Evaluation Administrator needs a basic understanding of the relevant techniques and methods. This information can be found in Part 2. Unlike the steps set out in Part 1, the guides in Part 2 are stand-alone and provide a high-level understanding of a particular technical process.

Technical Guide 1: Using Measures and Assumptions Lists

Key Points / Highlights

Use of accurate and defensible technology assumptions is critical in planning and assessing conservation and demand management (CDM) programs. The assumptions on which CDM programs are found are contained in "measures and assumptions lists" (MALs). The assumptions can be used to screen and assess measures for possible inclusion in a conservation program before the program runs (ex ante). As well, the MALs are used after the program runs (ex post) to evaluate the savings generated by measures and projects undertaken as a result of participation in the program.

The Program Manager reviews input assumptions for measures that are under consideration for inclusion in a program. This information is used to generate energy and demand savings estimates and to provide input into program cost-effectiveness calculations conducted for program design. It is important to use the most recent measures and assumptions list.

Evaluation Managers are responsible for ensuring that information used in evaluations is up to date and accurate.

Prescriptive and Quasi-Prescriptive Assumptions

Input assumptions are either prescriptive or quasi-prescriptive in nature, depending on whether application-specific information is needed to better reflect variations in how the technology is used or operated.

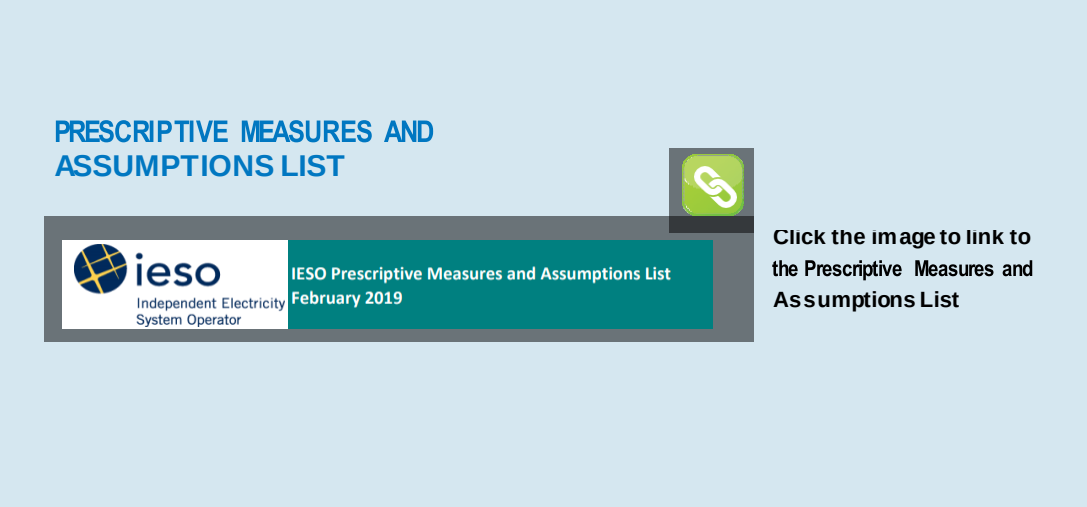
Measures that are included in MALs are typically substantiated with documented credible results or third-party verification, testing, or certification.

Measure-level assumptions are referred to as **"input assumptions".**

Conservation and Demand Management (CDM) programs are programs designed to reduce the amount of electricity participants consume.

Prescriptive measures are measures where the energy savings are pre-determined based on how the typical conservation program participant obtains resource savings as a result of implementing the measure (the savings are determined by applying fixed input assumptions into energy and demands savings equations).

Quasi-prescriptive measures are measures with resource savings estimates that vary depending on the technology or type of equipment and the context in which the measures are used. Quasi-prescriptive measures provide a methodology that allows for estimating resource savings for various scenarios, rather than relying on a fixed saving value for all scenarios.



Examples of key input assumptions on which measures included in MALs are include:

- 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)
- Seasonal and time-of-use (TOU) energy savings patterns (for example, periods emerging from system planning and/or regulatory rate structures such as summer, winter, and shoulder season TOU periods)
- Incremental cost data (for example, the cost differential between baseline equipment and high-efficiency equipment)
- Equipments' useful life and other assumptions about persistence

The measure-level assumptions are reviewed periodically, and the assumptions are updated as new knowledge, information, or technologies emerge.

Purpose and Scope of this Guideline

This guideline applies to all CDM programs that support or promote the installation of technologies with prescriptive or quasi-prescriptive assumptions and that are contained, or should be contained on the approved MALs.

This guideline provides information to CDM Program Managers, portfolio managers, and Evaluation Managers with regard to the use of input assumptions included in MALs, and to assist Program Evaluators in data collection, review and updating of measure-level assumptions.

Early in program planning and development Program Designers consult MALs to ensure that measures included in a program:

- are likely to produce reliable energy and/or demand savings
- are cost-effective and provide net benefits to society as demonstrated through the use of the cost effectiveness tests (**Technical Guide 2: Cost-Effectiveness Guidelines**)
- will satisfy other program objectives

Free ridership rates and other net-to-gross adjustment factors are not taken into account in MALs. Such factors are a function of program design and operation and must be determined and accounted for on a regular basis through program evaluation research. In the absence of better information, broad adjustment factor assumptions may be used for program planning and/or portfolio management purposes. But, any free ridership or other net-to-gross adjustment factors should be addressed by the evaluation and program input assumptions and revised as information is gained. These factors are discussed in *Net-to-Gross Adjustment Guidelines*.

Understanding & Using MALs

All parties involved in the planning, design, implementation and evaluation of resource acquisition CDM programs should be familiar with how MALs are used. When using input assumptions, either those included in MALs or that should be included in MALs, it is important to:

- Understand assumptions and processes used to develop the MALs
- Know of existing measure input assumptions
- Know of, or be able to locate, recent evaluations of comparable programs and assessments of similar technologies
- Have the technical ability to undertake a practical review of measure assumptions, if required
- Understand the need to substantiate measure assumptions and updates

MALs are typically approved by a regulatory board, commission, or authority that is accountable for ensuring that CDM program investments are cost effective and produce real savings.

Methods of Reviewing Input Assumptions

An input assumptions review is usually one of the first steps in developing a CDM program Evaluation Plan (**Step 7: Evaluation Plan Development Guidelines**). Reviewing input assumption may also be a part of a planning project-level measurement and verification (M&V) activities (**Technical Guide 4: Project-Level Energy Savings Guidelines**) to establish measurement techniques and procedures for calculating savings derived from projects.

Input assumptions for measures included in a typical MAL are:

- Description of the efficient technology
- Description of baseline technology (that is the technology that the efficient technology is replacing)
- Annual energy and demand savings
- Demand savings coincident with summer and/or winter system peak
- Seasonal energy savings patterns
- Effective useful life of the efficient technology (persistence)
- Incremental efficiency technology costs

Caution is required when using a MAL or measure assumption developed for use in other jurisdictions, especially where there are different codes, standards or market conditions. In all cases, the source of the assumptions for measures should be documented.

To provide an appropriate level of confidence in the MAL, periodic reviews of all underlying measure input assumptions are completed by independent research and through program evaluation activities. Any assumption update is based on the best available information.

Where insufficient data exists to complete an update to an assumption, the evaluation should use M&V to verify or re-estimate the assumption. New measure assumptions should be substantiated using literature reviews, program evaluations, case studies or third party testing, verification, or certification relating to the specific measure being investigated.

Documentation and Reporting

The Evaluator will 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.

Updating the Measures and Assumptions List

The IESO has an open, transparent, and flexible approach for reviewing and maintaining its MALs. Any stakeholder can submit measure revisions, or other measure considerations.

All requested updates/submissions related to MALs require verification. IESO staff use a standardized Measures and Assumptions Substantiation Form.

Review of Measures and Assumptions List Update Requests

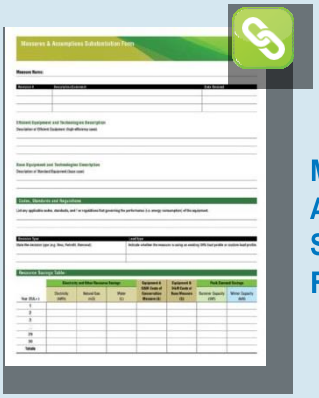
The submissions are reviewed based on the merits of the information provided. Following the review, submissions are either accepted as submitted, accepted with modifications, or rejected on specified grounds.

The review process time and approval is primarily dependent on the quality (relevancy and credibility) of the information provided to the IESO. Information referred to in substantiating the request must be available to, and accessible by, the IESO.

The IESO strongly encourages the inclusion in the submission an hourly (8760) annual load profile created from metered data or from a verified operating schedule. If unavailable, a description of the operating hours during weekdays and weekends for different seasons will be considered.

The Measures and Assumptions Substantiation Form

The IESO Form shows the information that is to be submitted when requesting an update of the IESO's Measures and Assumptions List. External stakeholders are encouraged to use the IESO form, or at least consider it as a guideline when making a submission.



MEASURES AND ASSUMPTIONS SUBSTANTIATION FORM

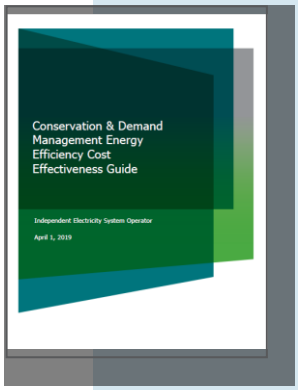
Summary of Actions

- Consult the MALs to see whether the measures are in them
- Conduct an input assumption review
- Consider whether the correct confidence in values in the MAL
- Consider whether to submit update of MALs

Technical Guide 2: Cost-Effectiveness Guidelines

Key Points / Highlights

The *Conservation and Demand Management Cost Effectiveness Guide* sets out the cost-effectiveness policy articulated in the EM&V Protocols. Evaluation Administrators and Evaluation Contractors must follow the requirements of the guide.



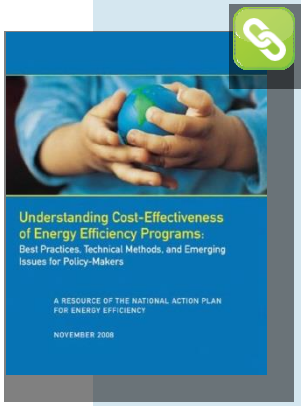
Cost-Effectiveness Guidelines

This Cost Effectiveness Guide ("Guide") describes standard industry metrics to assess the cost effectiveness of conservation and demand management (CDM) resources. The Guide may be updated from time to time. Cost effectiveness assesses whether the benefits of an investment exceed the costs.

Purpose

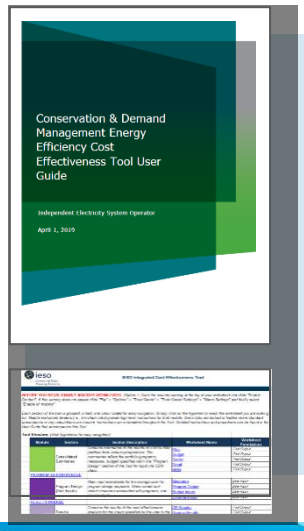
The purpose of the guide is to ensure program cost-effectiveness is considered by a broad range of stakeholders, including and but not limited to:

- Program Administrators
- Regulatory Agencies
- Administrative Agencies
- Policy Makers and
- Ratepayer Advocates



Reference: Understanding Cost-Effectiveness of Energy Efficiency Programs

A helpful tutorial on the common CDM-related cost tests can be found in the following document. Click the title to access the document.



CE Tool User Guide and CE Tool

These documents are intended to support IESO, LDC staff and other external service providers and/or delivery agents to calculate resource savings, budget and cost effectiveness metrics for new and existing conservation programs in Ontario.

Summary of Actions

- Review the *Conservation and Demand Management Cost Effectiveness Guide*
- Ensure Evaluation Administrators and Evaluation Contractors follow the requirements of the guide to assess program cost-effectiveness.

Technical Guide 3: Process Evaluation Guidelines

Key Points / Highlights

A process evaluation is an empirical examination of program design, development, delivery, and administration. Such a systematic assessment of program elements, from resource allocation through program outputs, ensures program stakeholders that the planned offer is realized.

Process evaluations yield both qualitative and quantitative findings on which practical advice can be offered to enhance the program the design and administrative processes and the program service delivery. Unlike audits, process evaluations should provide evidence of outstanding practices and the means by which these practices can be transferred to other program delivery agents.

Process evaluations gauge the effectiveness and appropriateness of the following:

- **Program Design** – the linkage between key program elements, as well as the reasonableness of program objectives and resource allocation.
- **Program Development** – the protocols and procedures that form the basic offer to be implemented; the training and technical assistance provided to program delivery agents; and the changes made to the program design.
- **Program Administration** – the controls established for program delivery; the procurement processes for program goods and services; and the mechanisms in place to evolve the program offer.
- **Program Delivery** – the services provided by program agents; the processes used in the field to deliver the program offer; the systems used to track and monitor program outputs; the actual program expenditures over the assessment time horizon; the quality of measure installation; and the levels of participant satisfaction maintained throughout the offer.

Collaborative Effort

Because of the need for collaboration among program delivery agents, contracted or external Program Managers, and the Program Administrator, process evaluations are complex. The Evaluation Administrator is responsible for fostering a cooperative relationship between the Evaluation Contractors who will be charged with carrying out the work and the program actors.

Experience has shown that attention to the following will help establish strong collaboration between program staff and the evaluation team:

- **Make introductions early:** The Evaluation Administrator should introduce themselves and the Evaluation Contractors to program staff as early as possible within the program development life cycle. Without early involvement, elements of program theory could be missed and the process evaluation could easily turn into, or be perceived as, a process audit.

- **Appreciate that program management and delivery staff are the experts.** Evaluation contractors are experts in assessment, not program operation. Only the program staff can offer the details needed to appreciate the available operational options and the choices made; without this expertise, the process evaluation cannot be developed and meaningful recommendations will not be identified. It is the Evaluation Administrator's responsibility to get the required information from program staff. (Information gathering is an essential competency of any process Evaluation Administrator.)
- **Recognize that observation affects operation.** It is important to remember that an effect cannot be measured without it being affected by the tool used to record the measurement. Process evaluations are a measurement of operational efficiency. As such, the presence of the Evaluation Contractor could affect the efficiency and efficacy of the process being assessed. Evaluation Administrators must be mindful of this when the Evaluation Contractor is formulating conclusions and recommendations.
- **Ensure findings are shared regularly.** After each field visit the Evaluation Contractor should share his/her findings with the Evaluation Administrator, who should then provide the information to the appropriate level of operational management. The responses offered by direct supervisors of those being observed will enlighten Evaluation Contractors about operation constraints and provide the basis for interpreting the evidence collected.

Process Data Collection

Collecting process evaluation data is relatively straightforward. The evaluation of a process begins by answering the five questions: who, what, when, where, and how.

What the Evaluation Contractor is looking for with respect to each question:

Who?	participant, service provider, Program Manager, etc.
What?	activity, materials, measures, behaviours, processes, etc.
When?	frequency, duration, size of interaction, etc.
Where?	home, office, internet, phone, etc.
How?	program policies, procedures, protocols, etc.

Process data should be recorded for each program element or program activity identified within the program logic model (see **Figure 1.0: The Basic Elements of a Logic Model**). The Evaluation Contractor should be confirming whether expenditures match the program budget and if the expected outputs resulted from the activities observed.

The processes evaluated should be readily distinguishable from each other. The process assessment should focus on observable behaviours, the materials leveraged, and how program materials were received by participants.

Each process chosen for assessment should be looked at thoroughly. However, not all processes can be included in the process evaluation. The Evaluation Administrator and the Program Administrator should have already set into place the critical research questions to be answered and the Evaluation Contractor need only examine the processes that fit within the scope of the study.

Process Evaluation Methods

Process evaluations consist of both quantitative and qualitative methods. Metrics for the quantitative assessment are often tracked by Program Administrators and program delivery agents within tracking systems and management reports. Qualitative data, on the other hand, must be observed or collected through survey/interview techniques.

In deciding who should collect the data, the Evaluation Administrator should balance cost and convenience against potential biases.

The methods listed below are frequently used when assessing processes, though other techniques may be recommended and used by the Evaluation Contractor:

- **Reviewing Field Notes:** These are brief records kept by program participants or delivery agents (typically recorded on forms). These forms may be part of the program delivery model or may be forms developed by the Evaluation Contractor. Examples of field notes include: activity logs, diaries, inspection notes, receipts, etc.
- **Creating a Case Study:** Case studies are created based on detailed records, often recorded by the Evaluation Contractor, of a small number of observed program activities.
- **Conducting Ethnographic Analyses:** This is a method of research that involves the Evaluation Contractor's direct observation of a program activity. This may include a "ride-along", which is where the Evaluation Contractor goes into the field with service providers and interacts directly with recipients of program measures and asks questions of program staff regarding their activity.
- **Conducting a Delphi Analysis:** This involves convening a panel of experts to explore a particular process or issue. The objective is to build a consensus 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, aimed at learning about focus groups members experience with a product or service offering of the program.
- **Using Questionnaires:** Using surveys conducted via phone, mail, e-mail, Internet/online or through comment cards with respondents answering questions outlined based on pre-defined questions.
- **Conducting Unstructured Interviews:** This technique is used to elicit information in complex situations where program participation-related motivations are likely to be multi-faceted and behaviours influenced by multiple factors. Unstructured interviews also work well when there is no single decision-maker or the actual decision-maker is not easily determinable (for example, a large industrial customer with significant energy efficiency investment).

The Process Evaluation Report

Keep in mind that the Process Evaluation Report can never be a compilation of all data recorded. Process evaluation reports should present summary data and should summarize important conclusions, as well as present recommendations based on the evaluation findings. Because there are many processes that get reviewed over the course of a program assessment and the scope of each assessment varies, there is no standard format for such reports. The contents and length of the report should be determined by what is most helpful to the Program Manager and by what meets the research requirements as defined by the Evaluation Administrator.

Determining what to include may not be easy since the Evaluation Administrator will look for detail while the Program Administrator likely wants only actionable items reported.

The Evaluation Administrator should work with the Program Administrator to define the types of information sought and ensure that the information and feedback is provided as quickly as possible and also included in the final process assessment.

Summary of Actions

- Ensure strong collaboration between program staff and the evaluation team by setting stage for good relationships
- Choose processes for assessment, realizing that not all processes can be assessed
- Decide who should collect process data, balancing cost, convenience and biases
- Consider the appropriate methodology when undergoing the process evaluation
- Ensure Process Evaluation Report contains all that is necessary

Technical Guide 4: Project-Level Energy Savings Guidelines

Key Points / Highlights

The objective of measurement and verification (M&V) activities at the project-level is to confirm that energy efficient measures supported by CDM programs are installed and are yielding the desired impacts, such as energy and demand savings.

Two broad categories of projects are covered in this guideline:

- those with program-supplied "deemed" savings assumptions (prescriptive or quasi-prescriptive) and,
- custom projects, which are projects that require M&V to confirm savings.

Energy efficient measures (also referred to as "energy conservation measures" (ECMs)) are a single technology, operational change, or action implemented by a customer at the customer's site. Measures can be supported or promoted through a demand-side management program. A "project" can consist of a measure or a combination of measures that, together, are designed to conserve energy. Keep in mind that measures or projects can also be undertaken voluntarily by customers, but this guideline deals with activities that are directly supported by CDM programs.

This guideline assists Program Administrators, as well as program participants, in selecting approaches and methods for estimating energy and demand savings of projects. Results can also be used to support:

- Good energy management practices by program participants
- The determination of cost-effectiveness of projects

This guideline applies to resource acquisition demand-side management retrofits, new construction, and operational change programs that result in direct energy or demand savings at a project level. Programs that produce indirect savings, such as capability building or market transformation programs, are not covered by this guideline. For details on Behavioural Program guidelines, refer to **Technical Guide 10: Behavioural-Based Evaluation Protocols**.

A balance must be found between the needs of the Program Administrator and eventual evaluation requirements and the costs of M&V borne by both participants and the program. On the other hand, the basic reporting needed for the program and evaluation purposes generally overlaps with good basic energy management on the part of energy users.

Under optimal circumstances, the Program and Evaluation Administrators would provide final approval of the program-level plan for project-level M&V. The approval of individual M&V plans, in the context of the operation of the program itself, is within the purview of the Program Administrator (and is subject to evaluation).

At the program-level, it is common to conduct project M&V studies on a representative sampling of projects, particularly for mass market programs, and to extrapolate these findings to estimate aggregate impacts at the program-level. Some programs may require M&V on the full range of projects implemented under the program. Further guidance on estimating savings at the program-level is provided in **Technical Guide 5: Gross Energy Savings Guidelines.**

Projects not directly supported by the efficiency program that are undertaken voluntarily by customers as a result of the program's influence (for example, increased awareness of energy efficiency opportunities) are accounted for in estimates of program "spillover" or other effects (**Technical Guide 8: Net-to-Gross Adjustment Guidelines**). Note that some of these results may need to be sampled for measurement and verification also.

Purpose and Scope of This Guideline

This guideline provides guidance for Program Administrators in selecting or, in some cases, prescribing evaluation methods to determine the energy savings from program-supported activities. The methods include:

- verifying the installation of energy efficient measures
- identifying factors that may affect prescriptive and quasi-prescriptive savings assumptions for measures
- improving the quality of prescriptive assumptions through technical reviews and,
- ensuring that an appropriate level of rigour is applied to M&V activities.

The Program Administrator is responsible for ensuring that the program design accommodates the need for any post-installation interaction with participants to facilitate project M&V. The Program Administrator also tracks program activity data and ensures that this information is available for the Evaluation Administrator in a usable format. Further, the Program Administrator may have to arrange for meetings or site visits to enable project M&V activities and then EM&V follow-up.

The Evaluation Administrator is responsible for providing oversight in the development of requirements for project M&V during evaluation planning (**Step 7: Evaluation Plan Development Guidelines**). The evaluation plan identifies which program-supported measures or projects will produce savings derived from prescriptive assumptions or through custom M&V methods. The evaluation plan also outlines the methods by which measure installations will be verified, as well as details regarding sampling strategies, data collection and analysis, and documentation of variances in baseline assumptions observed on site. Further, if required, the Evaluation Administrator can provide a technical review of assumptions or savings and, where appropriate, can recalculate the assumption in accordance with approved methodologies.

An Evaluation Contractor needs the following:

- Working knowledge of the International Performance Measurement and Verification Protocol (IPMVP) for energy efficiency projects
- Knowledge of measure-level assumptions (MAL) and use of measures and assumptions lists for prescribed savings (**Technical Guide 1: Using Measures and Assumptions Lists**)
- Knowledge of statistics and sample design methodologies to provide the desired levels of precision and confidence regarding the results
- Familiarity with ASHRAE or other guidelines for the measurement of technology-specific savings and,
- Certified Measurement and Verification Professional (CMVP) status is also highly desirable.

Methods Applied In Project-Level M&V

The following section outlines methods that are often used by an Evaluation Contractor in Project- Level M&V:

Review of Input Assumptions

If the prescriptive assumptions used as program inputs are new, are based on dated research or technologies, or are otherwise considered to be uncertain, a detailed review of the assumptions should be conducted. This review may occur during program planning and design, or during the program evaluation. Subsequent reviews of prescriptive assumptions are typically undertaken at least once every three years.

Detailed reviews or updates of prescriptive assumptions may also be triggered by changes in codes, standards and regulations, or by the natural introduction of more efficient products in the marketplace. A cursory review of all program input assumptions derived from the approved MALs (**Technical Guide 1: Using Measures and Assumptions Lists**) should help determine whether any major changes have occurred since the last detailed review.

When new and existing assumptions for a measure are under review as part of the evaluation or evaluation planning, the following should be considered for inclusion in the M&V study.

For **existing measures**, review input assumptions using:

- Billing, sub-metering, or engineering analyses on a sample of participants and non-participants
- Engineering calculations with M&V related to key assumptions
- Computer simulation models with M&V research related to key assumptions
- Calculations developed for quasi-prescriptive measures (for example, web-based applications) to compute savings based on customer-specific inputs

For **new measures**, determine input assumptions in advance of program implementation using information from:

- M&V study results from any relevant pilot projects
- Billing, sub-metering, and engineering analyses on a sample of potential participants
- Engineering calculations
- Computer simulation models
- Other quasi-prescriptive measure savings calculations, including ones developed by the IESO

During the process of verifying project-level savings, additional data can be collected on participant demographics, such as building or equipment operating characteristics or usage patterns. Further, findings from project-level studies can be used to substantiate differences between the baseline assumptions by improving information on efficiencies of replaced technologies, actual usage patterns, installation location, and so on, identified by participants and Evaluation Administrators.

Criteria for Selecting M&V Methods

When selecting the methods to use in a project-level M&V it is important to first differentiate the type of project. Keep in mind that programs may involve a blend of several classes of project or may involve situations not contemplated in this guideline. The Protocols should therefore be interpreted as necessary to reflect the spirit of the concepts embodied in this document.

Types of Projects

1. **Prescriptive projects** – these are projects where prescribed or "deemed" savings values are derived from the approved MALs (**Technical Guide 1: Using Measures and Assumptions Lists**) with additional documentation and analysis to establish the number of installations.
2. **Custom projects – equipment retrofit only** – these are projects where efficiency gains are achieved by the retrofit or replacement of equipment, without changes in operations.
3. **Custom projects – operational change only** – these are projects where energy consumption (and possibly demand) are reduced by changing the operating periods, settings, or methods, without modifications to equipment.
4. **Custom projects – equipment retrofit and operational change** – these are projects where the combination of equipment and operational changes may impact load and energy separately or energy directly.
5. **Custom projects – multiple energy conservation measures (ECMs)** – these are projects where three or more ECMs are implemented at a single site or facility. Multiple ECM's may enable the use of whole facility metering to determine savings.

Project Characteristics

Selection of the appropriate M&V method within any project type depends on a number of project characteristics. Five distinguishing characteristics can also be used to assist in selecting the M&V processes. These characteristics should be considered when developing M&V approaches for program-supported measures that do not exactly fit any of the basic project types described in this guideline.

1. Project Size

Project size may be based on:

- the incentive level (for example, dollars) for the particular energy conservation measure (ECM), per participant or for the whole program. When considering incentive levels:
  - small is under $10,000,
  - medium is from $10,000 up to and including $50,000 and,
  - large is greater than $50,000
- the participant's investment for the particular ECM, where:
  - small is under $10,000,
  - medium is from $10,000 up to and including $100,000 and,
  - large is greater than $100,000
- the savings (kWh or kW) expected by the participant for the particular measure(s) or project(s) installed.

The definitions of small, medium and large are intended as a guideline only. Program Administrators must provide definitions of the project size classes if these criteria are to be used as determinants of M&V methods.

2. Regularity of operating periods

Regularity of operating periods is a characteristic used where operating patterns are driven by routine events and the periods can be estimated with ease and accuracy. If operating periods vary irregularly because of variability in weather or plant production levels, precision must be applied when measuring the operating periods.

3. Persistence of savings

Persistence of savings is a characteristic used where the continuing success of the retrofit is uncertain (for example, control changes subject to human interaction). Note that it is inherently risky to base incentive payments and savings estimates on one-time observations. In these situations the reporting period should be extended and projects should be re-evaluated at least once.

4. Incentive base

Incentive base is a characteristic used when the basis for incentive payment is demand (kW). In such cases the analysis must consider the fraction of the equipment or the sub-system load that is normally operating when the site utility meter hits its monthly peaks ("diversity factor"). Energy savings (kWh) based incentives must consider the load of the equipment and normal annual operating hours.

5. Size of savings relative to utility meter total use

Size of savings relative to utility meter total use is a characteristic used where expected savings are small compared to total usage recorded on a meter; sub-meters may need to be added so that savings can be identified with reasonable precision. Suitable accuracy of meters and/or sampling strategy to yield reasonable results. Statistical analysis may be needed to select meters and sample sizes that will yield appropriate precision and confidence in findings.

The project characteristics are used to select appropriate M&V strategies from the following list:

- Using the Prescriptive Measures and Assumptions List (**Technical Guide 1: Using Measures and Assumptions Lists**)
- Conducting user survey or site investigation of the number of installations
- Carrying out site measurement by spot readings at representative times, or continuous readings through at least one full cycle of operations
- Estimating interactive effects between the energy efficiency measure and electricity uses not measured as part of the M&V
- Estimating diversity factors, or logging of load patterns and utility meter profiles at times of peak utility usage
- Reporting "Normalized Savings" (under long term "normal" conditions), rather than under actual conditions of the reporting period. Note that adjustments must be made to the baseline period and to the reporting period data to restate it under such normal conditions. The normal set of conditions is defined by each participant for its operations.
- Choosing the most appropriate IPMVP Option when retrofit isolation techniques are not suitable.

Methods for M&V on Prescriptive Measures/Projects

As noted, prescriptive measures/projects are defined as those for which energy or demand savings per item are contained in the MALs (**Technical Guide 1: Using Measures and Assumptions Lists**).

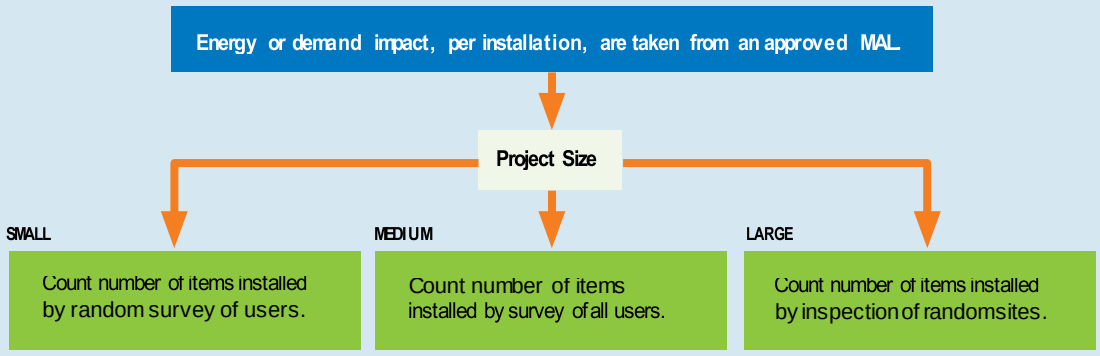
No field measurement is needed to determine the savings per measure or project. Gross impacts are determined by multiplying the per measure values derived from the measures and assumptions list by the number of installations.

The method of counting measures depends upon the size of the overall project:

Methods used for determining the number of measures in a project generally depend on the size of the project:

- **Small projects** – for small projects one can use participant self-reporting by questionnaire/survey of randomly selected participants
- **Medium projects** – for medium projects one can use participant self-reporting by questionnaire/survey of all participants
- **Large projects** – for large projects one can inspect randomly selected sites within homogeneous groups of all participant sites. Thus, achieving an overall precision of +/-10% at 90% confidence level

Figure 2.0
Prescriptive Projects



Methods for M&V on Custom Projects

Four categories of custom projects are considered here:

- projects involving equipment retrofits only
- projects involving operational change
- projects involving equipment retrofit and operational change
- projects involving multiple energy conservation measures

Depending on factors like the amount of anticipated savings, project size, or the incentive amount, the guidance and flow charts that follow are intended to help with the selection of appropriate methodologies for completing M&V on a measure or project basis.

Keep in mind that M&V plans and their reported findings are used to verify that: measures have been installed; are working as planned; and are generating savings. These custom project savings can be assessed by:

- isolating the retrofit,
- measuring the whole facility, or
- using computer simulations.

Installations can be verified through a combination of site visits and participant surveys to ensure reported results match actual impacts.

Methods for M&V on Equipment Retrofit Only Projects

These are custom projects involving only retrofit or the replacement of baseline equipment with more efficient equipment. In such projects no changes are made to operating periods, settings, or methods. If both retrofit and baseline equipment have load values shown in the MAL (**Technical Guide 1: Using Measures and Assumptions Lists**), these values are used for baseline and reporting period loads.

For equipment not on the MALs, one time measurement(s) must be made using meters of sufficient accuracy to allow the computed raw change in load to be reported with a precision of +/-10% and a confidence level of 90%.

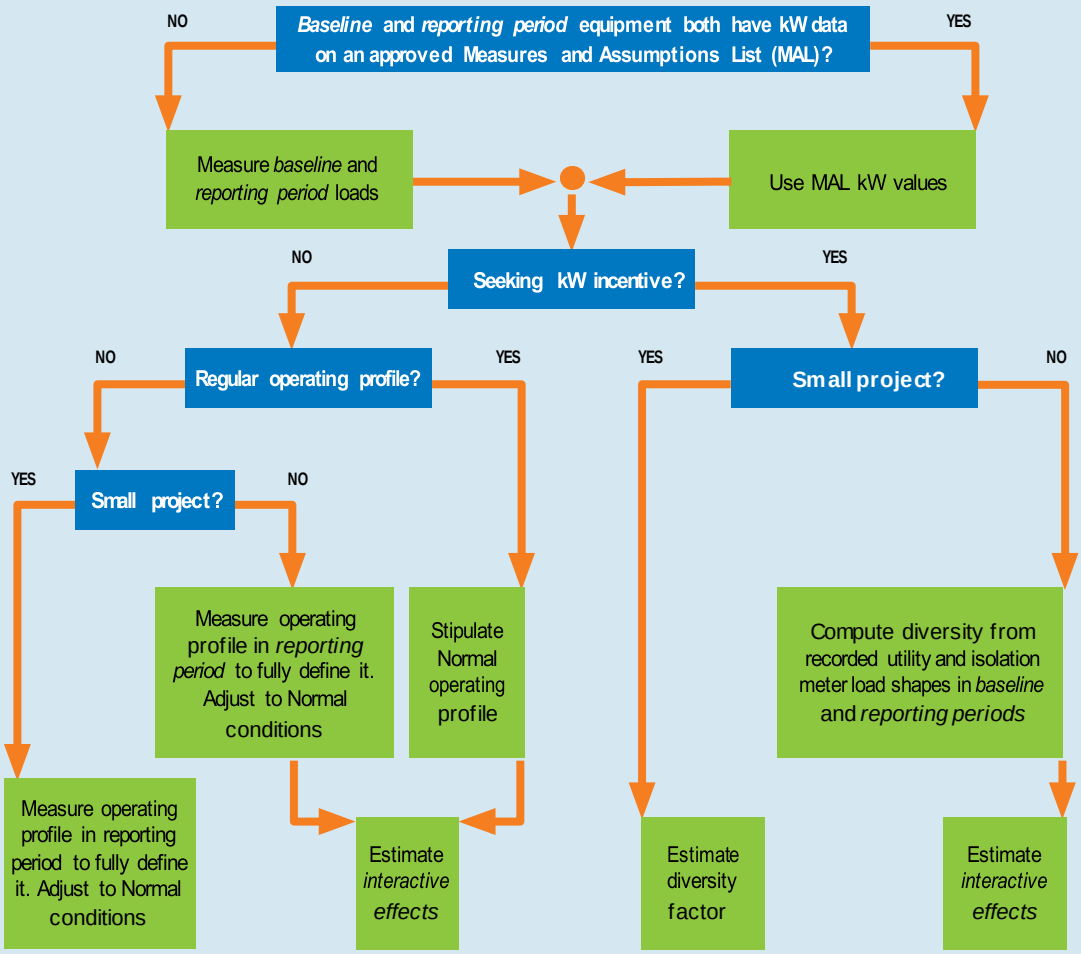
If an incentive is being used to impact energy **demand** (see also **Technical Guide 6: Demand Savings Calculation Guidelines**), the way to undertake M&V depends on the project size:

1. For small projects, multiply the baseline and reporting period loads by an estimated diversity factor.
2. For medium or large projects:
   - multiply baseline and reporting period loads by a diversity factor determined by recording the summer and/or winter demand profiles of the particular piece of equipment being retrofitted and the associated utility meter and,
   - estimate the interactive effects of the retrofit beyond the boundary of measurement.

If a **consumption incentive** is being used, the change in load is multiplied by the normal operating period. Again, the way to undertake M&V depends on the project size:

1. For small projects the normal operating period:
   - may be assumed, where the operating profile of the equipment before and after retrofit is implemented or,
   - Where the operating profile is not regular, M&V should be estimated from measurements taken at two separate points in time (at a minimum) representing the range of the normal operating pattern.
2. For medium or large projects:
   - the normal operating period should be estimated from continuous measurement throughout the full range of governing conditions after the retrofit is carried out and,
   - an estimate should be made of the interactive effects of the retrofit.

Figure 3.0:
Custom Projects: Equipment Retrofit Only



Methods for M&V on Projects Involving Only Operational Change

Such projects are custom projects that involve only changing equipment operating periods, settings, or methods. No equipment replacements or retrofits are involved. If the equipment whose operation is being changed has load values on a published MAL, the values on the list may be used. Otherwise measure equipment load once with a wattmeter having a precision of +/-5% or better, at a confidence level of 90%.

If a **demand incentive** is being used (see also **Technical Guide 6: Demand Savings Calculation Guidelines**) the way to undertake M&V depends on the project size:

1. For small projects, a diversity factor must be separately estimated for both the baseline and reporting periods and adjusted to normal operating conditions.
2. For medium or large projects:
   - determine separate diversity factors for both the baseline and reporting periods by recording the summer demand profiles of the particular piece of equipment and the associated utility meter and,
   - estimate the interactive effects of the project, beyond the measurement boundary.

If a **consumption incentive** is being used, the equipment load is multiplied by the change in operating periods between baseline and reporting periods derived as described below:

The baseline period's operating period is determined as follows:

1. If the operating profile is regular, measure it once and project it to normal conditions;
2. Otherwise, if operating profile irregular:
   - for small projects, measure the operating profile at two separate points in time representing the range of the normal operating pattern, being sure to adjust the operating profile to normal conditions.
   - for medium or large projects, measure the operating profile continuously for one cycle and adjust it to the operating profile of normal conditions.

The reporting period's operating period is determined as follows:

1. If the operating profile is regular, or the project is small, measure the operating profile once and adjust it to normal conditions
2. Otherwise:
   - for medium sized projects, measure the operating profile at two separate points in time representing the range of the normal operating pattern. Adjust the operating profile to normal conditions.
   - for large projects, measure the operating profile continuously for one cycle and adjust it to an operating profile of normal conditions.
   - for medium and large projects, estimate the impact of interactive effects beyond the measurement boundary.

Figure 4.0
Custom Projects: Operational Change Only 1 demand (kW) incentive

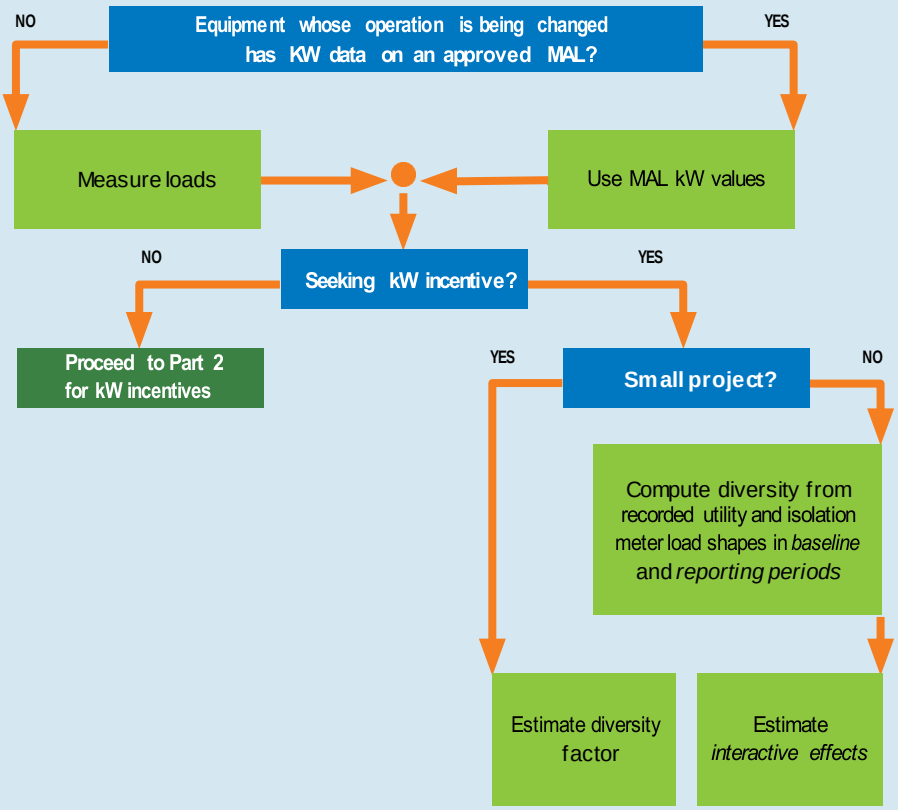
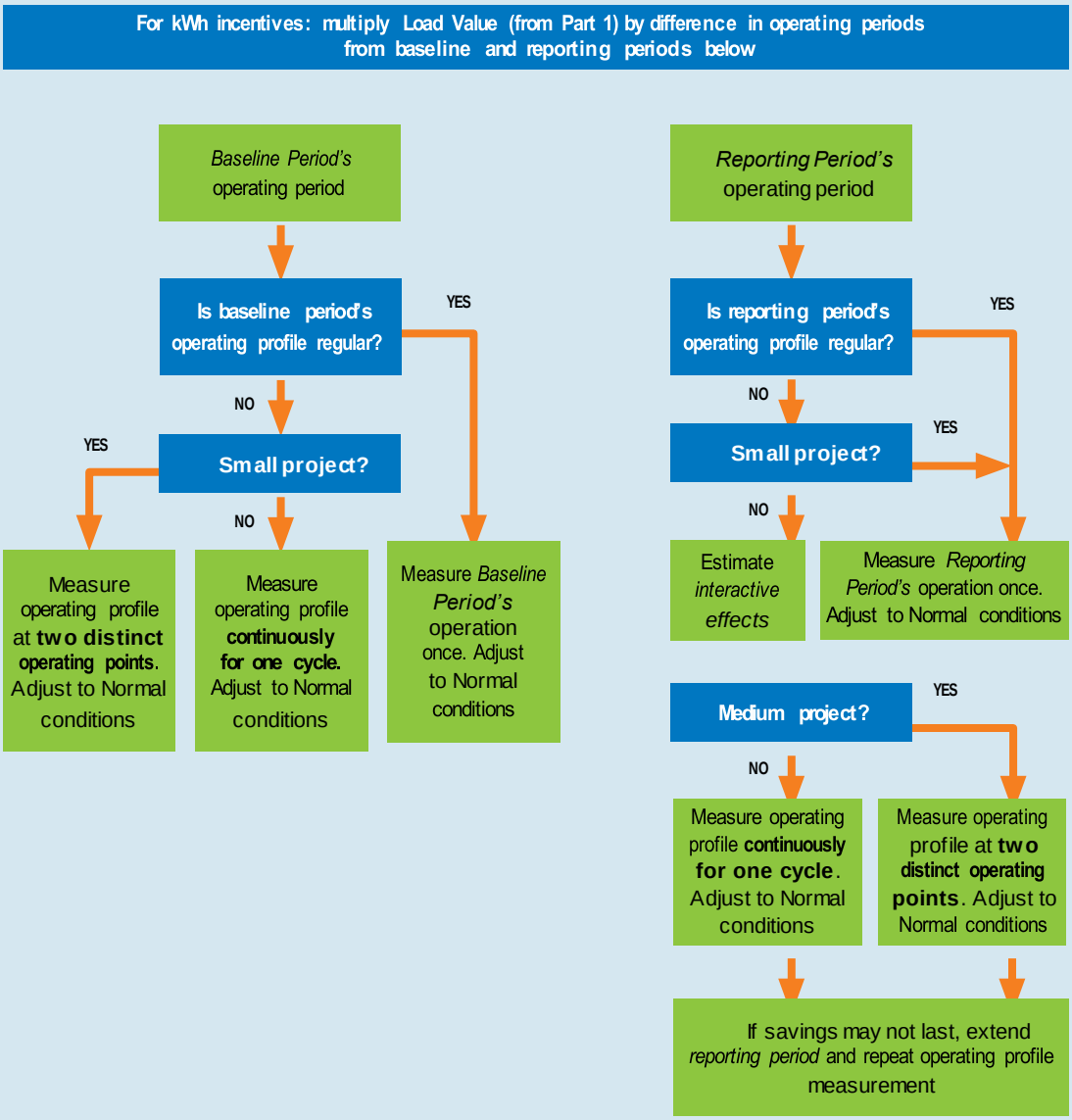


Figure 5.0
Custom Projects: Operational Change Only 2 energy (kWh) incentives



Methods for M&V on Equipment Retrofit and Operational Change Projects

These projects are custom projects involving both the retrofit or replacement of baseline equipment and a change in operational periods, methods, or settings.

There are two ways to undertake M&V for such projects:

1. If savings are highly likely to continue over time, or the project is small in size:

- If both baseline and reporting period equipment have load values shown in a current published MAL, use the MAL values to determine the loads; or
- If both values are not in a MAL, take one time measurement(s) using meters that are of sufficient accuracy to allow the computed raw change in load to be reported with a precision of +/-10% and a confidence level of 90%.

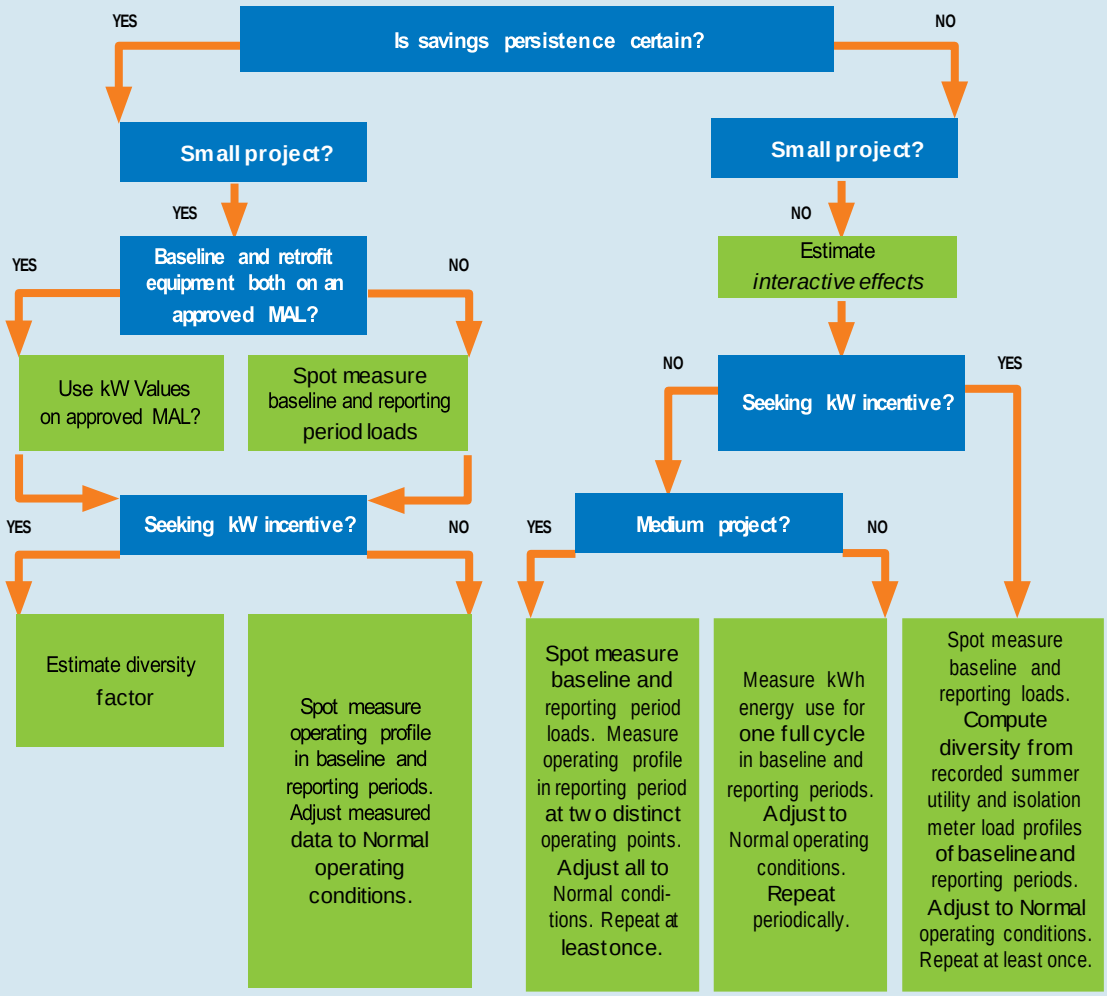
In either case:

- for kW incentives, estimate a diversity factor.
- for kWh incentives, measure operating profiles in baseline and reporting periods. Adjust all measured data to normal conditions.

2. If savings may not continue over time, or the project is medium or large in size:

- estimate interactive effects, and
- for kW incentives
  - take one time measurement(s) of baseline and reporting period loads using meters that are of sufficient accuracy to allow the computed raw change in load to be reported with a precision of +/-10% and a confidence level of 90%. Multiply the loads by diversity factors. Determine the diversity factors by recording the summer demand profiles of the particular piece of equipment being retrofitted and the associated utility meter. Repeat all reporting period measurement and recordings at least once.
- for kWh incentives:
  - for medium sized projects, take one time measurement(s) of baseline and reporting period loads using meters that are of sufficient accuracy to allow the computed raw change in load to be reported with a precision of +/-10% and a confidence level of 90%. Measure operating profiles at two distinct operating points. Adjust all data to normal conditions. Repeat reporting period measurements at least once.
  - for large sized projects, measure energy use for one full cycle of operations in baseline and reporting periods. Adjust all data to normal conditions. Repeat reporting period measurements periodically.

Figure 6.0
Custom Projects: Equipment Retrofit and Operational Changes



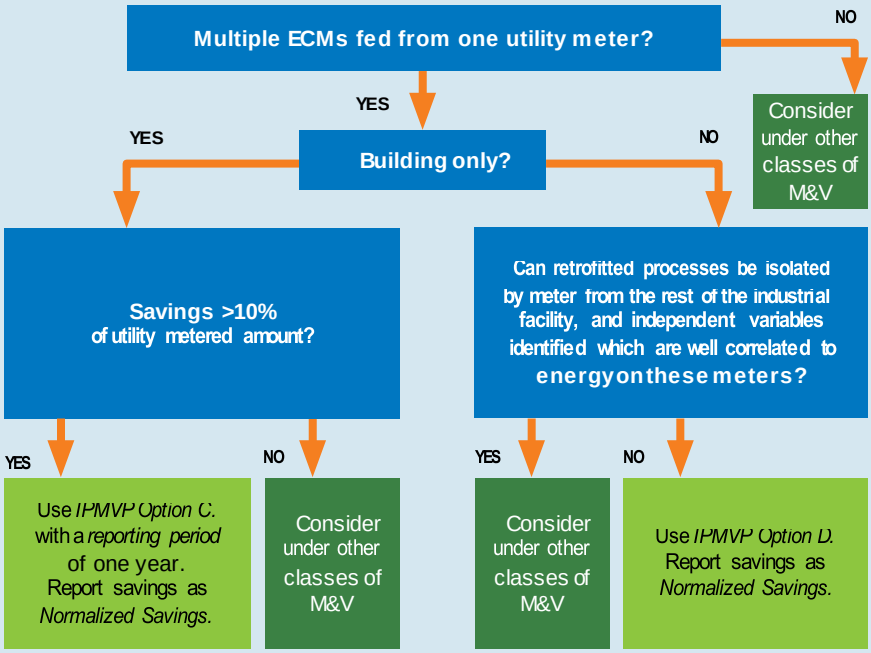
Methods for M&V on Multiple ECMs or "Blended" Projects

These projects consist of more than one energy efficiency measure. For these custom projects, special approaches can be used in certain circumstances to manage the M&V costs.

1. For buildings, where total expected savings of all ECM's is 10% or more of the affected building's consumption or demand as recorded on the utility meter, use IPMVP Option C Whole Facility. The reporting period should be one year. Savings should be reported under normal conditions ("Normalized Savings").
2. For industrial processes, where the ECMs cannot be isolated by an energy meter or reasonably correlated to independent variables related to production, use IPMVP Option D ("Calibrated Simulation"). This situation is likely to arise where multiple ECMs are installed in complex, integrated process plants. The plant must have a software-based hourly simulation model of the details of plant operations and energy use and a meter that records hourly energy for the portion of the plant being simulated. The software's calculation of energy use must be calibrated against actual hourly metered energy use. Such calibration must present a coefficient of variation of the root mean squared error (CVRMSE) of less than 30%. Savings should be reported under normal conditions ("Normalized Savings").

Site visits are required to verify measure installations for both prescriptive and custom measures. Participants included in this process are chosen using a sampling methodology with the objective of providing precision of +/- 10% and a confidence level of 90% (evaluation results that are within 10% of the actual result in 90% of cases). The evaluation team should also use the site visits to identify and document any variances in baseline conditions observed on site.

Figure 7.0
Custom Projects: Multiple Energy Conservation Measures (ECMs)



Documentation and Reporting on Project-Level Energy Savings Assessments

The M&V report should contain sections and complete descriptions of the processes used and results for all required elements in the M&V plan, including:

- Goals and objectives of the M&V Plan
- Sampling plan used to select buildings/ participants for examination, including physical and occupancy characteristics of the buildings visited or details (for example, regional) of the participants included
- Description of data collection and analysis procedures
- Estimated accuracy level of proposed assumption
- Verification and data quality procedures used to test the tracking systems
- Summary of the results and discussion of any variances or unexpected findings when the results were compared to the targets
- Documentation of the technical analysis or computer aided assumption reviews undertaken and the associated findings
- Recommendations for how the results should be used to adjust prescriptive/quasi-prescriptive or custom project savings
- Overall summary recommendations for improvement of process of the program in future years

In all cases, the Evaluation Administrator will provide final sign-off on the M&V report and its associated findings.

Timing of Analysis of Underlying Assumptions

To ensure the viability of the measures included in a program and to ensure their corresponding cost-effectiveness. The analysis of the underlining assumptions used to assess energy and peak demand savings is most beneficial when determined before program launch. However, if this is not feasible (for example, when billing analysis or meter reading is required) these assessments should take place within an appropriate period of time after program launch and the results should be used to update program impact forecasts.

The Program Administrator decides whether a third-party review of prescriptive assumptions is needed. The decision is based on a variety of factors including:

1. Previous independent review(s) of the input estimates
2. The expected magnitude of the programs savings
3. Third-party or intervener concerns over assumptions
4. Issues uncovered during literature review (**Technical Guide 1: Using Measures and Assumptions Lists**)
5. Scope of, and budget for, the evaluation

If an assessment is required, an Evaluation Manager investigates the current assumptions to either verify or re-estimate and the key inputs. Where appropriate, findings from this process should be used to update applicable measure assumptions (**Technical Guide 1: Using Measures and Assumptions Lists**).

Summary of Actions

- Review the input assumptions
- Select the project level M&V method based on the type of project and project characteristics
- Ensure M&V report contains necessary descriptions of processes used and results for all requirements
- Decide whether third-party review of assumptions is required

Technical Guide 5: Gross Energy Savings Guidelines

Key Points / Highlights

Energy savings as a direct result of CDM program activities is a key element to the establishment of energy efficiency as a reliable system resource.

Purpose and Scope of This Guideline

This guideline provides information about methods that can be used in CDM program evaluations to develop accurate estimates of the energy savings resulting from program activities. The goal is to produce energy savings estimates that are accurate within reasonable levels of precision and confidence (in most cases within 10% of the actual result at a 90% confidence level).

This guideline applies to all CDM programs that have the objective of producing direct energy savings (that is, resource acquisition programs).

The guide expands on the information set out in **Technical Guide 4: Project-Level Energy Savings Guidelines**. Therefore, where possible, we recommended that the same evaluation team perform or provide oversight for the requirements relating to both guidelines.

An Evaluation Administrator is typically responsible for fulfilling the requirements of this guideline through an approved evaluation plan (see **Step 7: Evaluation Plan Development Guidelines**). The evaluation plan details the methods for assessing program-specific energy savings. The Evaluation Administrator also provides the rationale for why the selected methodology has been chosen from the list of approved methodologies or why an alternative method has been proposed. The details of the methodological choices are usually developed in collaboration with the Evaluation Contractor.

Gross savings calculations are based on the difference between energy use and/or demand after the implementation of a program and an assumed set of baseline conditions that estimate what energy consumption and/or demand would have been in the absence of the program. Because there is no way to measure something that did not occur in the first place, there is no direct way to measure gross savings.

Gross savings are not discounted for free ridership or other adjustment/distortion factors (net savings).

This guideline pertains only to estimates of energy (GWh) savings. Demand (MW) savings are covered in **Technical Guide 6: Demand Savings Calculation Guidelines** and net savings are covered in **Technical Guide 8: Net-to-Gross Adjustment Guidelines**.

Gross energy savings is the change in energy consumption that results directly from program-related actions program participants take, regardless of the reasons why they participated

Evaluation Administrators should have the following skills:

- The ability to applying statistical and sample design methodologies
- Ability to calculate, using all relevant adjustment factors, program-specific cost benefit analysis (for example, total resource cost test)
- Strong research skill
- Practical abilities related to technically reviewing input assumptions.

Selecting an Approach

There are three general approaches for estimating gross savings:

- Deemed savings,
- Large-scale data analysis, and
- Custom M&V.

When choosing the methodology, the following factors should be taken into consideration:

- The program implementation strategy and the types of data that can be collected during the course of 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 MALs (**Technical Guide 1: Using Measures and Assumptions Lists**).
- The amount of energy savings expected to result from the program
- The professional judgement of the Evaluation Administrator
- Time and budget available for the evaluation

Basic Terms and Concepts

If one cannot measure the absence of energy use (savings), as noted, there is no way to directly measure gross energy savings. Energy savings can be estimated by comparing energy use before and after a CDM program is implemented. Equation 1 shows the general formula that applies when calculating energy savings for all energy efficiency programs.

Equation 1

Energy savings =

baseline energy use – reporting period energy use +/- adjustments

Where:

- **Baseline energy use** is the energy consumption that is estimated to have occurred before the program was implemented. The baseline period is selected to be representative of normal operations.
- **Reporting period energy use** is the energy consumption that occurs after the program is implemented.
- **Adjustments** account for independent variables that are beyond the program implementer or participant control. 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 use). Common independent variables that are adjusted for, include:
  - Weather normalization
  - Occupancy levels and hours (i.e. hours of operations)
  - Production levels (ie. operating cycles, shifts)

1. Deemed Savings Approaches

Deemed savings approaches use agreed upon values for program-supported measures with well-known and documented savings values. Deemed savings are determined using prescriptive and quasi-prescriptive assumptions and standard equations for determining gross savings. Applying deemed savings values to individual measures is addressed in **Technical Guide 4: Project-Level Energy Savings Guidelines**.

For prescriptive and quasi-prescriptive measures, the savings evaluation depends on:

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

For quasi-prescriptive measures, the savings evaluation depends on:

- 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)

For documentation and data collection purposes additional information that should be collected during the evaluation include:

- Customer address or location
- Information on technology being replaced or retrofitted
- Information about operation of new equipment (for example, hours of operation)

Prescriptive Approach Saving Calculations

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. Gross savings are calculated based on the number of participants and/or measures installed multiplied by the prescribed savings per participant or measure. The gross savings are calculated as shown in Equation 2.

Equation 2

$$PS_{gross} = N \times s$$

where,

PS_{gross} = Gross program savings (e.g., kWh)

N = Number of tracked participants (or measures installed)

s = Prescribed savings per participant or per measure (e.g. kWh per participant)

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. 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 bases and the savings is linear, then gross program 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 program savings are calculated as shown in Equation 3.

Equation 3

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

where,

PS_{gross} = Gross program savings (e.g., kWh)

N = Number of tracked participants (or measures installed)

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 MALs. In these instances, as shown in Equation 4, that the gross program savings are calculated from the sum of the savings calculated for each participant or measure installed.

Equation 4

$$PS_{gross} = \sum_{i=0}^{N} (ps_i)$$

where,

PS_{gross} = Gross program savings (e.g., kWh)

N = Number of tracked participants (or measures installed)

ps_i = Savings reported for the i^{th} participant using the quasi-prescriptive methodology

2. Large-Scale Data Analysis Approach

Large-scale data analysis applies a variety of statistical methods to measured facility energy consumption meter data (almost always whole-facility utility meter billing data) and independent variable data to estimate gross energy and demand impacts.[1] Meter analysis approach usually involves analysis of a census of project sites, versus a sample.

Most analyses of meter data involve the use of comparison groups. "Quasi-experimental design" has traditionally been used in assessing the impacts of programs. They compare the behavior of the participants to that of a similar group of non-participants – the comparison group – to estimate what would have happened in the absence of the program.

There are three basic large-scale meter data analysis methods employed for energy efficiency programs:

- **Time series comparison** - compares the program participants' energy use before and after their projects are installed. With this method the "comparison group" is the participants' pre-project consumption.
- **Use of comparison group** - compares the program participants' energy use after projects are installed with the energy use of non-participants. This method is used primarily for new construction programs, where there are no baseline data.

3. Custom M&V Approaches

Custom M&V approaches are used when no prescribed measures are found on the MALs for the types of measures included in a program. Custom M&V approaches require that gross 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 gross 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 approach (see **Technical Guide 4: Project-Level Energy Savings Guidelines)**.

1 National Action Plan for Energy Efficiency (2008). Understanding Cost-Effectiveness of Energy Efficiency Programs: Best Practices, Technical Methods, and Emerging Issues for Policy-Makers. Energy and Environmental Economics, Inc. and Regulatory Assistance Project. <www.epa.gov/eeactionplan>

For custom M&V approach evaluations, evaluators will need to collect the following information:

- Type(s) of equipment 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

M&V activities 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

At the project-level, the approach is typically outlined in an M&V plan that should be developed before project implementation. Programs that support custom measures are typically targeted to larger customers and are likely to involve fewer projects.

Gross savings can be determined by:

- Selecting a representative sample of projects for review
- Determining the savings generated by each project in the sample using one of the options described in the International Performance Measurement and Verification Protocol (IPMVP) and guidance provided in **Technical Guide 4: Project-Level Energy Savings Guidelines**.
- Applying the savings from the sample of projects to the entire population of projects

Documentation and Reporting Gross Energy Savings

A final evaluation report related to gross energy savings should include details as to how gross savings were determined. The final report should include information about:

- Methodology or methodologies used to assess gross savings
- Sampling plans and survey instruments used to collect data
- Precision and confidence of data and results
- Total gross savings and sample calculations
- Explanations, where possible, of variances between verified results and forecasted results for the program

The Evaluation Administrator reviews the final estimate of savings demonstrated through the study, which is provided by the Evaluation Contractor.

Timing of Gross Energy Savings Calculation

Completing a program-level estimate of gross savings takes time, the amount of which will depend on the analytical approaches selected and whether it will be necessary to gather and model a full range of data to complete the analysis (for example, 12 months of pre- and post-implementation electricity bills or one or more full operational cycles). The choice of the data collection period should be an explicit issue identified in the program evaluation plan (**Step 7: Evaluation Plan Development Guidelines**), as it relates to how frequently the calculation is made. Results should be reported in a timely manner to support the objectives of the evaluation.

Oversight and Responsible Parties

The Evaluation Administrator approves the gross savings methodologies used and is accountable for ensuring the analysis is completed on schedule by the Evaluation Contractor. The analyses are typically carried-out by the Evaluation Contractor and reviewed by the Evaluation Administrator. A broader evaluation team may be part of the review process. It is essential that the Program Administrator establishes a tracking system to facilitate this analysis and provides the Evaluation Administrator and Evaluation Contractor with all requested tracking system outputs and/or read-only access to the tracking system itself.

Summary of Actions

- Select an approach for estimating gross savings
- Ensure the appropriate equation is used for calculating gross savings
- Ensure the final evaluation report includes details of how gross savings were determined

Technical Guide 6: Demand Savings Calculation Guidelines

Key Points / Highlights

The Demand Savings Calculation Guideline establishes the framework for assessing demand savings attributable to specific conservation initiatives.

This guideline applies with regard to all energy efficiency programs designed to achieve energy or peak demand savings (Demand Response programs have a separate procedure).

The Evaluation Contractor is responsible for finalizing the methods used to estimate net demand savings for the program.

The Evaluation Administrator is responsible for reviewing the Evaluation Contractor's proposed plan for calculating demand savings and for signing off on that plan.

The Evaluation Contractor needs the following skills:

- Proficiency with statistical and sample design methodologies
- Familiarity with load shape analysis principles and assumptions
- Market research capabilities
- Technical ability conducive to the understanding of the operational functionality of efficiency measures (for example, peak demand effects)
- Ability to use models to forecast energy usage and ability to translate data into end-use and sector-level load shapes

Definition of Peak

The concept of peak demand is not simply the highest demand for electricity in a 24 hour period. Instead, the concept 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. So, the first step in determining peak demand (and peak demand savings) is determining the pre-defined blocks of hours during which demand is generally at its highest.

In order to maintain consistency from the program design and approvals stage, through to program operations and reporting, and finally to EM&V and verified savings, we use a before the fact (ex ante) definition of peak. Actual (ex post) system demand data is not used for the purposes of defining system peak, (it can, however be used as a reference to ensure that, over time, the ex ante definition of peak is valid.)

The hours that count towards savings targets should be known in advance and remain constant for the full program cycle. It is possible that actual system conditions will vary to a small extent over the framework period.

Though more accurate for in-year savings calculation purposes, normalized system forecasts are used to develop blocks of hours that ensure an extremely high likelihood that the top hour or top-10 hours of system peak will occur within the block(s). Benefits from the clarity and predictability of a block definition include: (a) better ability to track progress-to-target while a program is in-market and (b) greater likelihood that the Program Administrator and participants will comprehend the connection between various measures under consideration and the value they provide to the system (the basis for the cost-effectiveness of the programs).

Table 1.0

IESO EM&V Standard Definition of Peak for Calculating Demand Savings

Based on analysis of Ontario System Hourly Load data from 2003-2010, the defined summer and winter peak blocks for 2019-2020 are as follows:

Average Load Reduction over Entire Block of Hours

	Time	Months
SUMMER (Weekdays)	1pm - 7pm*	June
		July
		August
WINTER (Weekdays)	6pm - 8pm	January
		February
		December

*Daylight Savings Time-Adjusted

Because of Ontario's unique geography (vast distance from north to south and mid-latitude, full four season climate) and load characteristics, the system peak could occur in either season. Though summer peak has been dominant in recent years, it is not predicted to continue and there is a chance that the system will experience a winter peak.

Declaration of Peak Savings

Since both summer and winter peak savings have the potential to contribute to reducing the Ontario system peak, Evaluation Administrators should calculate both peaks. For example, automobile block heaters and space cooling/air conditioning provide straightforward examples of winter and summer peak-affecting measures (or initiatives or programs). Street lighting, though used all year around, would be highly coincident with the winter peak block period, but not at all with the summer block. Some measures or programs may be equally suitable for both blocks, so the selection of which one is not particularly important.

Note however, that savings for measures/programs that contribute to both block periods are not double-counted towards system peak. A declaration of the period that savings should be counted towards should accompany program funding approval. Peak demand savings results tracking and program evaluation then flow from that declaration.

Ontario also straddles summer and winter peak in terms of various parts of the province. Depending on the regions, some areas may remain summer peaking (for the foreseeable future) and northerly areas could remain winter peaking, despite the fact that the Ontario system peak could occur in either of the seasons. A program's deployment of summer or winter peak demand reduction is *not* dependent on the served areas peaking characteristic, but rather on the program's design to target a reduction in either summer or winter peak consumption.

Evaluation Administrators are encouraged to use the standard definition of peak described in **Table 1.0**, since it is the definition that will be used for verified savings calculation and

reporting purposes. Program administrators who choose to use a definition(s) of peak that varies from this one would be advised to employ a methodology to assess the gap between reported program savings and verified/evaluated savings. This gap should be predictable. In other words, a known risk factor that contributes to a gap between reported savings and eventual verified savings should be analyzed and documented so that there are no surprises at the end of the process.

Estimating Demand Savings during the Peak Period

Peak savings estimates are to be based on the average demand reduction across the total number of hours in the appropriate peak summer or winter block (refer to **Table 1.0: Definition of Peak for Calculating Demand Savings**) for block definitions). Note that because impacts across the total number of hours in each block are averaged, the peak blocks for the summer and winter do not comprise the same number of hours. Technologies that provide sustained demand reductions across the entire block have more value to the system than those that are variable. This is by design, since the chance of the actual (ex post) peak occurring in one hour versus another within the defined blocks is roughly equal. Therefore, measures or programs that better sustain savings across the span of the defined block have more value to the electricity system than those that provide a more limited sustained impact.

Maximum monthly demand reduction, typically described as "at design conditions" and/or the top facility hour of the month, in each of the three months (instead of the average of the entire block of hours) is for weather-sensitive measures because their load impact characteristics improve coincident with the system peak, since it is also weather-sensitive. For non-weather sensitive measures, using average impacts ensures that variable impacts are properly accounted for. But weather-sensitive measures are highly likely to produce their maximum impact at the same hour that was the actual top system peak hour (either summer or winter, depending on the measure). Weather-sensitive measures can therefore be properly accounted for their performance relative to periods of electricity system stress by using a much narrower – 3 individual hour in this case – definition of peak.

Other variable loads may also use this approach. Since the weighted average is structured to have no bias (advantage or disadvantage), Program Administrators and those managing M&V plans should feel free to compare and use this alternative approach. If the peak demand

Table 2.0
Alternate Definition of Peak

An alternative method can be used to calculate peak demand savings for weather-sensitive measures or for facilities with variable load characteristics. Peak demand savings are calculated on the basis of a weighted average of the maximum demand reduction in each of the three months that occurs within the blocks:

Weighted Average of the Monthly Maximum Load Reduction**

	Time	Months	Weighting[3]
SUMMER (Weekdays)	1pm - 7pm*	June	30%
		July	39%
		August	31%
WINTER (Weekdays)	6pm - 8pm	January	65%
		February	16%
		December	19%

*Daylight Savings Time-Adjusted
**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

savings credited are higher using one approach versus the other, one should use the approach that produces the higher impact. The higher impact should be used not simply because it is higher – it should be used because it will produce a more accurate assessment of the peak demand savings. For the purposes of preparing verified savings estimates, Evaluation Contractors should promote the method they believe produces the highest confidence result regardless of which approach was taken by Program Administrators.

Direct Methods for Computing Peak Demand

1. Collect hourly energy use data from a sample of participants before and after the measure installations, providing an estimate of the peak demand reduction performance of a specific measure
2. Collect hourly energy use data from a sample of locations where the efficiency measure has been installed and compare it to corresponding representative non-participant locations and use the variance to estimate the impacts on peak demand.

Indirect Methods for Computing Peak Demand

1. Allocate annual energy savings into one or more time of use periods using secondary data on average end use load shapes from past IESO evaluation results, forecasting models, or other relevant studies. Average demand savings can then be determined by dividing the energy use savings allocated to that period by the number of hours in that period.
2. Using the results of energy simulation models, allocate daily or annual energy savings for a measure or set of measures into time of use periods.
3. Estimate total peak savings for prescriptive measures installed based on the per measure values in the most recent MALs.

Valuation of Peak Demand Reduction

Since the cost-effectiveness of CDM program activity is premised on the avoided cost of generation, the power plant that would theoretically get built in Ontario to serve the marginal peak demand and energy that is being saved by the programs operates in both our summer and winter constrained periods. Setting aside some complexities of winter versus summer system capacity constraint characteristics, heat rates and other technical issues, the same "avoided cost dollars" build the same "peaking" plant that might operate primarily in the summer in years when Ontario's peak occurs in the summer and then switch to the winter if it was a winter peaking year. We don't hypothetically build a second plant to deal with a switch to winter peaking characteristics, either temporary or permanent.

Therefore, as is the case with savings impacts, double-counting avoided cost would be inappropriate. Given the accepted methodology for calculating cost-effectiveness, Program Administrators must use their earlier declaration of which peak block of hours the initiative is designed to impact. The value of energy savings is unaffected, but the avoided cost of capacity may vary by time period and therefore that value would be applied to the appropriate peak hour used for the avoided capacity cost calculation.

Report Content and Format

The initial elaboration of peak demand calculation issues should be addressed in the overall Evaluation Plan (Draft and/or Final). The final Evaluation Report should include the following:

- Clarification of program/measure-selected definition of peak demand
- Methodology used to assess program demand savings and program cost effectiveness
- Sampling plan (as well as the survey instrument) used to collect data and discussion of confidence interval
- Peak Demand Savings Results (Summer and Winter), including forecasts, reported energy savings, and verified energy savings levels (where applicable)
- Net Peak Demand Savings Results (Summer and Winter) adjusted for external factor including forecasts, reported and verified energy savings (where appropriate);
- Analysis of variances between forecast, reported, and verified demand savings

In all cases, the Evaluation Administrator must sign off on the estimation of peak demand savings demonstrated.

The Evaluation Administrator, once they have signed off on the peak demand analysis plan, as outline in the Final Evaluation Plan, is accountable for ensuring the analysis is completed on schedule. Once complete, the Evaluation Administrator must sign off on the estimation of peak demand savings. However, the analysis itself will be carried out by the Evaluation Contractor.

It is essential that the Program Designer establish an appropriate tracking system to facilitate this analysis and provide the Evaluation Contractor with all requested tracking system outputs. Once completed, the Evaluation Administrator informs program designers and delivery agents of key findings from the final demand impact analysis. This feedback is crucial, as it helps the Program Designer:

1. improve on existing program designs;
2. develop accurate initial peak demand savings forecasts; and
3. make decisions about funding and incentive levels provided for the program or similar programs.

Summary of Actions

- Choose method for estimating peak demand savings
- Sign off on Evaluation Contractor's proposed plan for calculating demand savings
- Provide Evaluation Contractor with requested tracking system output
- Ensure report provides required information and details
- Use key finding from report to consider ways of improving program design

Technical Guide 7: Market Effects Guidelines

Key Points / Highlights

To be a candidate for a market effects evaluation, the intended market effects should be a distinct part of the program strategy, an intended outcome of the program and have goals or targets forecasted. Ideally, the program administrator should be able to show that a share of the program budget or other resources was allocated with market effects as the intent.

Where substantive market effects are anticipated, simple net-to-gross ratios (NTGR) may prove inadequate. In their place, a market effects study should be commissioned to explore changes in market structure and attitudinal changes that contribute to a higher standard of practice.

Experimental Approach to Determining Market Effects

Evaluation Administrators should conduct in-depth interviews with Program Administrators, trade allies, and program participants to better appreciate the potential outcomes of the planned program design. These interviews should record changes made or changes expected in both the attitudes and abilities of each market actor as a result of the program offer.

The behaviours of market actors should be monitored and all significant changes recorded. The behavioural changes should then be correlated against the variables such as participant activities, perspectives and abilities.

And as a final step, the Evaluation Contractor must ultimately establish causal attribution leading from the program activities, through the realized outputs, accumulated through program outcomes and then to the intended impacts. This attribution pathway provides a foundation that allows Program Designers to assert that a program has broad market effects and creates market transformational savings.

Using the analytical approaches supported by this market transformational model, broad market effects can add significantly to program savings estimates and positively adjust net-to-gross ratios for both measures and programs affected by market changes.

Analytical Methods Used to Determine Market Effects

Market effects analyses require greater effort than the more typical cross-sectional analyses. Market effects, by their very nature, contribute savings year-upon-year following even a single market intervention.

In the later stages of market transformation, when the market interventions have ceased, the market effects evaluations serve as the program offer and leads to energy savings. The analytical methods applied to measure the market effects selected should take into account this longer-term horizon.

Methods for Analyzing Market Effects

- **Longitudinal analyses** – these enable Program and Evaluation Administrators to compare one pre/post period. Therefore, key market externals must be normalized to some comparable base-year or long-term trend.
- **Market characterization studies** – these serve as the data collection instrument for both cross-sectional and longitudinal assessments. These studies effectively capture a snapshot of the market that can be used as a benchmark and/or that can be analyzed to provide normalization factors for key variables and a time series of key program performance metrics.
- **Experimental studies** – these provide valuable explanatory findings that can be used to draw conclusions and formulate program recommendations. Even narrowly focused panel studies and Delphi analysis can help build expert consensus around key issues. In-field metering studies also contribute to establishing program behavioural outcomes (by helping clarify consumer electricity end-uses and the use of energy consuming appliances). Lastly, natural occurring and planned market experiments provide evidence of causal attribution and therefore should not be ignored.

Asserting the Existence of Market Effects

Evaluation administrators must carefully weigh the potential for market effects. Studies of market effects require a significant investment of both human and capital resources. In the event a potential claim of substantial market effects is absent, a market effects study should be narrowly scoped or avoided altogether.

Still, where substantial and transformational outcomes are expected, the Evaluation Administrator should be prepared to undertake a multi-year, multi-facetted study to capture the breadth of expected market effects. Also, the Evaluation Administrator should work closely with the Evaluation Contractor to ensure the use of methods that take into account causal attribution. Where methods allow for causal attribution, it may be found that the long-term market effects lead to savings greater than the annualized impacts sought directly from the program activities.

Summary of Actions

- Consider whether market effects are likely relevant; if they are, consider carrying out a market effects evaluation
- Consider what analytical method to use in evaluating market effects

Technical Guide 8: Net-To-Gross Ratio Adjustment Guidelines

Key Points / Highlights

The "net-to-gross ratio" (NTGR) is an adjustment factor applied to estimates of gross savings (**Technical Guide 5: Gross Energy Savings Guidelines**) to account for those energy efficiency gains that are only attributable to, and the direct result of, the conservation and demand management program in question. The NTGR represents the comparison between an estimate of savings achieved as a direct result of program expenditures and an estimate of savings that would have occurred even in the absence of CDM program.

Purpose of This Guideline

This guideline provides guidance for determining NTGRs for the estimation of program net impacts. Net savings estimates are the proportion of the gross savings that would have occurred in the absence of the program. Determination is usually done at the program level, but a more refined level of granularity may be warranted in some cases.

Several factors can reduce or, in some cases, increase the net impacts attributable to a program. Deciding which of these factors to account for in an analysis of net savings is influenced by the objectives of the evaluation. Factors that differentiate net savings from gross savings are also sometimes called "distortion effects", or net-to-gross (NTG) "adjustment factors", and can include the effects of free ridership, spillover, rebound effects, and transmission and distribution losses (described below). Free ridership is the most commonly evaluated adjustment factor, followed by spillover, and rebound effects.

Participant and non-participant surveys and tracking behavioural changes can help in determining net-to-gross ratio.

Net savings are of most interest to public or ratepayer-funded programs where the responsible party is interested in the influence of the program in producing incremental savings. In contrast, a government or private-sector in-house energy efficiency program or performance contract will be much more interested in total, or gross, savings.

Program benefits used in cost-effectiveness evaluations consider the program's **net savings** as opposed to **gross savings**.

Program and Evaluation Administrators of ratepayer or publicly-funded CDM programs will be interested in estimating the net savings attributable to these programs. Program Administrators should consider likely NTG factors during the design and development of a program and in designing the program logic model. NTG factors should be considered from a risk management perspective because factors such as free ridership, detract from the savings and cost- effectiveness of program investments, while other factors, such as spillover and transmission and distribution losses, can augment savings attributable to program activities.

In selecting an evaluation approach, Program Managers need to consider the level of effort to be devoted to studying net-to-gross factors (**Step 7: Evaluation Plan Development Guidelines**). The approach is tied to the program objectives, size, and scale of the program; the evaluation budget and time available; available resources; and specific aspects of the measures and participants in the program.

Net-to-Gross Ratio (NTGR) Basic Concepts

Energy and demand savings that occur due to CDM program activities are first determined as gross savings. Program net savings are then estimated by adjusting (discounting or augmenting) the gross savings by applying a set of net-to-gross "adjustment factors," such as free ridership rates, spillover effects, and rebound effects. The aggregate effect of these factors in a program impact evaluation is represented by the NTGR.

The value of the NTGR can vary dramatically depending on the type of program; how the program is implemented in the marketplace; the number of other programs that reach similar customer classes; or other market influences, such as codes and standards. For example, participants in some programs may be largely free riders whereas other programs may have virtually no free ridership.

To determine an estimated NTG value for program design, Program Administrators should incorporate free ridership rates and spillover effects, but may choose to disregard rebound effects. However, we recommend that all net-to-gross factors be considered when estimating the value of the NTGR, especially when these factors could be significant.

Some, though certainly not all, of the common net-to-gross factors that are used to calculate the NTGR are:

There are three general categories of free ridership:

- **Total free riders** – the total of consumers that would have installed the program-promoted measures at the same timeframe, regardless of program's existence
- **Partial free riders** – consumers who would have installed measures that are more efficient than baseline, but less efficient than the program promoted-measures, or who would have installed fewer of the program- promoted measures
- **Deferred free riders** – consumers who would have installed the program-promoted efficient measures, but at a later time

Free Ridership

Free ridership is a measure of program participants that would have implemented the program measure or practice even in the absence of the program. Savings do occur as a result of free ridership, but they may not be directly attributable to the program being evaluated, and thus these effects reduce the direct impact of the program.

Spillover Effects

Spillover effect occurs when the presence of an energy efficiency program influences customers to reduce energy consumption or demand, but the incremental savings are not directly a result of the program. Non-participant spillover is sometimes called "free drivership", which is the effect of people or companies that install energy efficiency measures as a result of the effects or influence of a program, but who never collect a rebate or incentive. These behavioural changes increase the effect of the program and can partially offset the effects of free ridership.

Program Enabled Savings (PES)

Program enabled savings are energy and demand savings resulting from additional energy efficiency actions that program participants or non-participant might have undertaken because of program influence, but for which they received no financial incentives. They are often referred to as "spillover" savings.

Types of program enabled savings can include:

- Operational/process changes
- Additional equipment retrofit
- Behaviour change

How can Program Enabled Savings be calculated:

For savings to be claimed, they must be quantifiable. Quantification must be transparent, assumptions clearly stated, and back-up documentation must be accessible. The following is a list of documentation that may be requested in order to calculate and validate savings claims:

- Description of the project with contact details
- Description of the Existing Condition/ Baseline
- Description of the Efficient Condition
- Annual Savings Estimate (kW, kWh)
- Persistence estimate
- Input assumptions used (with references), Engineering Calculations
- In service date
- Operating schedules
- Process modifications
- Project cost estimates

Rebound Effect

A rebound effect is an increase in energy-using behaviour following customer action to increase efficiency. This is sometimes referred to as "snap-back". An example of rebound is when customers increase their use of equipment after they have installed energy efficient equipment, or when customers use more energy when rates are low, such as during off-peak hours.[2] For example, curtailing residential air conditioning load during a set period reduces the consumption during that period, but there is a rebound effect if the customer increases their consumption by running the air conditioner harder and longer in the hours following the curtailment to make up for the increased heat and/or humidity in the home. This rebound effect can potentially offset a major part of the energy savings of a residential air conditioning load control initiative. Of course, in that case, the rebound effect might not be of much concern if the intention is to accomplish demand savings during specified times and there is a greater benefit to reduce demand.

2 This can occur under a time-of-use rate structure or a critical-peak-pricing regime.

Electricity Transmission and Distribution Losses

Because electricity is lost through the process of transmission and distribution of energy between a power plant and a consumer, when an efficiency project reduces the electricity consumption at a facility, electricity transmission and distribution losses are avoided. As a result, the amount of electricity saved by no longer having to be generated at a power plant is actually greater than the reduction experienced at the site (note that electricity transmission and distribution losses do not come into play in evaluations or net savings calculations because they are accounted for at the public reporting stage (see **Step 12: Provincial Reporting Standards**).

Other influences that come into play when determining gross savings include:

- the effects of multiple programs operating within a utility service area or region
- overlapping effects that can occur when marketing and promotion for energy efficiency programs are broadcast in neighbouring jurisdictions or service territories (through print media, radio or television) and,
- influence of energy efficient codes and standards that reduce the availability of low efficiency equipment can have the effect of increasing free ridership.

Approaches for Determining NTGRs

There are three approaches[3] for determining NTGRs:

1. Self-reported surveys and enhanced self-reporting surveys
2. Econometric methods
3. Agreed on net-to-gross ratios

All three approaches can be used with any type of CDM program, but econometric methods require large numbers of participants. Agreed on net-to-gross ratios are the least costly approach, followed by self-reported surveys and enhanced self-reporting surveys.

1. Self-reported surveys

Self-reported surveys ask participants a series of questions to get at what actions they would have taken in the absence of the program. Estimates of spillover effects can be developed by surveying 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.

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

Table 3 Sample free ridership survey question matrix (For illustration purposes only)

Survey Question									
Required financial help?	Yes	No	No	No	No	No	No	No	No
Previous experience with technology?	–	No	No	No	Yes	No	Yes	Yes	Yes
Planned to install measure without program?	–	No	No	Yes	No	Yes	Yes	No	Yes
Program influenced install decision?	–	Yes	No	Yes	Yes	No	Yes	No	No
Free rider score	**0.0**	**0.0**	**0.17**	**0.33**	**0.33**	**0.67**	**0.67**	**0.67**	**1.0**

Source: Adapted from BC Hydro, Power Smart Partners Program Free Ridership Case Study

3 National Action Plan for Energy Efficiency (2007). Model Energy Efficiency Program Impact Evaluation Guide. Prepared by Steven R. Schiller, Schiller Consulting Inc. https://www.epa.gov/sites/production/files/2017-06/documents/evaluation_guide.pdf

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 and **Table 2: Sample free ridership survey question matrix** illustrates an example of how these types of questions can be used in conjunction with a matrix to estimate free ridership.

- 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?

Enhanced self-reporting surveys

Enhances self-reporting surveys are used to improve the quality of information used to provide NTGRs 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 NTGR or a reasonable range to apply to the program being evaluated.

2. 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 NTGR, care must be taken to select a group that is comparable to the participant group.

3. 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 NTGRs 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 NTGRs are often periodically updated based on reviews and evaluations of net-to-gross factors.

Adjusting Gross Savings to Estimate Net Savings

The net program savings are calculated in a similar manner as the gross program savings with the difference being the number of tracked participants and/or measure is discounted (or increased) by NTG adjustment factors determined through the program evaluations. The net program savings are calculated as shown in Equation 1.

Timing Of Consideration Of NTG Factors

Net-to-gross factors should be examined during the evaluation planning stage (**Step 7: Evaluation Plan Development Guidelines**). The evaluation should seek to identify and to clarify, through participant surveys and follow-up activities, the net-to-gross factors and their relative magnitudes. Net-to-gross factors are determined once, at the time of the evaluation.

Equation 1

$$PS_{net} = \sum_{i=1}^{N} (NTGR_i \times N_i \times S_i)$$

where,

PS_{net} = Net program savings (kWh/Kw)

NTGR = Net-to-gross ratio (e.g., %)

N = Number of tracked participants/measures installed

S_i = Adjusted gross savings for the i^{th} participant/measure

Note that adjusted gross savings will vary according to the various types of measures (i.e. prescriptive, quasi-prescriptive, and custom) and should account for adjustment factors (i.e. realization rate, installation rates, etc.).

Summary of Actions

- Consider whether the gross savings estimated should be adjusted by a NTGR
- Consider whether there might be free ridership, spillover effects, or rebound effects
- If a net-to-gross ratio adjustment is appropriate, consider the best approach for determining the adjustment; for example, consider whether to use an agreed-to ratio, self-reporting or enhanced self-reporting surveys, or econometric methods.

Technical Guide 9: Guideline for Statistical Sampling and Analysis

Key Points / Highlights

Generally, when studying the impact of a program it is not viable to study every single program participant. Furthermore, with respect to a comparison group (or control group), it is nearly impossible and most often not feasible to study the entire range of eligible non-participants. Therefore, statistical sampling of the two populations (participants and non-participants) is used to gauge program effectiveness.

Questions to consider when drawing samples from a population under study:

1. **What if the sample population looks or behaves nothing like the larger population?** 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. To ensure accurate representation of a population that needs to take steps to avoid bias in the 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 sample design stage.
2. **Is it ever certain that the sample population would achieve the exact results as the population under study?** In the very best case, the sample only provides an estimate of 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.
3. **What if the sample population being studied is affected by influences beyond the scope of the program offer?** Statistically significant effects may be observed even where a program has not been implemented. For example, a commercial building or industrial account may be shown to have reduced energy consumption by 20% following an economic recession. These same accounts may or may not be participants in a program designed to achieve energy savings. The question becomes what portion of the 20% is attributable to the program and what percent is associated with the economic recession and other external factors. As such, it is important to recognize that a correlation does not necessarily indicate causation.

To deal with such questions, the industry relies on a research process known as a sample design. This guideline provides a primer on this subject and provides guidance for determining what design is best suited to serve the research objectives. Consult with a statistics professional before to implementing complex statistical analysis.

Defining the Study Population

When selecting a sample, the first question that must be asked is what is the population under study? To evaluate energy efficiency programs, the first step is to decide whether the savings estimate is to be assigned at the provincial, regional or individual utility level.

This is important because a small rural customer base, for example, may be primarily single family homes, farms, and some small commercial accounts. This population would not be representative of the Greater Toronto Area; nor is it likely to resemble Ontario as a whole. Therefore, it may not be accurate to formulate a provincial savings estimate by studying the program participants from this small rural customer base. Conversely, it may not be accurate to project savings for this small rural customer base from a broadly scoped study used to establish a provincial savings estimate. As such, it is essential to describe the characteristics of the population including, but not limited to, size and variance.

The Need for Strata

How the study population is defined will determine what conclusions can be drawn from the evaluation. As a result, it is sometimes necessary to stratify (sub-divide) the population. In the example above, a provincial savings estimate is desired plus the means to allocate savings to individual groups. Therefore, it may be practical to sub-divide Ontario into strata by individual group or by stratum of different groups with similar characteristics. By dividing the population into distinct and independent strata, researchers can draw inferences about the sub-populations that otherwise would be lost in a more broadly defined sample.

If the following conditions exist, applying stratification is likely appropriate:

- Variability within the defined strata are reduced
- Variability between the defined strata are maximized and,
- Variables used to stratify the population are strongly correlated with the desired dependent variable.

These three criteria may help show that the group is not the appropriate differentiating stratum, and it could be something else.

Advanced Stratification Options

To apply stratification, information about the characteristics of the population is required. Absent prior research, the researcher will have difficulty in defining appropriate strata. If that happens, the researcher may look to more advanced statistical methods to define the appropriate strata.

The two most common advanced approaches are over-sampling and post-stratification. With over-sampling, the researcher intentionally biases the sampling process to represent a known about the population, such that the resulting findings better represent the study population; even when the population itself cannot be appropriately sampled. For example, if it is known that there is a high non-response bias from a particular demographic of participants, the researcher may want to over-sample this population or sub-population to ensure that the actual number of responses received meets statistical requirements. In addition to over-sampling, a technique known as post-stratification may be used to develop estimates about sub-populations after the study is complete and can be used if characteristics about the sub-populations are unknown at the time the study is conducted. An example of this technique may be to simply over-sample a population to develop a provincial savings estimate for a program that can later be stratified to yield savings estimates by groups or strata, if desired.

Both over-sampling and post-stratification are advanced research methods and are fraught with potential pitfalls. If applied incorrectly, these two techniques could compromise compliance with the Protocols. These advanced techniques should be reserved for specific situations and used only after careful consideration of other options. In addition, use of the methods should be well documented in the experimental approach of the **Draft Evaluation Plan**.

Sample Selection

With the population and sub-populations defined, the researcher may turn his or her attention to selecting samples representative of the defined populations. These study populations are often referred to as the sample frame.

The sample frame is simply the pool from which a sample will be drawn; ideally, this will be from the entire study population. The worst-case scenario for a sample frame is to use a population of convenience, such as individuals who have participated in an initiative, to complete a questionnaire if they choose to (the reason using such a population is not a good idea, is because those who complete the questionnaire typically are people with strong opinions or higher degrees of knowledge about the subject and therefore are not necessarily representative of the entire population participating in the initiative). As a result it is important to use the appropriate sampling technique to address such biases during sample selection. Regardless of the sample methodology chosen, it is important to always keep in mind that a sample must be drawn to represent the population under study.

Of course, there are many other sampling techniques that could be employed in the study of conservation and demand management initiatives. The EM&V Protocols allow researchers to draw from the wide array sampling techniques available, however justification and documentation should be provided with regards to the sampling method employed.

The most common probability sampling techniques used to study energy efficiency and conservation programs are:

- **Simple Random Sampling:** This involves the random assignment of members from the study population to the study sample. This could be done, for example, using a computer to randomly assign 15% of program participants to the study sample.
- **Systematic Sampling:** This involves the systematic assignment of members from an ordered study population to a sample; for example, every 12th participant entering a program may be selected for the study sample.
- **Matched Random Sampling:** This involves the selection of 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 researcher 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 researcher is asked to sample a fixed number of members that meet specific criteria and assign them to a study sample; for example, a researcher may be asked to survey 400 middle-aged women and 300 middle-aged men. Quota sampling relies on the researcher's judgement and convenience in sample selection. Because of this, quota sampling is a non-proportional (biased) sampling technique.
- **Panel Sampling:** This involves the 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.

A Situation Requiring a Non-Probability Sample

If the goal is to study electricity use across the whole of Ontario, the broad scope of such an effort would require the population to be stratified. By doing so, several sub-populations could be identified based on similar characteristics and each can be studied independently of the other.

One such stratum could be industrial or manufacturing facilities, for example. Since the sub-population of industrial and manufacturing customers is typically not a homogeneous group, a non-probability sample may be employed for this stratum while using a random probability sample for the remaining strata. Because of the inherent differences between the energy use of the various industrial and manufacturing customers, a random sampling of this stratum could lead to unintended biases, namely, the selection of unusually large or abnormally small customers whose energy use are not representative of the stratum. In this case, a subject matter expert or a sector specialist may be better able to define a representative sample of the population. For example, the sector specialist may be able to isolate from the stratum some of the odd accounts and systematically select a sample from the remaining customers that can represent the group as a whole.

By allowing a sector expert to help with the sample selection, a more accurate study of the industrial and manufacturing sub-population can be realized than would be achieved based on a simple random sample. *Non-probability* samples must be carefully considered to ensure that sampling bias is explicitly identified and kept to a minimum.

Sizing the Study Sample

Some of the main advantages of sampling are:

- sampling is less expensive than conducting a census of the whole population;
- the data can be analyzed easier and there is greater flexibility in the analytical methods that can be applied; and
- sampling can lead to greater sensitivity for the study of populations and sub-populations (as required).

However, researchers should also be aware that the trade-off to studying a sample as opposed to the entire population can lead to errors and inferences being made about the population that may not be completely accurate. Thus, it is important for researchers to be comfortable with the level of precision that their sampling strategy can provide.

One consideration that must be addressed when sampling any population or stratum is the degree of precision desired for an estimate. Another factor is the confidence level sought. An evaluation contractor may have a requirement for the savings estimate to be ± 5% at a 95% level of confidence. That is to say a repeated sampling of the population would result in a mean savings estimate that is within 5% of the true mean of the population 95 times out of 100.

To determine the required size of the study sample, the researcher must consider the desired levels of precision along with some assumptions about the normal variance around the *population mean*. Generally, the mean of the population is not known; otherwise a study of that population would not be necessary. Where the mean of the population is unknown, the variance around that mean is also unknown.

Therefore, an assumption often has to be made regarding the coefficient of variance, which is the dispersion of a probability distribution. Typically, the coefficient of variance is set at 0.5%, when other studies are not available to inform the likely variance around the population mean sought. 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*.

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. In this section we outline 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 quite a simple exercise or highly complex depending on the nature of the study. Because one or more tests may be suitable, to address a research question we recommend 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. In this guideline we focus 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.

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.

Researchers should carefully document in the Draft Evaluation Plan the rationale behind the chosen test method and should outline all calculation methodologies applied.

Table 4.0 Common Statistical Tests for Normally Distributed Populations

Goal	Possible Outcomes	
	One (Measurement)	Two (Binomial)
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

Summary of Actions

- Define the study population
- Determine whether there is a need for stratification of the population chosen
- Decide on the sampling technique that will be used
- Decide on the sample size
- Decide whether to apply a statistical test
- Ensure the report includes information relating to the test method chosen as well as the rationale for choosing that test

Technical Guide 10: Behaviour-Based Evaluation Protocols

Key Points / Highlights

This document sets forth the basic protocols that are to be used in evaluating behavioral programs. Chapters 1 - 3 introduce the protocols, describe the philosophy behind their development and outline the types of programs that are governed by the protocols that are to be applied. Chapter 4 discusses the protocols that are to be used for cost benefit analysis, process evaluations and market effects studies. Chapter 5 introduces the basic research designs that are appropriate for assessing the impacts of behavioral interventions. Chapters 6 through 9, provide protocols for designing impact evaluations for Training/Capacity Building programs, Information Feedback programs and Public Information Programs. Finally, Chapter 10 provides protocols for analyzing data from experiments and other research designed to assess the impacts of behavioral programs.

When to Use this Guide:

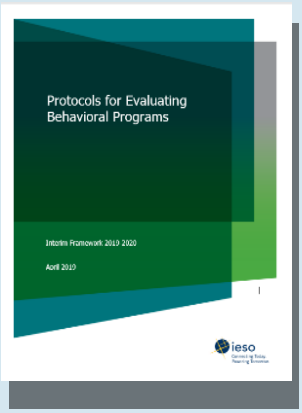
Behaviour-Based Evaluation Protocols should be employed when assessing the impact of behavioural programs on energy consumption. The following are examples of programs intended to alter behavior to achieve energy savings include:

- providing normative comparisons in which consumers are provided with comparisons of their household energy consumption with that of other purportedly similar households
- providing feedback technologies that allow consumers to observe their energy use at websites or from devices installed in their homes
- providing home automation technologies to consumers that help them consume less energy
- providing time varying rates that help consumers lower their energy consumption to reduce demand on the electric system while saving money on their bills
- providing financing for energy efficiency investments designed to encourage consumers to purchase more energy efficient equipment
- providing training to various market actors to enhance the likelihood that they properly size and install energy using equipment
- providing training to building industry professionals to assist them in designing and building energy efficient buildings

How to Use

These protocols are intended to be used by evaluators and policy makers to plan and carry out evaluations of behavioural programs. They describe best practices for evaluating such programs as well as the minimum information that must be reported regarding the selection of research methods and results. Four basic types of evaluations may be required in assessing the performance of behavioral intervention programs. They include:

- **Impact evaluations** – assessment of the impacts of capacity building programs on energy consumption;
- **Market effects evaluations** – assessments of the impacts of capacity building programs on various aspects of the market including changes in sales and prices of energy efficiency measures, prevalence of behaviors and opinions that influence energy consumption and actions that may be taken by market actors in response to the program;
- **Cost effectiveness evaluations** – assessments of the extent to which cost savings resulting from the program exceed the costs of delivering the program; and
- **Process evaluations** – assessments of the extent to which the process used to deliver the program was efficient and effective in accomplishing its intended purpose.



BEHAVIOUR-BASED EVALUATION PROTOCOLS

Glossary of General Program Evaluation Terminology

The definitions in this glossary are adapted from federal, provincial, and academic sources, many of which are listed in the bibliography at the end of this appendix.

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

Activities
A term used generically in logic modeling to describe the action steps necessary to produce program outputs.

Administrative Agency
An organization tasked with administering electric generation, transmission, distribution, reliability, and conservation programs within the Province of Ontario, such as the OPG, IESO, etc.

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 energy-demand savings outcome (gross savings) for a single unit of an installed energy-efficiency or renewable-energy measure that:
(1) has been developed from data sources and analytical methods that are widely considered acceptable for the measure and purpose, and
(2) 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 a deemed savings estimate is 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 (**Technical Guide 1: Using 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. If there are measures on deemed savings lists from other jurisdictions that are not on the official Lists in **Technical Guide 1: Using Measures and Assumptions Lists**, please request that they be analysed and added.

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. See also *Strength*.

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
The person responsible for developing an EM&V plan for a particular program or portfolio. This person is also the point-of-contact for EM&V contract management. This person is sometimes referred to as an Evaluation Manager.

Energy Conservation Measures (ECM)
An activity or set of activities designed to increase the energy efficiency of a facility, system or piece of equipment. ECM may also conserve energy without changing efficiency. An ECM may be applied as a retrofit to an existing system of facility, or as a modification to a design before construction of a new system or facility.

Evaluation Contractor
The individual(s) or firm(s) selected to implement the EM&V plan developed by the Evaluation Administrator. The Evaluation Contractor could also be referred to as the "Independent, Third-Party Evaluator" or the "Evaluator.

Ex ante load impact estimate
A load impact estimate representing a set of conditions or group of customers, or both, that differ from historical conditions (from the Latin word for "beforehand").

Ex post load impact estimate
A load impact estimate representing a set of conditions that actually occurred on a specific date or over some period of time for the customers that were enrolled in the program and called on that date or over that period of time (from the Latin word for 'something done afterwards').

Free driver (free drivership)
A non-participant who has adopted a particular efficiency measure or practice as a result of the evaluated program.

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.

8760s
Full year hourly consumption loads.

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
Energy effects created by 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.

Market Study Evaluation
A study that characterize energy markets, assess spatial and temporal changes in market structure and function that result from program interventions and other external influences (i.e., such as codes and standards, fuel price volatility, and environmental concerns).

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

Measures and Assumptions Lists
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. **Technical Guide 1: Using Measures and Assumptions Lists.**

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.

Peak demand

IESO defines peak demand as follows:

Table 1.0

IESO EM&V Standard Definition of Peak for Calculating Demand Savings

Based on analysis of Ontario System Hourly Load data from 2003-2010, the defined summer and winter peak blocks for the Interim Framework (2019-2020) are as follows:

Average Load Reduction over Entire Block of Hours

	Time	Months
SUMMER (Weekdays)	1pm - 7pm*	June
		July
		August
WINTER (Weekdays)	6pm - 8pm	January
		February
		December

*Daylight Savings Time-Adjusted

Persistence of savings

A critical element for many stakeholders is whether energy savings from the ECM and/or behavioral change continue over time. It is important to determine the value of the energy and demand savings beyond the initial program year. There are at least two different situations for which evaluators may assess persistence of savings

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.

Process Evaluation (or Assessment)
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 an 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. See *Measurement*.

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.

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.

Regulatory Authority
The entity with the mandate to oversee the actions of local distribution companies and administrative agencies; in Ontario this could be the Ontario Energy Board (OEB), the Environmental Commissioner of Ontario (ECO), or the Ministry of Energy, Northern Development and Mines (MENDM), or any combination of the three.

Reliability
The quality of a measurement process that would produce similar results on: (1) repeated observations of the same condition or event; or (2) multiple observations of the same condition or event by different observers.

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

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 savings of the participants. There can be participant and/or non-participant spillover.

Strength
A term used to describe the overall defensibility of the evaluation as assessed by use of scientific practice, asking appropriate evaluation questions, documenting assumptions, making accurate measurements, and ruling out competing evidence of causation.

Structured Interview
An interview in which the questions to be asked, their sequence, and the detailed information to be gathered are all predetermined. Structured Interviews are used where maximum consistency across interviews and interviewees is needed. Whereas unstructured interview is an interview used to elicit information in complex situations where questions can be changed or adapted to meet the interviewee's responses. Unlike structured interviews, it does not offer a limited, pre-set range of answers for an interviewee to choose, hence, the lack of consistency and reliability.

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.

Bibliography

1. California Public Utilities Commission, "The California Evaluation Framework," June 2004. www.calmac.org/publications/California_Evaluation_Framework_June_2004.pdf

2. DOE," The Performance-Based Management Handbook, Volume 4, November 2000, Appendix A. www.orau.gov/pbm/pbmhandbook/pbmhandbook.html

3. GAO, "Designing Evaluations," GAO/PEMD-10.1.4, March 1991, pp. 92-94. www.gao.gov/special.pubs/pe1014.pdf

4. GAO, "Performance Measurement and Evaluation: Definitions and Relationships," GAO/GGD-98-26, April1998.www.gao.gov/special.pubs/gg98026.pdf

5. OMB, "Instructions for the Program Assessment Rating Tool," pp. 7-10. www.whitehouse.gov/omb/part/2006_part_guidance.pdf

6. McLaughlin, J.A., and Jordan, G. B., "Logic Models: A Tool for Telling Your Program's Performance Story," Evaluation and Program Planning, Volume 22, Number 1, February 1999.

7. University of Wisconsin Extension, "Planning a Program Evaluation," February 1996, pp. 2-10. http://learningstore.uwex.edu/assets/pdfs/G3658-1.PDF

8. Ernest Orlando Lawrence Berkeley National Laboratory, Edward Vine, Jayant Sathaye, and Willy Makundi, "Guidelines for the Monitoring, Evaluation, Reporting, Verification, and Certification of Forestry Projects for Climate Change Mitigation", Environmental Energy Technologies Division, March 1999. https://escholarship.org/content/qt20h2r692/qt20h2r692.pdf

Protocols for Evaluating Behavioral Programs

Interim Framework 2019-2020

April 2019

ieso
Connecting Today.
Powering Tomorrow.

Acknowledgments

The Independent Electricity System Operator would like to acknowledge the work of Nexant Inc. in the development of this protocol, in particular the contributions of Dr. Michael Sullivan.

Table of Contents

1. Introduction

The protocols set forth in this document describe the basic approaches that the Independent Electricity System Operator (IESO) considers acceptable for assessing the impacts of behavioral programs on energy consumption.

Over the past 10 years, a variety of efforts have been undertaken to encourage energy conservation by changing the behavior of various market actors including service providers and consumers. Examples of programs intended to alter behavior to achieve energy savings include providing:

- normative comparisons in which consumers are provided with comparisons of their household energy consumption with that of other purportedly similar households;
- feedback technologies that allow consumers to observe their energy use at websites or from devices installed in their homes;
- home automation technologies to consumers that help them consume less energy;
- time varying rates that help consumers lower their energy consumption to reduce demand on the electric system while saving money on their bills;
- financing for energy efficiency investments designed to encourage consumers to purchase more energy efficient equipment;
- training to various market actors to enhance the likelihood that they properly size and install energy using equipment;
- training to building industry professionals to assist them in designing and building energy efficient buildings; and
- technical support to large organizations to assist them in identifying energy efficiency investment opportunities, designing and evaluating solutions and implementing them.

Following a recent discussion of evaluation measurement and verification for behavioral programs we define behavioral programs as those that seek to change energy use related behavior in an effort to achieve energy or demand savings.[1] These programs typically involve education, information feedback, training, awareness building or public appeals.

1 Annika Todd, Elizabeth Stuart, Charles Goldman and Steven Schiller "Evaluation, Measurement and Verification (EM&V) of Residential Behavior Based Energy Efficiency Programs: Issues and Recommendations (2012 (DOE/EE 0734

Four basic types of evaluations may be required in assessing the performance of behavioral intervention programs. They include:

- **Impact evaluations** – assessment of the impacts of the program on energy consumption;
- **Market effects evaluations** – assessments of the impacts of programs on various aspects of the market including changes in sales and prices of energy efficiency measures, prevalence of behaviors and opinions that influence energy consumption and actions that may be taken by market actors in response to the program;
- **Cost effectiveness evaluations** – assessments of the extent to which cost savings resulting from programs exceed the costs of delivering them; and
- **Process evaluations** – assessments of the extent to which the process used to deliver programs are efficient and effective.

Behavioral intervention programs are designed to change the *behavior* of market actors and thereby to cause changes in energy consumption. As such the evaluation of these programs poses special evaluation research design problems. In particular:

- Determining that a given intervention has caused a change in behavior requires the implementation of carefully designed research usually requiring experimental or quasi-experimental research techniques;
- The observation of change in behavior requires careful empirical measurements using surveys and other data that may be expensive to obtain;
- The impacts of behavior change sometimes take time to materialize (i.e., it may take longer for some parties to adopt behaviors than others);
- Efforts to change behavior do not always succeed with all parties subjected to behavioral interventions (i.e., some parties reject information or training);
- Improvements in practices adopted by some market actors as a result of training may cause other similar actors in the market to adopt those practices (i.e., spillover effects are possible);
- Behavior changes may have variable persistence; and
- Behavior changes can cause indirect changes in measure adoption rates for energy efficiency measures supported by other funding steams thereby necessitating an assessment of the attribution of the effects to the different programs that might be affected (i.e., design changes resulting from training of architects and engineers may alter the adoption rate of energy efficient appliances for which rebates are paid).

The above special considerations require the development new protocols for measuring the impacts of training and segment support on behavior and energy consumption.

1.1 The Purpose of the Behavior Protocols

These protocols are intended by to used by evaluators and program design and implementation staff to plan and carry out evaluations of behavioral programs. They describe best practices for evaluating such programs as well as the minimal information that must be reported regarding the selection of research methods and results. These protocols comprise a new component of the IESO EM&V Protocols and Requirements explicitly designed to meet the requirements for evaluating behavioral programs.

1.2 Underlying Philosophy of the Protocols

Guidance is provided concerning how best to meet the above described objectives in this document in the form of protocols. Miriam-Webster's Online Dictionary defines a protocol as: "a detailed plan of a scientific or medical experiment, treatment, or procedure." It is possible to specify protocols in three ways.

First, it is possible to prescribe the approaches that must be employed to evaluate programs. For example, California's Energy Efficiency (EE) protocols identify the specific methods that must be applied when estimating savings for EE programs in California. These are what are called prescriptive protocols because they require specific estimation procedures to be used in calculating impacts. A second type of protocol specifies the output that must be reported leaving decisions concerning research methods to be made by the researchers who are responsible for producing the required output. A third type of protocol primarily provides guidance concerning best practices and recommended approaches to research design and analysis, tailored to a particular subject matter area; for example, conservation and demand management (CDM) evaluation or outage cost estimation.

The protocols presented herein combine elements of all three types of protocols. They are intended to define the appropriate minimal requirements for carrying out valid evaluations of behavioral intervention programs while allowing researchers the leeway to design effective methods for achieving this goal.

In the discussion that follows, we focus most of our attention on research requirements for carrying out valid impact evaluations. By impact evaluations we mean evaluations intended to assess the changes in behavior and energy consumption that result from behavioral programs. We do so for the following reasons:

- Results of impact evaluations are crucial for determining whether the behavioral intervention programs are having the intended effects on behavior and energy consumption. This information is critically important for program planning and future decisions about program resource allocation.
- Research methods required to estimate the impacts of program interventions on behavior are very different from those that have been relied upon to quantify the effects of conventional energy efficiency programs. The paradigm for quantifying the impacts of behavior on energy consumption is based on observing the changes in behavior and energy consumption that occur when a behavioral interventions are provided; **not** on the reduction in energy consumption (adjusted for free ridership and spill over) arising from substitution of more efficient end use equipment for less efficient equipment. Protocols that have been adopted for studying the impacts of conventional energy efficiency programs simply are not appropriate for assessing the impacts of changes that arise from behavioral interventions. So, substantial effort must be dedicated to explaining and justifying those methods.

- When it is possible to estimate energy savings arising from behavioral interventions, the methods and procedures used to estimate program cost effectiveness are the same as those for conventional energy efficiency programs. In other words, what is different about estimating the cost effectiveness of behavioral programs is the way that energy savings from behavioral programs are estimated, not the manner in which cost benefit ratios are applied.
- Likewise, the methods and procedures used to carry out process evaluations and market effects studies are the same for behavioral programs as they are for conventional energy efficiency programs (or all other social programs for that matter).

There are "right ways" of assessing the impacts of behavioral programs on energy consumption and behaviors; and these methods and the reasons why they should be used are detailed in this document. As will be explained in detail below, these "right ways" often involve experiments designed to conclusively determine the extent of change energy consumption or behaviors as a result of exposure to the program.

However, we recognize there are sometimes intervening circumstances that make it impossible to achieve the ideal experimental design. It will be necessary to make decisions in the design process that give up some of the certainty about the outcome of interest in order to take account of practical considerations. The protocols are intended to provide guidance to research designers as they make these decisions. They call for both careful consideration of decisions that reduce the internal and external validity of experiments designed to assess program effects and careful documentation and explanation of the consequences of doing so at the reporting stage.

1.3 Description of Contents

This document sets forth the basic protocols that are to be used in evaluating behavioral programs implemented in Ontario. Chapters 1 - 3 introduce the protocols, describe the types of behavioral programs to which the protocols should be applied and discuss they types of evaluations that can be carried out for such programs. Chapter 4 discusses appropriate research designs for studying the impacts of the types of behavioral programs that are being carried out. Chapter 5 describes the protocols to be used in evaluating training and capacity building programs. Chapter 6 describes the protocols for evaluating the effects of feedback programs; and Chapter 7 describes the protocols that should be applied to evaluating the effects of education and information campaigns. Chapter 8 provides examples of the application of the protocols to three existing programs.

2. Types of Behavioral Programs

As conservation and demand management programs have emerged over the decades since the 1970s a distinction has developed between what are normally thought of as energy efficiency programs and conservation programs.

Energy efficiency programs are utility or third party sponsored policy initiatives designed to increase the market penetration of energy efficient equipment. They are programs that are designed to save energy by causing customers to use it more efficiently to provide the same level of comfort and convenience that would have been supplied by less efficient equipment. Examples of energy efficiency programs are lighting, refrigerator and air conditioner rebate programs in most markets.

Conservation programs, on the other hand are designed to cause parties to act in ways that save energy by reducing demand for it (e.g., properly installing equipment, investing in more energy efficient alternatives, setting thermostats lower in winter and higher in summer, turning off unneeded lights, loading laundry and dish washing machines to full capacity, replacing machine drying clothes with line drying, etc.).

For reasons that are unimportant to understanding the definition of behavioral programs that will be employed in these protocols, there has been a tendency for program planners and evaluators to think of energy efficiency programs and impacts as initiatives that are principally concerned with the effects of equipment on energy consumption; and to think of conservation programs as initiatives that are principally concerned with the effects of behavior or habits on energy consumption. It follows from such reasoning that savings from energy efficiency programs are deemed to arise principally from the difference in energy consumption for a lower level of energy efficiency with equipment that has higher efficiency. While savings from conservation programs are deemed to arise principally from changing behavior so that there is less demand for energy.

Whatever advantage the foregoing reasoning might have had in the preceding decades, it should be obvious that this definition of the problem has outlived its useful purpose. Today, most third party and utility sponsored programs contain important behavioral components; and in most senses can be considered to be behavioral programs.

To reflect the increasing importance of behavior change in achieving energy savings, for purposes of these protocols, we expand on the definition of behavior based energy efficiency programs adopted in the recent SeeAction report[2]. The definition of behavior based energy efficiency programs advocated in that report was:

> "Behavior based energy efficiency programs are those that utilize strategies intended to affect consumer energy use behaviors in order to achieve energy or peak demand savings. Programs typically include outreach, education competition, rewards benchmarking and feedback elements.
>
> Such programs may result in changes to consumers' habitual behaviors (e.g., turning off lights) or one time behaviors (e.g., changing thermostat settings). In addition, these programs may target purchasing behavior (e.g., purchase of energy efficient products or services) often used in combination with other programs)…"

2 Op cit 1

In our view, the above definition is too limited. In addition to consumers the scope of the target markets for behavioral programs should to include operators, installers, lenders and other market actors so that the revised definition is:

Behavior based energy efficiency programs are those that utilize strategies intended to affect energy use behaviors by *consumers, operators, installers, lenders and other market actors* in order to achieve energy or peak demand savings. Programs typically include outreach, education competition, rewards benchmarking and feedback elements

Such programs may result in changes to habitual behaviors (e.g., turning off lights) or one time behaviors (e.g., changing thermostat settings). In addition, these programs may target purchasing behavior (e.g., purchase of energy efficient products or services) often used in combination with other programs) *as well as other behaviors related to the selection, installation and operation of building systems.*

While there are a number of different kinds of behavioral programs, there is an immediate need to develop protocols for three basic types of behavioral programs. These types include:

- Training/Capability Building Programs;
- Information Feedback Programs; and
- Education/Awareness Campaigns;

These programs differ fairly dramatically in terms of the behavioral outcomes of interest and the mechanisms that will be used to stimulate impacts. As a result, the details of the measurements that must be taken to assess impacts and approaches to experimental design may differ somewhat from program type to program type. In the following sections, the different types of behavioral programs are discussed in detail along with current examples of such programs in the utility industry.

2.1 Training/Capability Building Programs

Training and capability building programs are designed to cause energy savings by providing training to installers and building operators by ensuring that systems for which they have responsibility are properly installed and operated. These kinds of programs have been in existence for literally decades in most localities that have established serious public efforts to enhance building energy efficiency. As a matter of fact, they were some of the first efforts that most utilities undertook to encourage efficient energy use in buildings.

While it is self-evident that training key market participants should lead to improvements in the operating efficiency of critical building systems, there is a surprising lack of empirical evidence supporting the proposition that such training encourages the installation of more efficient equipment or causes buildings to be operated more efficiently. Outcome measures of interest for training/capacity building programs include:

- Subscription rates to training courses (i.e., 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 the courses;
- Pass or certification rates for students taking courses; and
- Observed of the energy efficiency of systems installed or operated by students before and after they were trained.

2.2 Information Feedback Programs

Feedback is an important element in any effort to control human behavior. As the old management saying goes, one cannot manage what one cannot measure. Correspondingly, feedback based energy saving programs have been under development in the utility industry for decades. Early examples of feedback programs include monthly volumetric electric bills; and reports to customers attempting to characterize the sources of their energy use and recommend actions to lower their bills (e.g., Xencap). While the above feedback mechanisms have been in the market for many years, more recently, attention has been focused on the following evolving feedback strategies:

- **Periodic printed reports based on normative comparisons** – periodic (monthly, semi-monthly or quarterly) reports to customers comparing their energy use and costs with that of customers who are reputed to be neighbors or to be similar to the target customer.
- **Periodic Bill Alerts** – weekly messages by email, SMS and IVR informing customers of their usage up to a given date possibly in relation to a pre-established usage goal
- **Triggered Bill Alerts** – messages to consumers by email, SMS and IVR informing consumers that their usage is abnormally high or will exceed some designated value that they have identified in advance.
- **Web based feedback** – providing information about customer usage and tips on the web.
- **In Home Displays** – devices that communicate with advanced meters through Zigbee, Wi-Fi or internet and display electricity and/or gas consumption in various formats in near real time.
- **Home Area Networks** – devices that allow customers to control thermostats, lights and motor loads in their homes and businesses using internet and smart phone apps.
- **Optimizing thermostats** – similar to home area networks except that they are designed to analyze customer demands for heat and cooling based on response to thermostat setting changes and discover and schedule the optimal operating schedule based on occupancy and observed temperature preferences.

All of the above feedback mechanisms are being tested in utilities throughout the world using more or less robust evaluation practices. Some have been shown by replication to reliably and significantly alter customer energy consumption.

2.3 Education/Awareness Programs

Education and awareness programs have been a central part of efforts to encourage energy conservation and the efficient use of energy for decades. These programs vary in size and scope from societal level efforts like the Energy Star Change a Light, Change the World Campaign program in the US (sponsored by the US Environmental Protection Agency) to smaller scale efforts by local and regional governments, local distribution companies and service organizations focused on specific market segments (i.e., schools, municipal governments, business organizations, etc). These education/awareness programs have in common the fact that they typically involve a highly structured approach to developing and transmitting specific messages to specific target populations using well developed communications strategies. They usually involve:

- **Planning** – including defining the goals and objectives of the education/awareness effort, assessing resource requirements, obtaining resources and cooperation from organizational leadership, assembling a project team, etc.
- Careful design and implementation of an information campaign including:
  - identification of specific opinions, perceptions and behaviors that are to be affected by the campaign;
  - formulation of specific messages that are to be transmitted using surveys focus groups and other measures to evaluate message content intended to change behavior;
  - identification of channels to be used to transmit messages;
  - determination of actions needed to bring about the information campaign; and
  - management of the campaign.

Evaluation of results including estimation of changes in behavior by comparing survey responses from the target population before and after exposure to the information campaign and change in energy use when possible

Outcome measures for education/awareness programs normally include observed changes in reported behaviors, opinions, perceptions and knowledge regarding the issues that are the targets of the campaigns. However, in some circumstances it may be possible and desirable to directly measure changes in energy consumption arising from education/awareness campaigns. This can occur, for example for programs targeted at changing the energy use of organizations using information campaigns.

3. Types of Evaluations

In evaluating behavior intervention programs four types of evaluations may be undertaken including:

- **Impact evaluations,**
- **Market effects evaluations,**
- **Process evaluations, and**
- **Cost effectiveness evaluations.**

The methods and procedures required to assess the impacts of behavioral interventions on behavior and energy consumption are quite different from those ordinarily used in evaluating energy efficiency programs. The objective of behavioral intervention programs is to alter behavior and thereby to alter energy use. The impact of the programs is two pronged behavior change impact resulting in energy savings impact. Both of these aspects of behavioral intervention programs should be thought of as program impacts; and they should be directly measured. The protocols outlined in chapters 5-7 of this document outline the protocols that are to be used in assessing the impacts of behavioral programs.

Although behavior has been classified within the market effects paradigm historically, very little else from the market effects paradigm is useful in evaluating behavioral programs and the cost of true market effects evaluations makes them unattainable in the context of most behavioral program evaluations. So it is best to simply treat the behaviors of interest as program impacts.

Evaluation research projects for behavioral programs may also involve process evaluations, cost effectiveness evaluations or even market effects studies. The methods required to carry out these types of evaluations differ dramatically from one another and from the methods used in evaluating behavioral interventions. However, the methods and procedures for carrying out market effects evaluations, cost effectiveness evaluations and process evaluations for behavioral programs are the same as those used in the evaluations of all other types of energy efficiency programs. So there is no need to develop new protocols for carrying out these types of evaluations in the context of behavioral intervention programs. Indeed, it is appropriate and necessary that the protocols for carrying out these kinds of studies for behavioral programs be the same as those used for other types of energy efficiency programs, so that the results of studies of these behavioral programs can be compared with those of standard energy efficiency programs.

In the event that behavioral programs require process evaluations, cost effectiveness analysis and market effects studies, standard protocols from the IESO EM&V Protocols and Requirements should be applied.

The appropriate protocols for these types of evaluations are as follows:

- **Process Evaluation Protocol** – IESO EM&V Protocols and Requirements, Process Evaluation Guidelines.
- **Market Effects Protocol** – IESO EM&V Protocols and Requirements, Market Effects Guidelines
- **Cost Effectiveness Protocol** – IESO Conservation and Demand Management Cost Effectiveness Guidelines

4. Research Designs for Observing Impacts of Behavior Programs

This chapter is a basic introduction to the research design alternatives that are appropriate for assessing the impacts of behavioral intervention programs on behavior and related energy consumption.

It is designed to be read and used by program managers and analysts who need to understand the basic principles involved in program evaluation and the basic research strategies that are appropriate when evaluating behavioral programs. For parties seeking a more in-depth treatment of the subjects taken up in this chapter we recommend reading the following books and technical reports:

- *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* by William Shadish, Thomas Cook and Donald Campbell; Haughton Mifflin 2002.
- Evaluation Measurement and Verification (EM&V of Residential Behavior Based Energy Efficiency Programs: Issues and Recommendations by Annika Todd, Elizabeth Stuart, Charles Goldman and Steven Schiller; SEEAction Network 2012
- Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols: by Michael Sullivan and Stephen George; EPRI Report 1020855 2010

The first resource above is an excellent high level discussion of evaluation research design with particular attention to the application of quasi-experimental designs to situations when it is impossible to carry out randomized experiments. The second resource is an excellent discussion of the issues that arise when evaluating programs designed to change behavior. The third resource provides protocols that are particularly useful for evaluating programs designed to alter consumer behavior using feedback.

The material in this chapter draws heavily from these resources and attempts to present a high level summary of all of the issues found in those resources.

4.1 Measuring Changes in Behavior – the Problem

Behavioral programs as set forth in the foregoing chapter are designed to cause changes in energy use related behaviors by individuals and organizations. The behaviors of interest are myriad. Examples might include:

- Consumer decisions to purchase more efficient equipment;
- Consumer decisions to use more or less electricity;
- Consumer decisions about the timing of their electricity use;
- Practices used by HVAC sales and service technicians to specify the size and design of new and replacement HVAC systems;
- Actions taken during the installation, maintenance and operation of mechanical and lighting equipment;
- Choices of building envelope materials, mechanical systems and lighting systems made by designers and builders of low-rise residential buildings which produce an embedded level of energy efficiency;
- Choices of building practices that influence energy consumption; and
- Choices made by large organizations to identify and adopt energy efficiency improvements.

As explained above, behavioral intervention programs are designed to change specific behaviors within the above categories by applying social science theories that suggest that changing the conditions under which behavior is occurring will modify it. It is reasonable to imagine that these interventions are capable of causing market actors to change their behavior resulting in a change in energy consumption. But in reality, we don't know and cannot predict *how much* behavior change or change in energy consumption will occur without testing the effect of the intervention on the target persons or organizations. *The central problem in evaluating behavioral programs is to discover how much change (if any) results when behavioral interventions are presented.*

In virtually all cases in which an effort is made to change behavior, to measure the impact of a program on behavior we must discover *what would have happened* if the program had not existed. By comparing the behavior that is exhibited when the behavioral interventions are present (e.g., training or support) with the behavior that is exhibited in the absence of the interventions we can determine how much change in the outcome variable of interest (behavior or energy consumption) occurred as a result of exposure to the intervention.

The most robust strategy for assessing the impacts of an intervention on behavior is to create an experiment in which it is possible to (1) ensure that the intervention occurs before the behavior change occurs; and (2) ensure that no other causal factors may have produced the change in behavior that is observed. Experimentation is not always possible, and when it is not, there are alternative methods -- generally referred to as quasi-experimental techniques – that can be used with some success to assess the impacts of interventions on behavior. These techniques are almost certainly inferior to experiments in virtually all cases and require much more skill and talent on the part of researchers to reach valid conclusions, but sometimes they are all that can be done.

The protocols set forth in this document call for the use of both types of research designs – depending on the situation. When possible, experimental designs involving random assignment of target market actors should be used. When this is not possible, quasi-experimental techniques should be used.

These protocols are intended to provide guidance in the development of all kinds of training and support programs. As such they rest on the assumption that the evaluator understands the basic tenants of research and experimental design. The remainder of this chapter reviews the logical underpinnings of these techniques.

4.2 Principles of Experimental Design

Three conditions must be met in order to *conclusively prove* that a behavioral intervention (e.g., providing training or support) has caused a change in behavior (e.g., use of best practices in design and installation of HVAC systems):

- The behavioral intervention has to *precede* the behavior change in time.
- The behavioral intervention must be *correlated* with the behavior change – that is, when the intervention is present the behavior change occurs, and when it is not present, the behavior change does not occur.
- No other plausible explanations can be found for the behavior 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 (i.e., behavior 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 behavior before and after exposure to a treatment (e.g., training). This is known as a pretest-posttest design. This kind of design is seldom employed because of weaknesses described below. However, it is useful as a framework for discussing the sources of inferential error that can arise when certain critical elements of experimental design (i.e., randomization of exposure to experimental treatments) are ignored.

During a pretest-posttest 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 behavior 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 cause us to think the 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 behavior 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 behavior is observed at two points in time, the subject of our observation has gotten older and it is possible that something about the aging process has caused the change in the behavior that is observed, and not the treatment. Maturation can influence behavior 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 something 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 we observe a difference in behavior 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 behavior 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 know they are being observed; and it is unusual for experiments with humans to be undertaken without their being aware of it. They are particularly likely to occur with repeated measures (e.g. classroom tests) in which it is possible for subjects to learn the correct answers during the testing process.

- **Instrumentation** – when we observe a difference in behavior at two points in time, it is possible that the calibration of the instruments used to measure the behavior has changed – producing the appearance of a behavior 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 behavior (for the worse). Calibration problems are even more likely to occur with survey instruments and other self-administered behavioral 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 we observe a difference in behavior 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 cause us to believe 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 behavior change related to training.
- **Censoring** – censoring is like 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 behavior 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 (e.g., 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 behavior when nothing more than censoring has occurred.

The above inferential problems all occur because conditions other than the treatment can cause changes in behavioral outcome measures (e.g., installation practices or annual energy consumption) when the effect is measured by comparing observations of a *single group* at two points in time (i.e., 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 (e.g., trainees, consumers or organizations) at two points in time, the impacts of the experimental variable are observed by comparing the behaviors 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 problems mentioned above will affect both groups in about 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 behavior. In doing so, the threats to experimental validity described above will be completely eliminated.

Of course, the assumption that both groups are similar is a very big "if". The *drawback* to inferring cause from differences between groups is that the *groups may not have been exactly the same to begin with*. If they were not, then any observed difference between them could simply reflect the pre-existing difference. This last major threat to internal validity is called selection:

- **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 a potentially very important source of inferential error that must be controlled in experiments involving capacity building.

The above seven problems are what have been described as threats to the internal validity of experiments. If left uncontrolled, they 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 (i.e., 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 are, at the minimum, a waste of time and money. 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 behavior in a desired manner. They are discussed below.

4.2.1 Control

Control is completely central to the design of experiments. By taking control of the timing and exposure of subjects to experimental factors thought to change behavior, it is possible to ensure that the experimental factor occurs before the onset of the desired behavior. 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 behavior can be controlled in a variety of ways to observe their effects. Often, 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 behavior (e.g., one might imagine for example training programs targeted at the same audience lasting different periods of time or being presented in different formats). So it is possible to imagine experiments that range from very simple comparisons between the behaviors 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 to groups exposed to different levels of the treatment variable. When groups are compared to observe an effect of a treatment, the most fundamental assumption is that the groups are sufficiently similar at the outset of the experiment so that any difference after exposure to the experimental factors can be deemed to have resulted from the factor and not some 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 for all intents and purposes statistically identical before the experimental factor (treatment) is presented. Typically this is done by *randomly assigning* subjects to comparison groups (i.e., 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.

4.2.2 Stratification

In evaluating the impacts of a behavioral intervention on energy use related behavior 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 (i.e., sales technicians and installation technicians,). Breaking up experimental groups (i.e., treatment and control groups) into sub-groups based on criteria that are observable in advance of an experiment is called stratification.

Table 4-1 describes a simple experiment involving stratification on trade.

Table 4-1: Simple Stratification Example

	Training	No Training
Sales staff	n1	n5
Installers	n2	n6

In addition to providing useful information about the effects of experimental treatments within sub-populations of interest (e.g., 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 change in behavior (particularly energy use) 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 change in behavior within the sub-groups – ignoring the differences between the sub-groups.

4.2.3 Factoring

Sometimes behavioral interventions consist of treatments that contain more than one factor. For example, it is often the case that behavioral interventions intended to change energy consumption contain a technology component (e.g., a field computer or device that simplifies application of a given installation protocol) and an information component (e.g., 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 the technology and the training in changing the behavior 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 (e.g. 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. Table 4-2 describes a simple factoring experiment in which two treatment variables with two levels are examined.

Table 4-2: Simple Two Factor Experiment Example

	Technology	No Technology
Training	n1	n3
No Training	n2	n4

In the experiment, subjects would be randomly assigned to one of four groups n1-n4 in sufficient numbers to be able to estimate the differences in the outcome behaviors 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 while 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 temptation to create such complicated experiments involving many factors and strata should be approached cautiously because of the inherent difficulties encountered in carrying out complex experiments.

4.2.4 Replication

Perhaps the single most important tool for evaluating the impacts of behavioral interventions is replication. Replication is said to occur when the conditions involved in an experiment are repeated in order 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 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 something individual investigators should consider in designing evaluations it is a very powerful tool that should be used to assess the veracity of research findings at the program level; and in evaluations of behavioral 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 (e.g., among the various Local Distribution Companies (LDCs) implementing programs) sequentially or simultaneously. Evaluators carrying out behavioral experiments across multiple LDCs should be encouraged to design their experiments as replications of a single administration.

4.3 True Experiments

True experiments are research designs in which the evaluator has control over the exposure of experimental subjects to treatments. There are three kinds 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 behavioral interventions on energy use related behavior. They are discussed below.

4.3.1 Randomized Controlled Trials (RCT)

The RCT is an evaluation research design in which experimental subjects are randomly assigned to treatment and control groups; and the results observed for the groups are compared to discover whether the treatment has caused a change in behavior. The process of random assignment causes the resulting groups to be statistically identical on all characteristics prior to exposure to the treatment to within a knowable level of statistical confidence given the sample sizes being employed. This is true because each and every observation being assigned to both groups has the same probability of being assigned to each group (i.e. 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 and so on and so on for virtually all the variables one can imagine – whether we can observe these variables or not.

Of course, because sampling is involved, the above statement is true to the extent that relatively large samples are involved and even then only to within a certain level of statistical confidence. Indeed, anything can happen in the real world – which means that even with truly random assignment with large samples it is possible to create treatment and control groups that are not statistically identical. So it is good practice to check to make sure 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 measurement 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 behavior change. Several reasons underlie this designation. They 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 specification of estimation equations or error structures. They are often as simple as a difference in differences calculation. 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 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 the course of 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, outmigration 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 (i.e., more than 1 or 2 percentage points) then it is necessary to analyzed 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 (e.g., 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 the lottery are not too high, this strategy can overcome objections that stakeholders may have to randomly assigning subjects to test conditions. This approach is particularly useful when the experimental treatment (e.g., 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 to 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 which is scheduled to receive the treatment in the second year. This approach can be implemented without causing significant customers dissatisfaction. However, because the control group must also receive the treatment in the second year, it will result in higher cost for equipment and support than the recruit and deny approach.

4.3.2 Randomized Encouragement Designs (RED)

Sometimes regulatory or administrative considerations require that *all* subjects who are eligible to receive some behavioral intervention must receive it if they desire it. For example, administrative policy might dictate that all qualified HVAC technicians have access to training that would result in their receiving a certificate that can provide 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 behavioral 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 (e.g., training) 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 measurement of the difference in the behavior between those who were intended to be treated and those who were not intended to be treated. Because encouragement was randomly assigned, any difference between the encouraged and not encouraged group must necessarily have resulted from the fact that the encouraged group contains more parties who received the treatment. Because we know the acceptance rate in the encouraged group, it is possible to inflate the observe 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 a "silver bullet" for overcoming the difficulties that are often cited with the application of RCT designs in evaluations related to energy use behavior. Unfortunately this is not the case. As in the case of the RCT design, there are certain cautions that 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 that changed behavior or energy consumption, it is logically possible that encouragement stimulates some other actions that either enhance or attenuate the observed effect of the treatment; and 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 sample sizes required to detect effects using a RED design. In a RED, the measurement of the impact of the treatment on behavior is diluted because some (in many cases most) of the parties who were encouraged to be treated did not accept the treatment. So, it is possible that only a small portion of the subjects who are encouraged to be treated actually 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. So, 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 actually received the treatment. So, if the actual behavior 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 behavior change and the effectiveness of encouragement, the RED design may require much larger sample sizes in treatment groups than the conventional RCT. In cases where the effect of the treatment on behavior and the acceptance rate for the treatment are in the single digits, the sample sizes required to detect the resulting difference between the behavior 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 and even then these designs should be used only when relatively large impacts on behavior and energy use are expected.

4.3.3 Regression Discontinuity Designs (RDD)

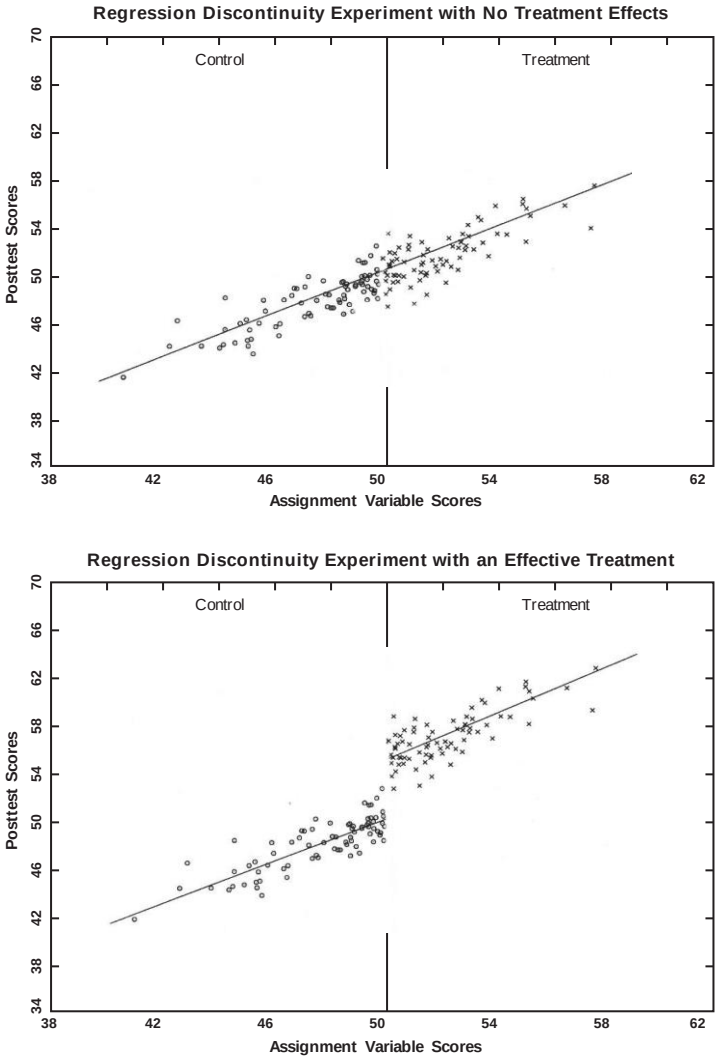
In the two true experimental designs discussed above (RCT and RED) subjects are randomly assigned to experimental groups – thereby establishing their statistical similarity. Under certain circumstances, assignment of subjects to treatments can be non-random provided subjects are assigned to treatment and control groups precisely on the basis of 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 doesn't vary systematically with the outcome variable at all. It doesn't matter. In fact, it can be shown that the RCT is just a special case of the RDD where the assignment variable is a random number (e.g., 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. Figure 4-1 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 discernible 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 Figure 4-1 by the difference on the horizontal axis between the projections of the two intersection points on the vertical discontinuity indicator.

Figure 4-1: Example of Regression Discontinuity



The RDD is an extremely powerful tool that can be used when subjects must be assigned to treatment conditions based on some pre-existing qualification. It controls all of the possible alternative explanations for the observed program effect. However, there are certain important caveats that 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 (i.e., 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 in fact 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, resulting 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 and one in which the effect of the experimental treatment will be seriously underestimated. Protecting against this possibility requires estimating non-parametric (nonlinear) regression functions, which imposes an additional complexity.

4.4 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 behavioral 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. These problems commonly occur in experiments involving training.

When 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 on the basis of a pre-existing interval level variable) as is the case with the true experiments all plausible alternative explanations (e.g., 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 as 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 an 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.

There are several types of quasi-experimental designs that are particularly important in behavioral 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
- Interrupted time series designs
- Within subjects designs

4.4.1 Non-equivalent Control Groups – Matching

In true experiments, subjects are assigned to treatment and control groups in such a way that they are either known to be statistically identical prior to 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 rests 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 exactly 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 (e.g., 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 a very old idea 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 favors 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 something happens (e.g., signing up for training) based on information about readily observable independent variables that may be correlated with selection into the treated group (e.g., 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 exactly 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.

4.4.2 Within Subjects

All of the preceding experimental designs rest on the comparison of the behavior exhibited by groups of subjects who have been exposed to treatment with behavior exhibited by groups that have not been exposed to a treatment (control groups). The difference between the behaviors 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 really 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 really 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 circumstance 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 (e.g., training) by comparing the behaviors 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 and every experimental subject is exposed to all levels of the experimental factors under study as well as the absence of the experimental factor (i.e., the control condition). Under the appropriate conditions this is a very powerful quasi-experimental design because it completely eliminates the possibility of selection effects *because it completely eliminates the control group.*

4.4.3 Interrupted Time Series

Another quasi-experimental design that is appropriate to studies of the impact of behavioral interventions on energy use related behavior is the interrupted time series design. An interrupted time series design consists of repeated measures of the behavior of interest before and after a treatment has been administered. This design is particularly useful when variables related to usage or other frequently measured behaviors 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 behavior) 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 (i.e., it is definitely known that treatment commenced abruptly at a time certain).
- 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 statistical analysis is required (as it almost always is).

Like all comparisons that rest entirely on observing the difference in behavior 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 (i.e., 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 (e.g., 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 (i.e., 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.

5. Evaluating Training/Capacity Building Programs

Capacity building programs are social interventions designed to lower energy consumption in residential and commercial buildings by providing training and technical assistance to various market actors who design, install, operate and service systems that influence energy consumption in buildings; and by providing expert advice to organizations to assist them in identifying and implementing energy efficiency improvements.

Below are some examples of capacity building programs:

- **Residential builder training** – training and incentives designed to encourage residential builders to incorporate energy efficiency and green attributes into new residential buildings. Program is targeted at company executives, designers, marketing staff, site superintendents, framers and insulators.
- **HVAC installation optimization training** – training and incentives to HVAC contractors to encourage them to apply best practices in designing and installing residential and small commercial air conditioning and heat pump installations.
- **Energy Manager Training** – training for energy managers working in large commercial or industrial organizations.
- **Energy Efficiency Service Provider Support Initiative** – support to energy service providers and support organizations for delivering energy services to various market segments (e.g., health care, refining, forestry, mining, etc.). Services will include: identification of savings opportunities, preparation of energy management plans, assistance in identifying and promoting incentive programs and applying for incentives, promotion of effective energy management practices, and delivery of training, outreach and advice regarding opportunities for energy savings.

While it is self-evident that training key market participants should lead to improvements in the operating efficiency of critical building systems, there is a surprising lack of empirical evidence supporting the proposition that such training can encourage the adoption of more efficient technology, ensure that equipment is properly installed, will cause buildings to be operated more efficiently or cause significant energy saving measures to be adopted by organizations. This is so because the existing paradigm for evaluating energy efficiency programs doesn't provide for a reasonable means for quantifying the impacts of these and other efforts to alter energy consumption by changing behavior.

Training programs are, as the name suggests, generally involve classroom training courses intended to enhance the ability of various actors in the market to cause reductions in energy use. The training varies dramatically from market actor to market actor, but the intended outcome is the same – reductions in energy consumption. Segment support programs provide specialized consulting services to different market segments (e.g., government and industries) to assist them in identifying opportunities for achieving energy savings, planning, financial assessments, management presentations and other

services that may enhance the rate at which energy efficiency investments are achieved. The objective of these programs are to inject expertise into organizations to help them overcome institutional and other hurdles that may impede the adoption of energy efficiency projects in complex investment environments. Different evaluation strategies are required for these two types of programs

To assess the effects of training programs on the market one must:

- **Establish the current state of the art and resulting energy efficiency for the market actions of interest.** For example, in the case of HVAC installation it is necessary to determine what the typical installation practices in the market are for establishing system sizing, matching coils to air handling systems and determining appropriate air flow before training is offered. This effort will provide an understanding of the need for training as well as the magnitude of the energy savings that could result from a program designed to improve practices. This can be done in a variety of ways. It is usually done by interviewing practitioners to discover the practices they are using. Delphi groups, focus groups and surveys are used to collect information. In some cases (as in the case of HVAC contractor training) this work may have already been done at the time the evaluation is undertaken. In other cases this may not be the case and it will need to be undertaken.
- **Estimate the effectiveness of the training program in changing the knowledge, skills and abilities of those exposed to training.** This is an empirical study designed to determine the effectiveness of the training program in changing knowledge, opinions and practices in the market. For example, in the case of an HVAC installation training program this might be done by observing installations that were done before and after training; or by classroom exercises and tests intended to test the knowledge of trainees before and after exposure to the training.
- **Estimate the average improvement in energy efficiency that results from providing training to the target market.** For example, in the case of the HVAC installation contractor training, this could be done by analyzing the difference in estimated energy efficiency of installations completed by each trainee before and after exposure to the treatment. This will produce an estimate of the average uplift in energy efficiency (e.g., annual kWh savings, SEER) that results from exposure to training.
- **Assess the persistence of the effect of the training.** It is possible that trainees will cease to use the practices they learn in training as time passes. Therefore, it is important to follow up with trainees after significant time has passed (i.e., 1-2 years) to determine how much the effect of the program is decaying. This may suggest the need for refresher courses or other actions to resent the effect of the program; or at a minimum an adjustment will have to be made in the long term expected savings resulting from the program.
- **Observe any spillover effects that may have occurred because of training.** It is possible (even likely) that useful practices learned directly in training will be transferred from trainees to other workers as time goes on. This should be expected because skilled workers often use first-hand experience to teach their colleagues useful practices. For example, in the case of the HVAC installer training, it might very well be the case that journeyman HVAC workers who receive the training will train the apprentices in their companies or even other apprentices in their trade working in different companies to apply the techniques they learn in the classroom.

A number of empirical measurements are required to address the above issues. Most of the measurements required to evaluate training programs involve surveys of trainees taken before and after exposure to training; survey measurements of parties who do not undergo training (i.e., control groups); and in some cases survey measurements of physical facilities (e.g., installed systems affected by the actions of trainees. In many cases it will be possible and highly desirable to carry out experiments in which the outcomes of market actions taken by those who have received training (e.g., installations) are compared with outcomes of market actions taken by those who have not received training.

Unlike the training initiatives described above the segment support programs are designed to improve energy efficiency by providing consulting expertise to specific organizations (e.g., municipalities, schools, hospitals, industries, etc.) to help them identify cost effective energy efficiency investments and implement them. The outcome measures of interest for these initiatives is not a better educated and more qualified workforce but an accelerated rate of adoption of energy efficient technology by specific organizations. In other words, the effect of the segment support programs is not to improve the knowledge of the organizations that are being served by EE specialists, it is to use the efforts of these specialists to overcome institutional barriers that impede adoption of more energy efficient technologies in organizations. This sort of program is particularly challenging to evaluate because very little about the implementation of the program can come under the control of the evaluator. That is, it is difficult to craft a true experimental design that can be practically implemented in the context of such a program.

To assess the impact of segment support programs one must:

- Identify the market segments that should be or are being targeted (e.g., municipal governments, state governments, universities and colleges, school systems, forest products, mining, mineral extraction, real estate, etc.) and the organizations inside those segments that have significant potential for energy efficiency improvements. The purpose of this task is to identify the potential targets of the program. This information is useful both in directing the work of the energy efficiency solutions providers and in assessing the extent which their efforts are being directed at high value targets for evaluation purposes.
- Estimate the effectiveness of the service delivery system in overcoming barriers to the identification and adoption of energy efficient technology. This is a very challenging problem. Energy savings potential will vary dramatically from sector to sector and within sector from organization to organization. Moreover, the service can only be delivered to organizations that volunteer to accept it and it is undoubtedly the case that organizations that volunteer are inherently more likely to identify and implement energy efficiency improvements than those that do not. Correspondingly, it will be very difficult to identify organizations to serve as control groups for purposes of identifying the effectiveness of the program. Probably the best way to establish control groups for the segment support programs is to divide up the service area geographically and make segment support available to some areas and not to others. In this way it would be possible to compare the rates at which organizations of different types are implementing energy efficiency improvements for the different geographical locations.

So for example, if there are 50 municipalities in one area and 50 in another, and segment support is only offered in one area and not in the other, it would be possible to compare the rates at which the municipalities in the different areas are implementing energy efficiency improvement plans as well as the resulting savings. Any effort to quantify the effectiveness in the absence of the establishment of such a control group will be subject to selection effects and therefore will produce a biased estimate of the effect of the program.

- Estimate the uplift in energy efficiency that results from providing assistance. Plans that are actually implemented will usually incorporate rebates or incentive payments and the calculations required to obtain these incentives can be used to estimate the resulting energy savings. It should be possible to assess the claimed savings resulting from the plans made by organizations and if necessary to verify the accuracy of those claims. The magnitude of the uplift must be judged in terms of the increase in energy savings over and above the savings that occur in locations where the segment support programs are not offered.

5.1 - Protocol 1: Define the Situation

The first step in research design is to develop a clear understanding of the purpose of the evaluation research and the context in which it is being carried out. In general, it is expected that the evaluator and project manager for the behavioral intervention will work collaboratively to answer the questions raised in this protocol. So, the application of this protocol is actually a task in which the parties who are carrying out and evaluating the training program work collaboratively to literally define the research design.

Describe the Capacity Building Program:

- **Type of Program** – Training or Segment Support
- The target population (i.e., in the case of training identify market actors that are targeted, in the case of segment support identify the specific market segments that are being targeted)
- The behavior(s) that is/are targeted for modification (e.g., design practices, system specification, building design, construction practices, installation, operations, organizational decisions, etc.)
- The mechanism(s) that is/are expected to change behavior (e.g. education, feedback, training, indoctrination, organizational change etc.)
- Whether presentation of the hypothesized behavioral change mechanism(s) is/are under the control of the evaluator (i.e., whether the evaluator can decide which members receive the behavior change mechanism and which do not)
- The outcomes that will be observed (i.e., adoption of technology, adoption of practices, sales of efficient technology, energy consumption, rebate requests, information system access attempts).

The answers to the above questions should be no more than a page in length each and should describe the behavioral program in sufficient detail to permit discussion of the experimental design alternatives with stakeholders.

While all of the above questions are important for identifying an appropriate research design for a behavioral outcome evaluation, none are more important than question no. 4 – i.e., whether the exposure to the behavior change mechanism can be brought under the evaluator's control. If the presentation of the treatment can be controlled, then it is possible to employ true experiments and reach definitive conclusions about the effectiveness of the behavioral mechanism at relatively low cost. If it is not possible to control the presentation of the treatment, then it will be necessary to evaluate the program using quasi-experimental techniques which are inherently less reliable than the true experiments and rest on assumptions that may or may not be tenable.

Exposure to the treatment may be outside the evaluator's control for a variety of reasons. For example, the program may have already been implemented or be underway when the evaluator is first introduced to the problem. So, the treatment may have already been presented to the target audience. It is also sometimes the case that regulators prescribe the delivery of treatments – requiring that all eligible parties receive a given behavioral treatment (e.g., access to training); and sometimes utility management are reluctant to deprive parties who are seeking access to behavioral programs – either because they do not want to disappoint them or because they want to achieve maximum effect of the behavioral intervention. These and other considerations may limit the control of the delivery of the experimental treatment of subjects in impact evaluations. The type of and robustness of the experimental design that can be implemented depend entirely on the extent of control the evaluator has over the assignment of subjects to treatments.

Program managers and other stakeholders often resist controlling the delivery of treatment to customers. They suspect or know that depriving customers of treatments they desire can create an unpleasant customer experience that may cause problems for them and their superiors. So it will often be necessary to educate these parties about the need for controlled experiments; and to convince them to accept the highest level of control possible. For this reason it is appropriate and necessary to plan to carry out the work required to implement Protocol 1 collaboratively with the project manager. The answer to the question that follows is critical to the eventual design of the evaluation and will in large measure govern the usefulness of the study results.

Table 5-1 identifies the level of control you believe is possible in assigning the treatment to subjects and why.

Provide a brief discussion of factors that led you to this conclusion.

This discussion should not exceed five pages and should carefully state your reasons for concluding that your level of control is as indicated in section 5.1.4. The purpose of this element of the protocol is to demonstrate that the evaluation team has carefully analyzed the design of the program in an effort to identify opportunities to create randomized experimental groups and has reached their decision on the level of control based on a good faith effort to attempt to achieve maximum control over the assignment of subjects to treatment and control groups and that you and your client understand the consequences of the level of control you have identified.

Table 5-1: Appropriate Experimental Designs Based on Ability to Control

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions	Randomized Controlled Trial (RCT)
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control 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 (e.g., income, usage, building size, etc.)	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	Quasi-experimental designs

5.2 - Protocol 2: Describe the Outcome Variables to be Observed

Among other things, Protocol 1 (Section 5.1.1) requires the evaluator to describe the behaviors that are to be modified by the intervention. Observations of two basic outcomes will be required – behavior changes and energy savings. Behaviors of interest will vary with the design of the intervention. For example, the training for HVAC contractors is designed to change several very specific behaviors carried out by sales and installation technicians – procedures used to estimate equipment size requirements, procedures used to select the size of coils, procedures used to establish air flow and several other activities. For other training programs the behaviors of interest may be different. For segment support programs offering EE solutions, the behaviors will be very different – including changes in the behavior of organizations such as adopting energy efficiency investment plan and operating plans and investments in recommended energy efficiency investments.

In Protocol 2, the evaluator is required to explicitly describe the measurements that will be used to observe the behaviors of interest before, during and after exposure to the intervention. There are two broad categories of measurements that arise in the context of evaluating behavioral interventions – observations of behavior or actions taken in response to interventions and observations of the impacts of the intervention on energy consumption.

Protocol 2 consists of a series of questions that are designed to produce an exhaustive list of outcomes that will be measured in the evaluation. As discussed earlier, this list may evolve iteratively if the initial evaluation design and the budget required to assess all of the treatments and outcomes of interest exceeds what is available, and therefore not everything of interest may be pursued.

In general, this protocol is designed to identify all of the different types of physical measurements that must be taken in order to assess the impacts of the behavioral intervention. These measurements might include:

- Measurements from tracking systems recording the progress of marketing efforts indicating who received program offers, what channels the offers were transmitted through, how many offers were sent, what content they received and if and when they responded to the offers.
- Records of participation in rebate and other programs that may identify actions taken by subjects in response to the program
- Measurements from surveys of consumers or other market actors taken before and after exposure to treatments.
- Measurements from tests given to trainees before and after exposure to training.
- Measurement of energy consumption before, during and after treatment for treatment and control groups

Please describe the behavioral outcomes of interest in the study, the operational definitions that will be used to measure them.

Complete Table 5-2 in as much detail as possible describing all of the behavioral and energy savings outcomes that are expected to occur as a result of the program along with operational definitions of each outcome.

Table 5-2: Table Caption

Behavioral Outcome	Operational Definition
Training Programs · e.g. HVAC Installation Contractor Training Program · Improved performance in carrying out best practices in calculating system size requirements and applying other technical and non-technical practices involved in installation.	**Behavior Measures** · Comparison of actual work before and after training or treated and control trainees, · written test of trainee knowledge before and after training, · comparison of knowledge and opinions (as measured by test) of trainees and comparison group
Training Programs · Energy savings resulting from improved performance from training	**Savings Measures** · Comparison of average SEER of systems installed by treatment and control groups before and after training · Estimated annual, monthly, hourly energy savings given average SEER difference · Estimated difference in peak kW if any by hour · Other energy consumption measurements
Segment Support Programs · e.g. EE solutions support to Municipal Governments	**Behavior Measures** · Rate of acceptance of assistance in treatment groups · Expressed interest in assistance for control groups · Comparison of rate of adoption of different types of energy efficiency solutions (e.g., energy efficiency plans, financial analysis, management presentations, measures adopted) for treatment and control groups
Segment Support Programs · Energy savings resulting from solutions	**Savings Measures** · Comparison of annual energy consumption for treatment and control organizations before and after treatment

5.3 - Protocol 3: Delineate Sub-segments of Interest

Capacity Building programs are sometimes targeted at multiple audiences (e.g., trades or disciplines in the case of training programs and market segments in the case of EE solutions segment support programs). If there is a desire to understand how the program affects different market segments, it is important to recognize these different segments during the design process. Protocol 3 requires the evaluator to identify all of the segments that are of interest in the study.

Complete the following table in as much detail as possible describing all of the segments that are of interest in the evaluation. Be careful to limit the segments to those that can be observed for both the treatment and control group before subjects are assigned to treatment groups. For example, it is possible to determine in advance of treatment whether a person working in a given HVAC contracting firm is a sales agent or an installer. This might be a useful segmentation variable, as there is some evidence that these two disciplines approach the installation of new equipment differently. It is also important to limit the number of segments so that 30-100 observations can be taken within each segment and treatment level.

Please describe all of the segments that are of interest in the study.

In Table 5-3, please use one line for each segment of interest.

Table 5-3: Segments of Interest

Segments of Interest
Training Programs (e.g., different jobs, different sized organizations, different business types, etc.)
Segment Support Programs (e.g., different types of organizations (municipal governments, school systems, state government departments), different industries (forest products, light manufacturing, etc.)

5.4- Protocol 4: Define the Research Design

Protocol 4 is designed to guide the experimental design process by asking evaluators to answer key questions designed to identify the theoretically correct design, as well as the practical realities that confront real-world social experimentation. When completing these questions, it may be useful to refer to Section 5 of this document as a guide to selecting the experimental design that best supports the treatments, objectives, and practical realities associated with the specific experiment under consideration.

Please answer the following questions.

Please use Table 5-4 to complete your answers.

Table 5-4: Questions on Behavior Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?		
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?		
How long of a pre-treatment period of data collection is required?		
Is a control group (or groups) required for the experiment?		
Is it possible to randomly assign observations to treatment and control groups?		

Using the framework outlined in Chapter 4, describe the evaluation research design that will be used during the evaluation.

This description should explain what type of research design will be used (e.g., RCT, RED, Regression Discontinuity, Non-Equivalent Control Groups, Within Subjects, etc.) It should describe the treatment groups and control groups and any segmentation (e.g., by trade or industry group) that is contemplated. In the case of true experiments, the design should be presented in a table of the kind presented in Section 5.2.2 where treatments are described on the column headings and segments are described on the rows. If random assignment is either inappropriate or impossible to achieve, the description should explicitly discuss how suitable comparison groups will be identified or how the design otherwise provides a comparison that allows an assessment of the impact of the treatment on behavior and energy consumption.

5.5 - Protocol 5: Define the Sampling Plan

Once the appropriate experimental design has been selected, a sample plan must be developed. Obviously, experimental design and sampling go hand in hand. While an in depth discussion of sample design would lead us far afield of the focus of research design, there are certain critical issues that have to be addressed in any sample design used to study the impacts of behavioral interventions. They are:

- Are the results of the research intended to be extrapolated beyond the experimental setting to a broader population (e.g., all parties involved in the installation of HVAC systems in the region served by IESO)?
- Are there sub-populations (strata) for which precise measurements are required (e.g., sales agents and installation technicians)?
- What is the absolute minimum level of change in the dependent variable(s) that is meaningful from a planning perspective (e.g., 1.5 SEER point improvement in performance of installed HVAC systems)?
- How much sampling error is permissible (e.g., + or - .1 SEER point)?
- How much statistical confidence is required for planning purposes (e.g., 90%)?
- Are pre-treatment data available concerning outcome variable(s) of interest?

The answers to the above questions will greatly influence the design of the samples to be used in the study. They cannot and should not be answered by the sampling statistician. The answers to these questions must be informed by the policy considerations. They have to be made by the people who will use the information to make decisions given the results. Once these requirements have been developed, a sampling expert can then determine the sample composition and sizes needed to meet the requirements.

Defining the Target Customer Population

Often it will not be necessary to extrapolate the results of the experiment to a larger population of interest. That is, it may not be necessary to generalize the results from a given experimental test of a training program to all possible parties who might be exposed to it. Instead, the purpose of the experiment may simply be to observe the effect of the treatment on the population of parties who were exposed to it. In this case it is not necessary to sample observations from the entire population of possible participants.

However, if the results of the experiment are to be statistically extrapolated to a larger population outside the experiment, then it is necessary to draw a representative (i.e., random) sample from the available population, and the sample has to be structured so that it is possible to calculate meaningful estimates of the population level impacts using appropriate sampling weights. To calculate weights for purposes of extrapolation, it is necessary to have a list of the members of the population of interest, to sample randomly from that list before assigning customers to treatment and control conditions, and to carefully observe any selection effects that might emerge in the sampling process so that the extrapolation can be adjusted to take account of them.

If precise measurements are needed for specific sub-populations (e.g., certain trades or organizations in different industries), then it may be necessary to over-sample these customers to ensure that enough observations are present in relevant cells to precisely estimate the impacts of the treatment. These are called sampling strata or blocks as described in Section 3.

Precision of the Estimates

A critical requirement in developing a sample design for any sort of experiment is a clear understanding of the minimum threshold of difference (between treated and not treated customers) that is considered meaningful from the point of view of those who will be using the results in 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 in order 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 sizes of samples required in an experiment is the quantity of sampling error that is tolerable from the point of view of planning.

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 isn't one (because of sampling variation). This is called a Type I error. Or one can incorrectly conclude that there isn't 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.

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 isn't 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 isn't 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 things 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 sure we want to be that we are not incorrectly finding a statistically significant difference, the less sure we can be that we 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 sample design. The statistical power used in calculating required sample sizes for experiments is subjective and, in modern times, 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 isn't a difference of a specified magnitude when indeed there is one. For Capacity Building experiments, statistical power 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.

Please answer the following questions pertaining to sample planning:

1. **Are the measurements from the experiment to be extrapolated to a broader population?**

a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.

b. If no, describe the list of parties from which the sampling will be obtained.

---

2. **Are precise measurements required for sub-populations of interest?**

a. If yes, describe the sub-populations for which precise measurements are desired.

---

3. **What is the minimum threshold of difference that must be detected by the experiment?**

---

4. **What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?**

---

5. **Will participants be randomly assigned to treatment and control conditions or varying levels of factors under study?**

a. If yes, do you expect subjects to select themselves into the treatment condition?

b. If so, how will you correct for this selection process in the analysis and sample weighting?

---

6. **If subjects will not be randomly assigned to treatment and control conditions or varying levels of factors under study:**

a. Describe the process that will be used to select customers for the treatment group(s).

b. 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.

---

7. **If no control group is used, explain how the change in the outcome variables of interest will be calculated.**

Please indicate the proposed sample sizes (within the treatment cells) for the study.

If experiments are contemplated (true or quasi-experiments) please use the table format provided in 4.2.2 to describe the distribution of sample across treatment cells and strata.

5.6- Protocol 6: Identify the Program Recruitment Strategy

Most capacity building programs will require outreach to the community of eligible participants to recruit them to participate in training or support programs. At a minimum, the evaluation must carefully describe the recruiting process used to attract program participants.

Please answer the following questions in Table 5-5 regarding the recruiting process and its outcome.

Table 5-5: Questions on Recruiting Process and Outcome

Question	Answer
Describe the eligibility criteria for the program	e.g., participants must be actively employed HVAC sales or installation technicians with more than 5 years of experience in the industry
What is the estimated number of eligible parties in the region under study	e.g., 10,000 total (sub-groups unknown)
How were participants recruited to the program	e.g., flyers were mailed to all currently licensed HVAC contractors in the region
Were participants randomly assigned to treatment and control conditions	e.g., yes, because of limited availability ½ of interested parties were randomly admitted into the program in the first year and the reminder was asked to wait for training until the following year
If there were sampling strata indicate the number of participants recruited into each strata and group	e.g. 100 sales technicians in treatment, 100 HVAC installers in treatment, 100 sales technicians in control and 100 HVAC technicians in control

It is sometimes the case that multiple recruiting processes are being tested during the evaluation program and that one of the objectives of the evaluation is to evaluate recruitment strategy alternatives and identify the most cost-effective approach for purposes of program design, taking into consideration both the number of enrollees as well as the average savings per customer.

If different recruiting strategies are being tested as part of the program please answer the following questions:

- Describe each of the recruiting options that are being tested in the program including how potential participants are being identified, how they are being contacted, what they being told, whether they are being offered incentives and any other pertinent information.
- Describe the research design that is being used to assess the effectiveness of alternative recruiting strategies including: the type of experimental design being employed (e.g., RCT, RED), how customers are sampled for the recruitment and how many potential participants are being selected for each recruiting test.
- Describe how the results of the recruiting strategy tests will be analyzed statistically.

5.7- Protocol 7: Identify the Length of the Study

In evaluating a behavioral intervention it is important to understand the expected time required to carry out the various aspects of the intervention, the expected onset time for the effect of the treatment and its expected persistence after initial treatment. These considerations will determine the length of time that is required to assess the impact of the treatment and thereby determine the length of time for which the situation must be observed.

Please answer the following questions pertaining to the experimental time frame.

1. Is it possible to observe the impacts of the treatment for at least two years?
2. If no, how will the persistence of the effect be determined?
3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?
4. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?
5. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?
6. What is the expected amount of time required for subjects to receive and understand the information being provided to them?
7. What is the expected amount of time needed by subjects to implement behavioral changes in response to the information provided?
8. What is the minimum amount of time the effect of the treatment must persist to cost-justify investment on the part of the utility?
9. If the duration of the experiment is shorter than the expected persistence of the treatment how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?
10. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?
11. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

5.8 - Protocol 8:
Identify Data Requirements and Collection Methods

Please complete the following table identifying the data requirements and data collection methods for each data element required in the evaluation. The table describes three types of data – energy consumption data, data describing the behaviors in question and other data.

Table 5-6 should be completed for as many measurements that will be taken during the course of the study. For example, if the SEER of an installed AC unit is to be collected as part of the evaluation then it should be described under energy consumption. The description of the variable should include a definition of the variable in sufficient detail as to permit third parties to understand what the measurement is. It should describe the frequency with which the measurement will be taken. For electricity consumption, the variable might be once or twice (as in the case of SEER measurements), or it might be monthly, hourly or even momentarily in the case of electricity consumption or demand. The method of measurement should describe how the data will be collected in as much detail as is required to explain the data collection process. 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, then the technology as well as installation and data collection protocols should be described.

Table 5-6: Measurements Taken During the Study

Energy Consumption	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Behaviors of Interest	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Other Data	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	

Behavior data is information describing the impact of the program on target behaviors. Examples of behavior data that might be appropriate for training programs include: classroom tests of knowledge, skills or abilities before and after training, observations of actions taken by trainees before and after training (e.g., installations or operating condition). Behavior data for segment support might include interviews with organization members concerning the impacts of the segment support program offerings on the operations of the target and control organization.

Other data includes all kinds of other data that might be useful in evaluating the impacts of the training or segment support programs including: 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 the behavioral intervention has taken place.

6. Protocols for Evaluating Feedback Programs

In recent years significant efforts have been undertaken to develop and test different information feedback strategies to cause customers to adjust their behavior related to energy consumption.

A wide variety of techniques have been developed or are under development including: normative comparisons designed to present consumers with a comparison of their household energy use with that of other households; in home display devices that are intended to inform consumers of their energy consumption in near real time; adaptive thermostats that are capable analyzing the energy use related habits of consumers and adapting household systems to those habits and so on.

In some cases these interventions have been shown to be effective. However, what works on one population doesn't necessarily work on another and variations in the technical design of in home devices makes it impossible to infer the performance of all devices from tests conducted on one of them. Therefore, there is the need to carry out robust testing on feedback techniques to determine whether they are effective and if so whether the impacts they produce are justified in light of the costs.

6.1 - Protocol 1: Define the Situation

The first step in research design is to develop a clear understanding of the purpose of the evaluation research and the context in which it is being carried out. In general, it is expected that the evaluator and project manager for the behavioral intervention will work collaboratively to answer the questions raised in this protocol. So, the application of this protocol is actually a task in which the parties who are carrying out and evaluating the feedback program work collaboratively to literally define the research design.

Describe the Feedback Program(s) to be tested:

- **Type of Program** – Type of feedback (e.g., neighbor comparison, IHD, HAN, etc.)
- **The target population (e.g. households or businesses** – if these target populations have specific characteristics that will narrow the population of interest down from all customers such as usage thresholds or SIC categories they should be described in detail)
- The behavior(s) that is/are targeted for modification (e.g., thermostat settings, use of lighting, time of use, website access, acceptance of home energy audits or other services, etc.)
- The mechanism(s) that is/are expected to change behavior (e.g. normative comparisons, cognitive dissonance, commitment, etc.)
- Whether presentation of the hypothesized behavioral change mechanism(s) is/are under the control of the evaluator (i.e., whether the evaluator can decide which members receive the behavior change mechanism and/or when)
- The outcomes that will be observed (i.e., acceptance of treatment, energy use related behaviors, purchasing behavior, energy consumption, timing of energy consumption).

The answers to the above questions should be no more than a page in length each and should describe the behavioral program in sufficient detail to permit discussion of the experimental design alternatives with stakeholders.

While all of the above questions are important for identifying an appropriate research design for a behavioral outcome evaluation, none are more important than question no. 4 – i.e., whether the exposure to the behavior change mechanism can be brought under the evaluator's control. If the presentation of the treatment can be controlled, then it is possible to employ true experiments and reach definitive conclusions about the effectiveness of the behavioral mechanism at relatively low cost. If it is not possible to control the presentation of the treatment, then it will be necessary to evaluate the program using quasi-experimental techniques which are inherently less reliable than the true experiments and rest on assumptions that may or may not be tenable.

Exposure to the treatment may be outside the evaluator's control for a variety of reasons. For example, increasingly feedback devices such as IHDs, HAN systems, and Optimizing Thermostats are being sold over the counter and through the internet directly to consumers. It is impossible to control who obtains such devices and therefore impossible to randomly assign customers to treatment or control groups. It might be possible to randomly assign encouragement to customers, but that would be difficult to orchestrate. It is also sometimes the case that regulators prescribe the delivery of treatments – requiring that all eligible parties receive a given behavioral treatment (e.g., access to website information concerning energy consumption and energy saving tips); and sometimes utility management are reluctant to deprive parties who are seeking access to behavioral programs – either because they do not want to disappoint them or because they want to achieve maximum effect of the behavioral intervention. These and other considerations may limit the control of the delivery of the experimental treatment of subjects in impact evaluations. The type of and robustness of the experimental design that can be implemented depend entirely on the extent of control the evaluator has over the assignment of subjects to treatments.

Program managers and other stakeholders often resist controlling the delivery of treatment to customers. They suspect or know that depriving customers of treatments they desire can create an unpleasant customer experience that may cause problems for them and their superiors. So it will often be necessary to educate these parties about the need for controlled experiments; and to convince them to accept the highest level of control possible. For this reason it is appropriate and necessary to plan to carry out the work required to implement Protocol 1 collaboratively with the project manager. The answer to the question that follows is critical to the eventual design of the evaluation and will in large measure govern the usefulness of the study results.

In Table 6-1, identify the level of control you believe is possible in assigning the treatment to subjects and why.

Table 6-1: Table Caption

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions	Randomized Controlled Trial (RCT)
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control 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 (e.g., income, usage, building size, etc.)	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	Quasi-experimental designs

Provide a brief discussion of factors that led you to this conclusion.

This discussion should not exceed five pages and should carefully state your reasons for concluding that your level of control is as indicated in section 6.1.4. The purpose of this element of the protocol is to demonstrate that the evaluation team has carefully analyzed the design of the program in an effort to identify opportunities to create randomized experimental groups and has reached their decision on the level of control based on a good faith effort to attempt to achieve maximum control over the assignment of subjects to treatment and control groups and that you and your client understand the consequences of the level of control you have identified.

6.2 Protocol 2: Describe the Outcome Variables to be Observed

Among other things, Protocol 1 (Section 6.1) requires the evaluator to describe the behaviors that are to be modified by the intervention. Observations of several basic outcomes will be required. These include:

- The acceptance rate of feedback;
- Changes in appliance acquisition behavior;
- Changes in energy use related behavior; and
- Changes in other behaviors (e.g., knowledge, opinions and attitudes).

Specific behaviors of interest will vary with the design of the intervention. For example, some feedback techniques are provided to all customers by default. This is virtually always the case with written normative comparisons. In other cases, customers may be offered feedback technology a zero cost or reduced cost and make the decision whether or not to accept it. These two very different deployment strategies require the collection of very different outcome measures for measuring customer acceptance.

In Protocol 2, the evaluator is required to explicitly describe the measurements that will be used to observe the behaviors of interest before, during and after exposure to the intervention. Protocol 2 consists of a series of questions that are designed to produce an exhaustive list of outcomes that will be measured in the evaluation. As discussed earlier, this list may evolve iteratively if the initial evaluation design and the budget required to assess all of the treatments and outcomes of interest exceeds what is available, and therefore not everything of interest may be pursued.

In general, this protocol is designed to identify all of the different types of physical measurements that must be taken in order to assess the impacts of the behavioral intervention. These measurements might include:

- Measurements from tracking systems recording the progress of marketing efforts indicating who received program offers, what channels the offers were transmitted through, how many offers were sent, what content they received and if and when they responded to the offers.
- Records of participation in rebate and other programs that may identify actions taken by subjects in response to the program
- When enabling devices 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 treatments.
- Measurements of drop-out rates and reasons for departing the program.
- Measurement of energy consumption before, during and after treatment for treatment and control groups

Please describe the behavioral outcomes of interest in the study, the operational definitions that will be used to measure them.

Complete Table 6-2 in as much detail as possible describing all of the behavioral and energy savings outcomes that are expected to occur as a result of the program along with operational definitions of each outcome. The table shows an example of the level of detail that is required for feedback experiments involving Normative Comparisons and Feedback.

Table 6-2: Behavioral Outcome and Operational Definition

Behavioral Outcome	Operational Definition
Normative Comparisons · Customer acceptance · Energy related knowledge, skill and opinions · Appliance acquisition behaviors · Energy use related behavior.	**Behavior Measures** • Customer subscription rate (for opt-in delivery) and opt-out rate (for default delivery) from tracking system • Surveys of treatment and control customers' knowledge, skills and opinions, reported appliance acquisition behavior and reported energy use related behavior before and after treatment
Normative Comparisons · Energy savings resulting from providing normative comparisons	**Savings Measures** · Observed differences in monthly or annual energy consumption and demand (kWh, therms) for treatment and control groups before and after treatment from billing systems
Other Feedback Strategies (i.e., IHD, HAN Optimizing Thermostats) · Customer acceptance · Device commissioning · Device utilization · Energy related knowledge, skill and opinions · Appliance acquisition behaviors · Energy use related behavior · Usability · Persistence	**Behavior Measures** · Customer acceptance rate from tracking system · Device commissioning rate from MDMS or other tracking system · Interviews/focus groups with customer service agents · Interviews with customers regarding commissioning problems · Surveys of treatment customers regarding satisfaction with acquisition/installation process · Surveys of treatment customers and control customers' knowledge, skills and opinions, reported appliance acquisition behavior and reported energy use related behavior before and after treatment · Focus groups with treatment customers regarding usability and persistence
Other Feedback Strategies (i.e., IHD, HAN Optimizing Thermostats) · Energy savings resulting from providing technology	**Savings Measures** · Observed differences in monthly or annual energy consumption and demand (kWh, therms) for treatment and control groups before and after treatment from billing systems
Website · Customer acceptance · Website access · Website utilization · Opinions about website · Energy related knowledge, skill and opinions · Energy use related behavior · Usability · Persistence	**Behavior Measures** · Website access from tracking system · Page views from tracking system · Return rate from tracking system · Focus groups with customers regarding usability · Surveys of treatment customers regarding satisfaction with website content and performance · Surveys of treatment customers and control customers' knowledge, skills and opinions, reported appliance acquisition behavior and reported energy use related behavior before and after treatment

6.3 - Protocol 3: Delineate Sub-segments of Interest

Feedback programs are sometimes targeted at multiple audiences (e.g., customers on time varying rates, disadvantaged customers, customers with certain heating or cooling devices, etc.). If there is a desire to understand how the program affects different market segments, it is important to recognize these different segments during the design process. Protocol 3 requires the evaluator to identify all of the segments that are of interest in the study.

Complete the following table in as much detail as possible describing all of the segments that are of interest in the evaluation. Be careful to limit the segments to those that can be observed for both the treatment and control group before subjects are assigned to treatment groups. For example, it is possible to determine in advance of treatment whether a household is on a rate that qualifies for a discount or if it is on time varying rates. It is not possible to determine the approximate annual income of the household. The former are good candidates for stratification, while the later are is not. It is also important to limit the number of segments so that 30-100 observations can be taken within each segment and treatment level.

Please describe all of the segments that are of interest in the study.

Please use one line for each segment of interest in Table 6-3.

Table 6-3: Segments of Interest

Segments of Interest
IHD, HAN, Optimizing Thermostats, (e.g., rates, usage categories, assisted customers, etc.)
Website (e.g., Current MyAccount customers, engaged customers, behavioral segments etc.)

6.4 - Protocol 4: Define the Research Design

Protocol 4 is designed to guide the experimental design process by asking evaluators to answer key questions designed to identify the theoretically correct design, as well as the practical realities that confront real-world social experimentation. When completing these questions, it may be useful to refer to Section 4 of this document as a guide to selecting the experimental design that best supports the treatments, objectives, and practical realities associated with the specific experiment under consideration.

Please answer the following questions.

Please use Table 6-4 to complete your answers.

Table 6-4: Questions on Behavior and Energy Consumption Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?		
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?		
How long of a pre-treatment period of data collection is required?		
Is a control group (or groups) required for the experiment?		
Is it possible to randomly assign observations to treatment and control groups?		

Using the framework outlined in Chapter 4, describe the evaluation research design that will be used during the evaluation.

This description should explain what type of research design will be used (e.g., RCT, RED, Regression Discontinuity, Non-Equivalent Control Groups, Within Subjects, etc.) It should describe the treatment groups and control groups and any segmentation (e.g., customer type, usage category, etc.) that is contemplated. In the case of true experiments, the design should be presented in a table of the kind presented in Section 4.2.2 where treatments are described on the column headings and segments are described on the rows. If random assignment is either inappropriate or impossible to achieve, the description should explicitly discuss how suitable comparison groups will be identified or how the design otherwise provides a comparison that allows an assessment of the impact of the treatment on behavior and energy consumption.

6.5 - Protocol 5: Define the Sampling Plan

Once the appropriate experimental design has been selected, a sample plan must be developed. Obviously, experimental design and sampling go hand in hand. While an in depth discussion of sample design would lead us far afield of the focus of research design, there are certain critical issues that have to be addressed in any sample design used to study the impacts of behavioral interventions. They are:

- Are the results of the research intended to be extrapolated beyond the experimental setting to a broader population (e.g., all households eligible to receive the technology in the region served by IESO)?
- Are there sub-populations (strata) for which precise measurements are required (e.g., usage categories or other segments)?
- What is the absolute minimum level of change in the dependent variable(s) that is meaningful from a planning perspective (e.g., 5% reduction in electricity or gas consumption)?
- How much sampling error is permissible (e.g., + or – 1%)?
- How much statistical confidence is required for planning purposes (e.g., 90%)?
- Are pre-treatment data available concerning outcome variable(s) of interest?

The answers to the above questions will greatly influence the design of the samples to be used in the study. They cannot and should not be answered by the sampling statistician. The answers to these questions must be informed by the policy considerations. They have to be made by the people who will use the information to make decisions given the results. Once these requirements have been developed, a sampling expert can then determine the sample composition and sizes needed to meet the requirements.

Defining the Target Customer Population

Often it will not be necessary to extrapolate the results of the experiment to a larger population of interest. That is, it may not be necessary to generalize the results from a given experimental test of a technology to all possible parties who might be exposed to it. With large scale feedback technologies targeted at the general market, extrapolation is an important consideration. However, in testing emerging technologies like IHDs, HAN devices and Websites, thoughts about extrapolation are futile. Virtually anyone who agrees to participate in a test of a new technology is an early adopter and there is no reason to believe that impacts of technology on this market segment foretell how the technology will be taken up in the general market. So, it is possible that in many cases the purpose of the experiment will simply be to observe the effect of the treatment on the population of parties who were exposed to it. In this case it is not necessary to sample observations from the entire population of possible participants.

However, if the results of the experiment are to be statistically extrapolated to a larger population outside the experiment, then it is necessary to draw a representative (i.e., random) sample from the available population, and the sample has to be structured so that it is possible to calculate meaningful estimates of the population level impacts using appropriate sampling weights. To calculate weights for purposes of extrapolation, it is necessary to have a list of the members of the population of interest, to sample randomly from that list before assigning customers to treatment and control conditions, and to carefully observe any selection effects that might emerge in the sampling process so that the extrapolation can be adjusted to take account of them.

If precise measurements are needed for specific sub-populations (e.g., customer types or size categories), then it may be necessary to over-sample these customers to ensure that enough observations are present in relevant cells to precisely estimate the impacts of the treatment. These are called sampling strata or blocks as described in Section 3.

Precision of the Estimates

A critical requirement in developing a sample design for any sort of experiment is a clear understanding of the minimum threshold of difference (between treated and not treated customers) that is considered meaningful from the point of view of those who will be using the results in 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 in order 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 sizes of samples required in an experiment is the quantity of sampling error that is tolerable from the point of view of planning.

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 isn't one (because of sampling variation). This is called a Type I error. Or one can incorrectly conclude that there isn't 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.

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 isn't 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 isn't 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 things 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 sure we want to be that we are not incorrectly finding a statistically significant difference, the less sure we can be that we 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 sample design. The statistical power used in calculating required sample sizes for experiments is subjective and, in modern times, 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 isn't a difference of a specified magnitude when indeed there is one. For Capacity Building experiments, statistical power 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.

Please answer the following questions pertaining to sample planning:

1. **Are the measurements from the experiment to be extrapolated to a broader population?**

a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.

b. If no, describe the list of parties from which the sampling will be obtained.

---

2. **Are precise measurements required for sub-populations of interest?**

a. If yes, describe the sub-populations for which precise measurements are desired.

---

3. **What is the minimum threshold of difference that must be detected by the experiment?**

---

4. **What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?**

---

5. **Will participants be randomly assigned to treatment and control conditions or varying levels of factors under study?**

a. If yes, do you expect subjects to select themselves into the treatment condition?

b. If so, how will you correct for this selection process in the analysis and sample weighting?

---

6. **If subjects will not be randomly assigned to treatment and control conditions or varying levels of factors under study:**

a. Describe the process that will be used to select customers for the treatment group(s).

b. 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.

---

7. **If no control group is used, explain how the change in the outcome variables of interest will be calculated.**

Please indicate the proposed sample sizes (within the treatment cells) for the study.

If experiments are contemplated (true or quasi-experiments) please use the table format provided in 4.2.2 to describe the distribution of sample across treatment cells and strata.

6.6 - Protocol 6: Identify the Program Recruitment Strategy

Sometimes feedback programs are operated on an opt-in basis. That is, the treatment is given only to volunteers. When this is true, the recruitment strategy can affect the outcome of the evaluation. At a minimum, the evaluation must carefully describe the recruiting process used to attract program participants.

Please answer the following questions in Table 6-5 regarding the recruiting process and its outcome.

Table 6-5: Recruiting Process Questions

Question	Answer
Describe the eligibility criteria for the program	e.g., households in single family dwellings located in climate zones X and Y
What is the estimated number of eligible parties in the region under study	e.g., 1 million
How were participants recruited to the program	e.g., flyers were mailed to all currently eligible households
Were participants randomly assigned to treatment and control conditions	e.g., yes, because of limited availability ½ of interested parties were randomly admitted into the program in the first year and the reminder was asked to wait for training until the following year
If there were sampling strata indicate the number of participants recruited into each strata and group	e.g. 500 customers were sampled in each of 4 sampling strata

It is sometimes the case that multiple recruiting processes are being tested during the evaluation program and that one of the objectives of the evaluation is to evaluate recruitment strategy alternatives and identify the most cost-effective approach for purposes of program design, taking into consideration both the number of enrollees as well as the average savings per customer.

If different recruiting strategies are being tested as part of the program please answer the following questions:

- Describe each of the recruiting options that are being tested in the program including how potential participants are being identified, how they are being contacted, what they being told, whether they are being offered incentives and any other pertinent information.
- Describe the research design that is being used to assess the effectiveness of alternative recruiting strategies including: the type of experimental design being employed (e.g., RCT, RED), how customers are sampled for the recruitment and how many potential participants are being selected for each recruiting test.
- Describe how the results of the recruiting strategy tests will be analyzed statistically.

6.7 - Protocol 7: Identify the Length of the Study

In evaluating a behavioral intervention it is important to understand the expected time required to carry out the various aspects of the intervention, the expected onset time for the effect of the treatment and its expected persistence after initial treatment. These considerations will determine the length of time that is required to assess the impact of the treatment and thereby determine the length of time for which the situation must be observed.

Please answer the following questions pertaining to the experimental time frame.

1. Is it possible to observe the impacts of the treatment for at least two years?
2. If no, how will the persistence of the effect be determined?
3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?
4. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?
5. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?
6. What is the expected amount of time required for subjects to receive and understand the information being provided to them?
7. What is the expected amount of time needed by subjects to implement behavioral changes in response to the information provided?
8. What is the minimum amount of time the effect of the treatment must persist to cost-justify investment on the part of the utility?
9. If the duration of the experiment is shorter than the expected persistence of the treatment how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?
10. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?
11. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

6.8 - Protocol 8: Identify Data Requirements and Collection Methods

Please complete Table 6-6 identifying the data requirements and data collection methods for each data element required in the evaluation. The table describes three types of data – energy consumption data, data describing the behaviors in question and other data.

Table 6-6 should be completed for as many measurements that will be taken during the course of the study. For example, if electric and gas consumption are to be collected as part of the evaluation then they should be described in separate entries under energy consumption. The description of the variable should include a definition of the variable in sufficient detail as to permit third parties to understand what the measurement is. It should describe the frequency with which the measurement will be taken. For electricity consumption, the variable might be monthly, hourly or even momentarily in the case of electricity consumption or demand. The method of measurement should describe how the data will be collected in as much detail as is required to explain the data collection process. 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, then the technology as well as installation and data collection protocols should be described.

Table 6-6: Data Requirements

Energy Consumption	Description of Variable
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Behaviors of Interest	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Other Data	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	

Behavior data is information describing the impact of the program on target behaviors. Examples of behavior data that might be appropriate for feedback programs might include: reported recent history of appliance purchases, an inventory of energy saving actions taken since the start of the behavioral intervention, perceptions and opinions about energy use, reported conversations among the family or with neighbors about energy consumption, etc..

Other data includes all kinds of other data that might be useful in evaluating the impacts of the feedback programs including: 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 the behavioral intervention has taken place.

7. Protocols for Evaluating Education/Awareness Campaigns

Education and awareness campaigns are designed to change behavior or facilitate change in behavior by providing information to consumers.

Such campaigns assume that consumers are reasoning beings who use information about the consequences of their actions to formulate and undertake actions (behaviors) to achieve desired outcomes. There are very well developed social science theories expressing the causal relationship between perception, belief, intention and action. That is, there are well developed theories about how opinions are shaped and how opinions shape behavior. These theories -- generally referred to under the heading of Reasoned Action Theories -- hold that it is possible to educate people about the consequences of their actions, make them aware of the extent to which their actions are normatively acceptable and encourage them to formulate intentions to behave in a manner that is more in line with positive consequences and more normatively acceptable. Through this causal chain, consumers and other actors in the energy market are expected to change their behavior. Of course, the underlying social science theories can be much more complicated than this, but in broad outline terms, they all share these basic tenants.

Education and awareness campaigns have been in existence in the energy policy arena for at least four decades. Indeed, the first efforts to systematically change energy use related behavior were primarily education campaigns. These early efforts focused on informing consumers of the availability of energy efficient technology alternatives, of the economic benefits of energy efficiency and conservation, of the societal consequences of energy consumption and so on. They were carried out by government and utilities under the assumption that once consumers knew the facts they would behave appropriately.

Education and awareness campaigns can have a wide variety of goals. They can be designed to cause widespread changes in energy consumption. For example, in 2001 in California serious power shortages created the need for dramatic reductions in electricity consumption on the part of businesses and households. During that period, the California state government, in partnership with utilities and local governments implemented a wide spectrum public education and awareness campaign designed to encourage consumers to lower their energy consumption overall – and in particular on hot summer days. This campaign consisted of newspaper, television and radio advertising, bill inserts and other specialized marketing collateral designed to explain the seriousness of the situation, inform consumers of the offer to reduce electric bills by 20% for consumers who lowered their consumption (year on year) by 20%, and provide them with tips about how to reduce their energy use.

This Flex Your Power campaign was relatively large involving about $45 million in paid and earned advertising over a two year period. However, there are many examples of more modest efforts designed to accomplish less ambitious goals. For example, in California small and medium sized commercial and industrial firms are being defaulted to time of use rates between November of 2012 and November of 2014. An intensive education/awareness campaign is being used to inform customers when they will be defaulted and of the actions they can take to lower their costs either by reducing their energy consumption overall or by restricting their use during the peak hours in the afternoon. This is a relatively small and focused education effort that each year involves educating about 150,000 customers, costing only a few million dollars each year.

Education and awareness campaigns can be targeted at all levels of society. They can be national campaigns such as DOE's Energy Star Program, campaigns carried out by state and local governments as described above, campaigns focused on individual organizations or businesses – even campaigns focused on schools and neighborhoods.

One can imagine a very large number of examples of education and awareness campaigns with differing goals, messages, target audiences and contact strategies. However, the critical evaluation questions that must be answered for virtually all of these campaigns are the same. Namely,

- What were the beliefs, opinions, attitudes, intentions and behaviors of the target audience prior to exposure to the education or awareness campaign;
- What were the beliefs, opinions, attitudes, intentions and behaviors of the target audience after exposure to the education or awareness campaign; and most importantly
- Did the education campaign cause any observable change in the beliefs, opinions, attitudes, intentions and behaviors?

Beyond these basic questions it is possible to address a number of other interesting and important questions in the context of evaluating an education or awareness campaign. These include:

- What combinations of message, format and channel were most effective in educating or informing important market segments?
- Did the education campaign have an impact on targeted customers' belief that their behavior was normatively acceptable?
- Did exposure to the education campaign increase the likelihood that consumers expressed the intention to engaged in desired energy use related behavior?
- Did exposure to the education campaign increase the likelihood that consumers engaged in desired energy use related behavior?

While the ultimate objective of education and awareness campaigns may be to cause a change in energy consumption on the part of the target population by providing education, it is very difficult to conclusively demonstrate a causal connection between attitude change and behavior change. The causal linkage between education and action is mitigated through a number of important intervening factors that can significantly interfere with the expression of desired energy use related behavior. For example, it is possible that a target consumer receives the intended education and that the education has the desired effect of causing the consumer to intend to exhibit an energy conserving behavior, but that the consumer is prevented from doing so by circumstances in the market (e.g., lack of resources or control of the situation). For this reason, it may be difficult or impossible to directly quantify the impact of behavior change achieved in this manner on energy consumption.

7.1 Protocol 1: Define the Situation

The first step in research design is to develop a clear understanding of the purpose of the evaluation research and the context in which it is being carried out. In general, it is expected that the evaluator and project manager for the behavioral intervention will work collaboratively to answer the questions raised in this protocol. So, the application of this protocol is actually a task in which the parties who are carrying out and evaluating the feedback program work collaboratively to literally define the research design.

Describe the Education or Awareness Program(s) to be tested:

- The underlying behavioral science theory linking the information that is to be transmitted to the outcome behavior of interest (e.g., Theory of Reasoned Action diagram describing beliefs that are to be changed, social reinforcements that are to be given (if any), intentions that are to be affected if any and outcome behaviors of interest.)
- The target population(s) (e.g. household heads, children, business leaders, employees, etc.) – if there is a geographic catchment within which education or awareness is to be achieved it should be specified (i.e., city, state, nation, business, neighborhood, etc.)
- The information that is to be imparted to the target population (e.g., impacts of energy use on climate, cost of wasting energy, options for reducing energy consumption while maintaining comfort, benefits of changing timing of demand etc.)
- The behavior(s) that is/are targeted for modification (e.g., thermostat settings, use of lighting, time of use, website access, acceptance of home energy audits or other services, etc.)
- Whether presentation of the educational material is under the control of the evaluator (i.e., whether the evaluator can decide who receives the educational material and/or when)
- The outcomes that will be observed (e.g. awareness of messages, acceptance of messages, belief about normative support for action, expressed intention to engage in desired behavior, change in energy use, etc.).

The answers to the above questions should be no more than a page in length each and should describe the behavioral program in sufficient detail to permit discussion of the experimental design alternatives with stakeholders.

While all of the above questions are important for identifying an appropriate research design for a behavioral outcome evaluation, none are more important than question no. 4 – i.e., whether the exposure to the behavior change mechanism can be brought under the evaluator's control. If the presentation of the educational treatment can be controlled, then it is possible to employ true experiments and reach definitive conclusions about the effectiveness of the behavioral mechanism at relatively low cost. If it is not possible to control the presentation of the treatment, then it will be necessary to evaluate the program using quasi-experimental techniques which are inherently less reliable than the true experiments and rest on assumptions that may or may not be tenable.

The challenge in evaluating the effects of wide spectrum educational campaigns is that such campaigns are often carried out within media markets and it is impossible to restrict educational messages to customers within markets. However, Ontario is served by about 13 media markets so conducting educational campaigns in different randomly chosen media markets could provide a powerful platform for testing the impacts of education campaigns.

Exposure to the treatment may sometimes fall outside the evaluator's control. For example, it is often the case that education or awareness campaigns are carried out in emergencies or are required by law or good administrative practice. It may not be appropriate to randomly withhold advance notice to customers in emergencies or to those that will experience a rate change that might cause them to experience high bills that could have been avoided with advanced notice. Such situations will challenge the research designers and project managers 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 subjects to treatments.

Program managers and other stakeholders often resist controlling the delivery of treatment to customers. They suspect or know that depriving customers of education could create an unpleasant customer experience that may cause problems for them and their superiors in the future. So it will often be necessary to educate these parties about the need for controlled experiments; and to convince them to accept the highest level of control possible. For this reason it is appropriate and necessary to plan to carry out the work required to implement Protocol 1 collaboratively with the project manager. The answer to the question that follows is critical to the eventual design of the evaluation and will in large measure govern the usefulness of the study results.

In Table 7-1, identify the level of control you believe is possible in assigning the treatment to subjects and why.

Table 7-1: Identify Level of Control

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions	Randomized Controlled Trial (RCT)
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control 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 (e.g., income, usage, building size, etc.)	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	Quasi-experimental designs

Provide a brief discussion of factors that led you to this conclusion.

This discussion should not exceed five pages and should carefully state your reasons for concluding that your level of control is as indicated in section 7.1.4. The purpose of this element of the protocol is to demonstrate that the evaluation team has carefully analyzed the design of the program in an effort to identify opportunities to create randomized experimental groups and has reached their decision on the level of control based on a good faith effort to attempt to achieve maximum control over the assignment of subjects to treatment and control groups and that you and your client understand the consequences of the level of control you have identified.

7.2 - Protocol 2: Describe the Outcome Variables to be Observed

Among other things, Protocol 1 (Section 8.1) requires the evaluator to describe the behaviors that are to be modified by the intervention. Observations of several basic outcomes will be required. These include:

- Beliefs and opinions related to energy consumption;
- Beliefs about what is normatively appropriate energy use related behavior;
- Beliefs about whether their energy use related behavior is normatively appropriate;
- Perceptions of energy use related behaviors of others;
- Attitudes about energy consumption, comfort, convenience, etc.;
- Awareness of the education and awareness messages;
- Awareness of channels through which messages were transmitted;
- Reported energy use related behaviors
- Household/business energy use.

Specific behaviors of interest will vary with the design of the intervention. For example, interventions that are created in response to emergency conditions may focus on changing perceptions of the emergency conditions (e.g. drought, supply disruptions) and appropriate behaviors while other interventions may focus on perceptions of longer range issues such as climate change or reliability.

In Protocol 2, the evaluator is required to explicitly describe the measurements that will be used to observe the behaviors of interest before, during and after exposure to the intervention. Protocol 2 consists of a series of questions that are designed to produce an exhaustive list of outcomes that will be measured in the evaluation. As discussed earlier, this list may evolve iteratively if the initial evaluation design and the budget required to assess all of the treatments and outcomes of interest exceeds what is available, and therefore not everything of interest may be pursued.

In general, this protocol is designed to identify all of the different types of physical measurements that must be taken in order to assess the impacts of the behavioral intervention. These measurements might include:

- Measurements from surveys of consumers or other market actors taken before and after exposure to education campaigns;
- Measurements from tracking systems recording the details of the education campaign including when populations were exposed to education materials, what channels the messages were transmitted through, how many messages were sent and what content was used;
- Records of response to programs (if appropriate);
- Measurement of energy consumption before, during and after treatment for treatment and control groups

Please describe the behavioral outcomes of interest in the study, the operational definitions that will be used to measure them.

Complete Table 7-2 in as much detail as possible describing all of the behavioral and energy savings outcomes that are expected to occur as a result of the program along with operational definitions of each outcome. The table shows an example of the level of detail that is required for feedback experiments involving Normative Comparisons and Feedback.

Table 7-2: Behavioral Outcome and Operational Definition

Behavioral Outcome	Operational Definition
Beliefs About Own Energy Consumption · Beliefs and opinions related to energy consumption; · Attitudes about energy consumption, comfort, convenience, etc.; · Beliefs about whether subject's energy use related behavior is socially normal; · Awareness of the education and other r elated messages; · Awareness of channels through which messages were transmitted;	**Behavior Measures** · Surveys questions about beliefs held by subjects about their energy use before and after exposure to the educational treatment for treatment and control customers
Beliefs about Normative Energy Consumption · Beliefs about what is normatively appropriate energy use related behavior; · Perceptions of energy use related behaviors of others;	**Behavior Measures** · Surveys questions about beliefs held by subjects about what energy use related behavior and opinions are normatively correct before and after exposure to the educational treatment for treatment and control customers
Reported Energy Use Related Behavior · Reported intention to take actions to reduce energy consumption · Reported appliance purchases · Reported thermostat settings · Reported use of lighting and other appliances	**Behavior Measures** · Surveys questions about reported energy use related behaviors before and after exposure to the educational treatment for treatment and control customers
Energy Use · Energy savings resulting from providing technology	**Savings Measures** · Observed differences in monthly or annual energy consumption and demand (kWh, therms) for treatment and control groups before and after treatment from billing systems

7.3 - Protocol 3: Delineate Sub-segments of Interest

Education/Awareness programs are sometimes targeted at multiple audiences (e.g., customers on time varying rates, disadvantaged customers, customers with certain heating or cooling devices, etc.). If there is a desire to understand how the program affects different market segments, it is important to recognize these different segments during the design process. Protocol 3 requires the evaluator to identify all of the segments that are of interest in the study.

Complete the following table in as much detail as possible describing all of the segments that are of interest in the evaluation. Be careful to limit the segments to those that can be observed for both the treatment and control group before subjects are assigned to treatment groups. For example, it is possible to determine in advance of treatment whether a household is on a rate that qualifies for a discount or if it is on time varying rates. It is not possible to determine the approximate annual income of a household. The former are good candidates for pre-stratification, while the later are not. It is also important to limit the number of segments so that at least a few hundred observations can be taken within each segment and treatment level.

Please describe all of the segments that are of interest in the study.

In Table 7-3, please use one line for each segment of interest.

Table 7-3: Segments of Interest

Segments of Interest

7.4 - Protocol 4: Define the Research Design

Protocol 4 is designed to guide the experimental design process by asking evaluators to answer key questions designed to identify the theoretically correct design, as well as the practical realities that confront real-world social experimentation. When completing these questions, it may be useful to refer to Section 4 of this document as a guide to selecting the experimental design that best supports the treatments, objectives, and practical realities associated with the specific experiment under consideration.

Please answer the following questions.

Please use Table 7-4 to complete your answers.

Table 7-4: Behavior and Energy Consumption Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?		
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?		
How long of a pre-treatment period of data collection is required?		
Is a control group (or groups) required for the experiment?		
Is it possible to randomly assign observations to treatment and control groups?		

Using the framework outlined in Chapter 4 describe the evaluation research design that will be used during the evaluation.

This description should explain what type of research design will be used (e.g., RCT, RED, Regression Discontinuity, Non-Equivalent Control Groups, Within Subjects, etc.) It should describe the treatment groups and control groups and any segmentation (e.g., customer type, usage category, etc.) that is contemplated. In the case of true experiments, the design should be presented in a table of the kind presented in Section 5.2.2 where treatments are described on the column headings and segments are described on the rows. If random assignment is either inappropriate or impossible to achieve, the description should explicitly discuss how suitable comparison groups will be identified or how the design otherwise provides a comparison that allows an assessment of the impact of the treatment on behavior and energy consumption.

7.5 - Protocol 5: Define the Sampling Plan

Once the appropriate experimental design has been selected, a sample plan must be developed. Obviously, experimental design and sampling go hand in hand. While an in depth discussion of sample design would lead us far afield of the focus of research design, there are certain critical issues that have to be addressed in any sample design used to study the impacts of behavioral interventions. They are:

- Are the results of the research intended to be extrapolated beyond the experimental setting to a broader population (e.g., all households eligible to receive the education in the region served by IESO)?
- Will measurements of behavior change involving surveying be taken for only a subset of treatment and control customers?
- Are there sub-populations (strata) for which precise measurements are required (e.g., usage categories or other segments)?
- What is the absolute minimum level of change in the dependent variable(s) that is meaningful from a planning perspective (e.g., 5% increase in expressed positive opinions related to saving energy)?
- How much sampling error is permissible (e.g., + or – 1%)?
- How much statistical confidence is required for planning purposes (e.g., 90%)?
- Are pre-treatment data available concerning outcome variable(s) of interest?

The answers to the above questions will greatly influence the design of the samples to be used in the study. They cannot and should not be answered by the sampling statistician. The answers to these questions must be informed by the policy considerations. They have to be made by the people who will use the information to make decisions given the results. Once these requirements have been developed, a sampling expert can then determine the sample composition and sizes needed to meet the requirements.

Defining the Target Customer Population

With large scale educational interventions targeted at the general market, extrapolation is an important consideration. It will almost certainly be necessary in such interventions to study samples of treated and control group customers and to make inferences about the impacts of the educational intervention based on the differences between these samples. Correspondingly it will be necessary to draw representative (i.e., random) samples from the treated and control groups in such a way as to permit calculation of meaningful estimates of the population level impacts using appropriate sampling weights. To calculate weights for purposes of extrapolation, it is necessary to have a list of the members of the treated and control group populations, to sample randomly from those lists and to carefully observe any selection effects that might emerge in the sampling process so that the extrapolation can be adjusted to take account of them.

If precise measurements are needed for specific sub-populations (e.g., customer types or size categories), then it will be necessary to over-sample these customers to ensure that enough observations are present in relevant cells to precisely estimate the impacts of the treatment. These are called sampling strata or blocks as described in Section 3.

Precision of the Estimates

A critical requirement in developing a sample design for any sort of experiment is a clear understanding of the minimum threshold of difference (between treated and control group customers) that is considered meaningful from the point of view of those who will be using the results in 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. Because changes in attitudes and beliefs often result in small or negligible changes in energy consumption in the short run it is difficult to directly translate such changes into cost effectiveness calculations using energy savings. So it is not really possible to directly identify detection thresholds for attitude change for purposes of setting sample sizes (as it is when designing samples to detect a change in energy consumption).

Correspondingly it is probably more appropriate to fall back onto conventional expectations for statistical precision and power that are used in social science investigations. By convention, we recommend that all samples used in measuring changes in beliefs and attitudes related to education programs be designed to produce no more than plus or minus 10% sampling error. That is, the sample sizes should be selected so that a change of at least 10% in survey measurements is required to consider the education program effective.

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 taken from the populations of interest. One can incorrectly conclude that there is a difference between the treatment and control groups when there isn't one (because we are observing samples). This is called a Type I error – also known as alpha. Or one can incorrectly conclude that there isn't a difference when in fact there is one. This is called a Type II error – also known as beta. The challenge in designing experimental samples is to minimize both types of errors. This is done by choosing sample sizes that simultaneously minimize their likelihoods.

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 allowed sampling precision (as described above +/- 5%), and the sample size. This probability 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 isn't one. Be careful not to confuse the sampling precision (+-5%) with the probability of Type I error 5%. They are not the same thing. 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 studies of the impact of education, 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 isn't 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 things 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 sure we want to be that we are not incorrectly finding a statistically significant difference, the less sure we can be that we 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 sample design. The statistical power used in calculating required sample sizes for experiments is subjective and, in modern times, 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 isn't a difference of a specified magnitude when indeed there is one. For Capacity Building experiments, statistical power 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.

Please answer the following questions pertaining to sample planning:

1. **Are the measurements from the experiment to be extrapolated to a broader population?**

a. If yes, indicate whether the sample will be stratified and what variables will be used in the stratification.

b. If no, describe the list of parties from which the sampling will be obtained.

2. **Are precise measurements required for sub-populations of interest?**

a. If yes, describe the sub-populations for which precise measurements are desired.

3. **What is the minimum threshold of difference that must be detected by the experiment?**

4. **What is the acceptable amount of sampling error or statistical precision and acceptable level of statistical confidence (i.e., 90%, 95%, 99%)?**

5. **Will participants be randomly assigned to treatment and control conditions or varying levels of factors under study?**

a. If yes, do you expect subjects to select themselves into the treatment condition?

b. If so, how will you correct for this selection process in the analysis and sample weighting?

6. **If subjects will not be randomly assigned to treatment and control conditions or varying levels of factors under study:**

a. Describe the process that will be used to select customers for the treatment group(s).

b. 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.

7. **If no control group is used, explain how the change in the outcome variables of interest will be calculated.**

Please indicate the proposed sample sizes (within the treatment cells) for the study.

If experiments are contemplated (true or quasi-experiments) please use the table format provided in 4.2.2 to describe the distribution of sample across treatment cells and strata.

7.6 - Protocol 6: Identify the Program Recruitment Strategy

Information/education campaigns typically do not involve recruitment.

7.7 - Protocol 7: Identify the Length of the Study

In evaluating a behavioral intervention it is important to understand the expected time required to carry out the various aspects of the intervention, the expected onset time for the effect of the treatment and its expected persistence after initial treatment. These considerations will determine the length of time that is required to assess the impact of the treatment and thereby determine the length of time for which the situation must be observed.

Please answer the following questions pertaining to the experimental time frame.

1. Is it possible to observe the impacts of the treatment for at least two years?
2. If no, how will the persistence of the effect be determined?
3. Do pre-treatment data for the relevant variables already exist or must time be allowed to obtain pre-treatment data?
4. If pre-treatment data do not already exist, how long must the pre-treatment period be to support the experimental objectives?
5. If pre-treatment data do not already exist, can the experiment be conducted using only post-treatment data, and what adjustments to sample design will be required to employ a post-test-only design?
6. What is the expected amount of time required for subjects to receive and understand the information being provided to them?
7. What is the expected amount of time needed by subjects to implement behavioral changes in response to the information provided?
8. What is the minimum amount of time the effect of the treatment must persist to cost-justify investment on the part of the utility?
9. If the duration of the experiment is shorter than the expected persistence of the treatment how will the determination be made as to whether the effect of the feedback persists long enough to be cost effective?
10. How much time is needed between when the research plan is completed and approved, and when treatments are in place for experimental participants?
11. How much time is required between when the final data are obtained from the experimental observations and when the analysis can be completed?

7.8 - Protocol 8: Identify Data Requirements and Collection Methods

Please complete Table 7-5 identifying the data requirements and data collection methods for each data element required in the evaluation. The table describes three types of data – energy consumption data, data describing the behaviors in question and other data.

Table 7-5 should be completed for as many measurements that will be taken during the course of the study. For example, if electric and gas consumption are to be collected as part of the evaluation then they should be described in separate entries under energy consumption. The description of the variable should include a definition of the variable in sufficient detail as to permit third parties to understand what the measurement is. It should describe the frequency with which the measurement will be taken. For electricity consumption, the variable might be monthly, hourly or even momentarily in the case of electricity consumption or demand. The method of measurement should describe how the data will be collected in as much detail as is required to explain the data collection process. 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, then the technology as well as installation and data collection protocols should be described.

Table 7-5: Measurements

Energy Consumption	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Behaviors of Interest	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	
Other Data	
Description of Variable	
Frequency of measurement	
Method of Measurement	
Issues and Solutions	

Behavior data is information describing the impact of the program on target behaviors. Examples of behavior data that might be appropriate for feedback programs might include: reported recent history of appliance purchases, an inventory of energy saving actions taken since the start of the behavioral intervention, perceptions and opinions about energy use, reported conversations among the family or with neighbors about energy consumption, etc..

Other data includes all kinds of other data that might be useful in evaluating the impacts of the feedback programs including: 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 the behavioral intervention has taken place.

8. Example Applications of the Protocols for Specific Behavioral Interventions

In this section, example applications for the protocols that are specific to each of the different types of behavioral programs are presented.

8.1 Capacity Building Program

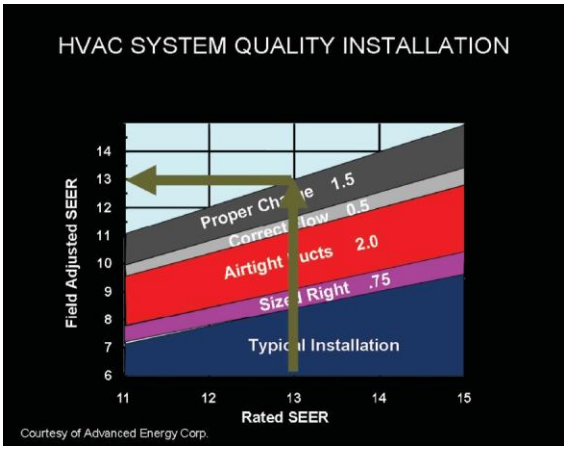
In this section, an example of the application of the evaluation protocols to a training program is presented. It is intended to show the level of depth that is required to meet the requirements of the protocols and to illustrate the types of information that are required to answer the questions in the protocols.

8.1.1 Introduction

The following example of a behavioral training program offered by a Heating, Ventilation and Air Conditioning (HVAC) association was designed to improve the efficiency of installed HVAC units by training parties responsible for designing and installing units in best practices that should be followed during the design and installation processes.

Figure 8-1 graphically displays the relationship between the rated SEER of AC equipment and the SEER that occurs as a result of installation practices – called the field adjusted SEER. It indicates that much of the technical potential for energy efficiency can be lost during the installation process for a variety of reasons that are under the control of the parties who specify the size of the components that are to be installed and those who carry out the installation. The figure indicates that as much as 40% of the technical potential for the energy efficiency of AC systems can be lost if proper design and installation practices are not followed.

Figure 8-1: Impacts of Installation Quality on Realized Energy Efficiency



A program for training personnel responsible for designing and installing AC systems has been developed and implemented. Successful completion of the training course is a condition of becoming a participating contractor in the AC incentive program being offered. The question is: how much impact does this training program have on the design and installation practices used in installing air conditioning systems both in terms of educating the delivery channel and in terms of energy saving.

8.1.2 - Protocol 1: Definition of the Situation

Type of Program

The HVAC Contractor Training Program is a classroom training program consisting of a one day course in best practices to be used in designing and installing HVAC systems. The training program is offered in both winter and spring.

In the program, qualified designers of AC systems and installers receive a one day training course in best practices used in the design and installation of HVAC systems. Subjects covered in the training include:

- Establishing the proper system size
- Matching the coil size to the outdoor condensing unit
- Determination of correct air flow rate
- Design of ducts and sealing practices
- Refrigerant charging
- Commissioning

The Target Population

The target population includes contractor personnel responsible for specifying the components that will be included in HVAC systems and personnel responsible for installing systems in the field.

The Behaviors Targeted for Modification

Parties involved in the design and installation of HVAC systems make a number of decisions that influence performance and efficiency. They do not always follow industry best practices because these practices are sometimes more time consuming and costly to carry out than are other less effective technical procedures. The behaviors that are targeted for change are:

1. **Practices used to identify the size of the air conditioning system to be installed (i.e., tons of capacity)** – to properly size an HVAC system the designer should make a heat gain calculation based on the area of the building, the amount of insulation in the walls and ceiling, the size and types of windows, the orientation of the house and the mount of shading. This design process is time consuming and expensive; and consequently simple it is often substituted by ineffective rules of thumb or simple replacement of pre-existing equipment.
2. **Use of appropriate procedures for matching coil size to exterior condensing** – using ASHRAE reference documents;
3. **Establishment of correct Air Flow over the coil** – using the manufacturer's specifications for the unit
4. **Properly designing and sealing ducts** – ensuring that ducts are installed by professional sheet metal workers and are sealed
5. **Correctly charging the system with refrigerant** – using manufacturers' specifications to established appropriate charge level based on local temperature and pressure conditions
6. **Procedures for commissioning HVAC units** – including proper system startup, cleaning and servicing of ductwork and providing documents and training to occupants concerning the use of the appliance.

The Mechanisms That Are Expected to Change Behavior

Training is designed to make designers and installers aware of the negative consequences of improper installation techniques for comfort and system performance and thereby to cause them to apply best practices in future installations.

Whether the Exposure to Training Can Be Controlled

Training cannot be denied to applicants for several reasons. First, contractors seeking to participate in the program offer are required to complete the training course before they are eligible to become a participating contractor. So, denying contractors access to the training would effectively deny them access to the program an anticompetitive practice that the program should probably avoid. Second, contractors have to schedule their participation into a limited number of available locations for training; and limiting access to contractors at specific locations would undoubtedly cause severe disruptions to the training program and increase the requirement of offering more training in more places than currently are planned.

Table 8-1: Ability to Control and Appropriate Experimental Design

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions – **NO**	Randomized Controlled Trial (RCT)
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **NO**	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **NO**	RCT using recruit and delay tactic
Able to randomly encourage subjects to accept treatment – NO	Randomized Encouragement Design (RED)
Able to assign subjects to treatment based on qualifying interval measurement (e.g., income, usage, building size, etc.) – NO	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	**Quasi-experimental designs**

The Outcomes that Will Be Observed

Several outcomes will be observed during the evaluation. They include:

1. the fraction of AC installation professionals that receive the training;
2. the extent to which professionals who are exposed to the training employ best practices in designing and installing systems
3. changes in attitudes about using best practices as evidence from measurements of beliefs, attitudes and opinions before and after training
4. the improvement in energy efficiency resulting from training of the professionals

8.1.3 - Protocol 2: Description of the Outcome Variables to Be Observed

Table 8-2: Behavioral Outcome and Operational Definition.

Behavioral Outcome	Operational Definition
Training Programs · Beliefs, attitudes and opinions about best practices recommended for designing and installing AC units · Application of best practices in calculating system size requirements and applying other technical and non-technical practices involved in installation.	**Behavior Measures** · Comparison of actual work before and after training for treated trainees, · Comparison of reported installation practices before and after training, · Knowledge and opinions (as measured by test) of trainees and comparison group
Training Programs · Efficiency of installed HVAC systems	**Savings Measures** · Comparison of SEER of systems installed by treated contractors before and after training · Estimated annual, monthly, hourly energy savings given average SEER difference

8.1.4 - Protocol 3: Sub-segments of Interest

According to market research carried out during the development of the training course, sales personnel and installers are responsible for different aspects of the AC installation or replacement process. Sales personnel are primarily responsible for specifying the system components (i.e., size of unit, condenser size, etc.) and installers are responsible for putting the system together in the field. In smaller organizations, the contractor may be responsible for all aspects of the design and installation. In any case, market researchers reported that installers are generally knowledgeable about best practices, but may not apply them because of practical barriers associated with concern about the willingness of buyers to accept increased time and cost associated with doing the job right. They also indicated that sales personnel sometimes did not have the technical training required to carry out best practices.

Therefore, it is appropriate to segment the training market according to these basic job categories listed in Table 8-3.

Table 8-3: Segments of Interest

Segments of Interest
Two different job classifications that are of concern in this training program. They are:
· **Sales/design personnel**– back office personnel who work with customers to specify the design and cost of the system that will be installed on the premises of interest
· **Installers** – field personnel who are responsible for installing and commissioning the HVAC system

8.1.5 - Protocol 4: The Proposed Research Design

Table 8-4 summarizes the situation leading to the proposed research design.

Table 8-4: Behavior and Energy Consumption Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?	NO	NO
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?	The pre-treatment measurements on behavioral indicators will be taken prior to commencement of classroom instruction	NO
How long of a pre-treatment period of data collection is required?	N/A	N/A
Is a control group (or groups) required for the experiment?	NO	NO
Is it possible to randomly assign observations to treatment and control groups?	NO	NO

It is not possible to control the assignment of trainees to treatment and control groups in this case. However, the program is being offered in successive years in the same geographical locations to the same populations of students (i.e., installers and sales personnel in HVAC contracting firms); and, given this situation, it is possible to compare the knowledge, opinions and installation practices used by parties who have received training with the knowledge, opinions and installation practices of those who have not. This effort requires:

- Surveying students concerning their knowledge, opinions and installation practices during the training period. This survey will be designed to observe the knowledge that student retained from the course, their beliefs about the importance of using best practices as well as their reported use of best practices. It should also contain a section designed to observe their report of the extent to which the training changed their practices.
- Surveying the following years students concerning their knowledge, opinions and installation practices prior to training. This survey will be more or less identical in content to the survey carried out with the previous years students
- Comparison of installations of HVAC systems completed in the summer and fall of a given year by parties who were trained within the same year to that of HVAC installations and trainees from the subsequent year. Careful engineering reviews of the subject installations before and after training should be carried out to determine whether:
  a. they have been properly sized;
  b. the coil has been properly matched with the outdoor condensing unit
  c. the air flow rate is correct
  d. the ducts are properly connected and sealed
  e. the refrigerant charge of the unit(s) is correct
  f. it was properly commissioned.

8.1.6 - Protocol 5: The Sampling Plan

All of the parties who seek training under the program will receive training and in an ideal world the experience of the entire population of students would be used to assess the impacts of the program. However, the measurements required to assess the effectiveness of the program are expensive. In order to compare the survey responses of parties who received training between years, it will be necessary to intensify follow up survey efforts with all parties to ensure that response rates are nearly identical for both groups. This is necessary because even small differences in response rates might be responsible for subtle differences in survey results between the two groups and thus invalidate the comparisons that are sought. Intensive follow up efforts may require repeated contacts with survey respondents and significant economic incentives. Such intensive survey efforts will lead to relatively expensive survey costs.

Moreover, comparisons of the installation practices before and after training must be carried out by qualified field engineers who will spend at least two hours at each site. This will lead to engineering evaluation costs of approximately $300 per site.

The sample sizes selected for this evaluation are sufficient to measure the prevalence of knowledge, opinions and installation practices to within plus or minus 10% precision with 95% confidence. The sample sizes required for each of the study elements are shown in Table 8-5.

Table 8-5: Study Element and Sample Size (Example)

Study Element	Sample Size
Survey of year 1 (Y1) trainees	· 100 sales personnel · 100 installers
Survey of year 2 (Y2) trainees	· To be completed on intake into the classroom for all year 2 trainees
Survey of installations	· 100 installations made by Y1 trainees in Y1 · 100 installations made by Y2 trainees in Y1 · 100 installations made by Y1 trainees in Y2 · 100 installations made by Y2 trainees in Y2

8.1.7 - Protocol 6: The Program Recruitment Strategy

Contractors are recruited to the training on a first come, first served basis. All contractors who seek to participate in the program must com- plete the training course prior to the cooling season.

All trainees will be compelled to complete the knowledge, opinions and practice survey prior to their training. However, it will be necessary to collect survey answers from prior trainees by surveying them after the fact of their training. This survey should be carried out using a combination of internet and telephone interviewing; and it should be assumed that a nominal incentive (i.e., $100) will be provided to parties who complete the survey.

It will also be necessary to obtain lists of installations that can be inspected to determine the degree to which trainees are adopting and maintaining best practices for trainees completed. To ensure the cooperation of contractors, it should be assumed that surveyors will provide a nominal incentive to contractors for each address they provide for evaluation. Each contractor will be asked to provide 10 addresses for review with a nominal incentive (i.e. $25 per address). Homeowners will also be provided with incentives for permitting evaluators to review their installation.

8.1.8 - Protocol 7: The Length of the Study

The extent to which trainees adopt and use the practices contained in the training can be observed immediately after training takes place. It will also be possible to observe the persistence of the practices that are adopted by examining installations that are made by year one trainees in the second year after their training. The period of the study is two years.

8.1.9 - Protocol 8: Data Collection Requirements

Table 8-6 describes the data collection requirements for the evaluation. It outlines three types of data that will be collected during the study energy consumption data measured at sites where trained and untrained installers are working; compliance with best practices measured at sites where trained and untrained installers are working and results of survey measurements of knowledge and reported applications of best practices before and after training.

Table 8-6: Data Collection Requirements

Energy Consumption		
Variable	Definition	Method
Rated SEER	Rated Efficiency	Manufacturer published
Adjusted SEER	Realized Efficiency	Field measured by Technician
Use of Best Practices		
Variable	Definition	Method
Best Practice Size	Unit Sized properly	Inspector observation
Best Practice Coil	Coil sized properly	Inspector observation
Best PracticeAir Flow	Air flow correct	Inspector observation
Best Practice Ducts Connected	Ducts performing properly	Inspector observation
Best Practice Ducts Sealed	Ducts performing properly	Inspector observation
Best Practice Charging	System properly charged	Inspector observation
Best Practice Commisioning	System properly started	Inspector observation
Initial Knowledge of Best Practices		
Variable	Definition	Method
Best Practice Size	Contractor understands best practices	Contractor responses to survey questions before training
Best Practice Coil	Contractor understands best practices	Contractor responses to survey questions before training
Best PracticeAir Flow	Contractor understands best practices	Contractor responses to survey questions before training
Best Practice Ducts Connected	Contractor understands best practices	Contractor responses to survey questions before training
Best Practice Ducts Sealed	Contractor understands best practices	Contractor responses to survey questions before training
Best Practice Charging	Contractor understands best practices	Contractor responses to survey questions before training
Best Practice Commisioning	Contractor understands best practices	Contractor responses to survey questions before training
Knowledge of Best Practices After Training		
Variable	Definition	Method
Best Practice Size	Contractor understands best practices	Contractor responses to survey questions after training
Best Practice Coil	Contractor understands best practices	Contractor responses to survey questions after training
Best PracticeAir Flow	Contractor understands best practices	Contractor responses to survey questions after training
Best Practice Ducts Connected	Contractor understands best practices	Contractor responses to survey questions after training
Best Practice Ducts Sealed	Contractor understands best practices	Contractor responses to survey questions after training
Best Practice Charging	Contractor understands best practices	Contractor responses to survey questions after training
Best Practice Commisioning	Contractor understands best practices	Contractor responses to survey questions after training
Reported Use of Best Practices		
Variable	Definition	Method
Best Practice Size	Contractor uses best practices	Contractor responses to survey questions before training
Best Practice Coil	Contractor uses best practices	Contractor responses to survey questions before training
Best PracticeAir Flow	Contractor uses best practices	Contractor responses to survey questions before training
Best Practice Ducts Connected	Contractor uses best practices	Contractor responses to survey questions before training
Best Practice Ducts Sealed	Contractor uses best practices	Contractor responses to survey questions before training
Best Practice Charging	Contractor uses best practices	Contractor responses to survey questions before training
Best Practice Commisioning	Contractor uses best practices	Contractor responses to survey questions before training
Reported Use of Best Practices After Training		
Variable	Definition	Method
Best Practice Size	Contractor uses best practices	Contractor responses to survey questions after training
Best Practice Coil	Contractor uses best practices	Contractor responses to survey questions after training
Best PracticeAir Flow	Contractor uses best practices	Contractor responses to survey questions after training
Best Practice Ducts Connected	Contractor uses best practices	Contractor responses to survey questions after training
Best Practice Ducts Sealed	Contractor uses best practices	Contractor responses to survey questions after training
Best Practice Charging	Contractor uses best practices	Contractor responses to survey questions after training
Best Practice Commisioning	Contractor uses best practices	Contractor responses to survey questions after training

8. Example Applications of the Protocols for Specific Behavioral Interventions

8.2 Education or Awareness Campaign

In this section, an example of the application of the evaluation protocols to an education/awareness campaign is presented. It is intended to show the level of depth that is required to meet the requirements of the protocols and to illustrate the types of information that are required to answer the questions in the protocols.

8.2.1 Introduction

The following is an example of an awareness campaign sponsored by a Public Utilities Commission over a three year period. During each of the campaign years a randomly chosen subset of approximately 1/3rd of all small and medium sized commercial and industrial customers served by investor owned utilities were defaulted on to time varying rates. A public information campaign was implemented to ensure that customers understand how costs change with time of day; that their electricity costs might change as a result of being assigned to the new rate; that there were actions they could take to avoid cost increases and that they could no longer receive service under their former rates,. In this campaign, customers who are about to be defaulted were informed by direct mail and telephone of the rate change; and what they might be able to do to control their energy costs.

As part of the ongoing effort to ensure that customers are informed, an evaluation of the effectiveness of the information campaign was undertaken. The objective of the evaluation was to determine how effective the information campaign was in informing customers of the impending rate change and what they might do about it.

8.2.2 - Protocol 1: Definition of the Situation

Type of Program

The Awareness Campaign was designed to inform selected non-residential customers that they are about to be defaulted onto time varying rates. The information campaign was carried out over three consecutive years prior to the default assignment of selected customers onto time varying rates in November of each year. In the months preceding November, customers receive bill inserts, direct mail letters and, for customers who might experience large cost increases telephone calls informing them of the impending change in their rates.

The purpose of the information campaign was to ensure that customers understand that their rates are going to change; that in some cases their electricity costs may increase; that they can lower their electricity costs by reducing their electricity consumption overall and by changing the time of day during which they used electricity. The information campaign also explained why the rate change was necessary and that customers will no longer be able to subscribe to flat rates.

The Target Population

The target population included non-residential customers that were assigned to time varying rates in each defaulting period (i.e., November of each year). Within these overall populations there was also a need to provide more intensive effort to inform customers that are likely to experience relatively large bill impacts.

The Behaviors Targeted for Modification

Defaulting non-residential customers to time varying rates is expected to cause them to lower their electricity consumption during peak hours possibly shifting consumption to periods before and after the peak period. Customers can make a wide variety of changes to reduce their electricity costs under time varying rates. These include:

- Pre-cooling their businesses to reduce the amount of energy required to run air conditioning during the peak;
- Replacement of inefficient equipment with equipment that will use less electricity during the peak; and
- Reducing their demand for electricity during the peak by turning off unneeded equipment.

To undertake any of the above actions, customers must be aware of the impeding change in their rates; understand how their electricity costs might be affected and understand how they can lower those costs.

The Mechanisms that Are Expected to Change Behavior

The information campaign was intended to make customers aware of the impending rate changes and inform them of the actions they can take to control their electricity costs. Customers were expected to change the timing and magnitude of their electricity consumption after they were informed.

Whether the Exposure to Education Can Be Controlled

Education cannot be denied to parties who were about to be defaulted onto a time of use rate. Indeed the entire purpose of the information campaign was to ensure that all parties who were about to experience a significant rate change, were aware of it and understood how to respond to it.

Table 8-7: Ability to Control and Appropriate Experimental Design

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions – **NO**	Randomized Controlled Trial (RCT)
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **NO**	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **NO**	RCT using recruit and delay tactic
Able to randomly encourage subjects to accept treatment – NO	Randomized Encouragement Design (RED)
Able to assign subjects to treatment based on qualifying interval measurement (e.g., income, usage, building size, etc.) – NO	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments	**Quasi-experimental designs**

The parties who were defaulted onto time varying rates in each year were a randomly selected subset of all non-residential customers. A randomly selected subset of non-residential customers was defaulted onto time varying rates in November of program year one. In November of the subsequent year, program year 2, another randomly selected subset of non-residential customers was defaulted; and another randomly selected subset was defaulted in program year three. While the evaluator was not in direct control of the assignment of customers to the year during which the information program was carried out, the random selection of customers to default each year and the annual timing of the notification and defaulting process, made it possible to interpret the results of the notification campaign as though it was a true experiment.

The Outcomes that Will Be Observed

The outcomes of interest for this program were the customers' understanding of how time varying rates work; their awareness of the fact that they were about to be defaulted on to time varying rates; their understanding that their electricity costs may change as a result of the change to time varying rates and their understanding of the options they have for controlling their costs when they were defaulted.

8.2.3 - Protocol 2: Description of the Outcome Variables to Be Observed

Table 8-8: Behavioral Outcomes and Operational Definition

Behavioral Outcome	Operational Definition
Awareness Campaign · Understanding of time of use rates · Awareness that they will be defaulted on to time varying rates in November of the assignment year · Understanding that their cost of electricity may change when they are assigned to time varying rates · Awareness of changes they can make in their operation in order to lower their electricity consumption · Recollection of the sources of information through which they received information.	**Behavior Measures** · Comparison of reported knowledge about time of use rates, awareness of impending change in rates, understanding of likely bill impacts and awareness of cost saving alternatives for customers who have been exposed to the awareness campaign and those who have not been exposed to the awareness campaign, · Information to be obtained by surveying parties who were and were not exposed to the awareness campaign in summer and fall of program year two.
Awareness Campaign · Change customer load shape	**Load Impact Measures** · Comparison of changes in load shapes for customers who have been defaulted on to time varying rates and those who have not – using interval data supplied by utilities

8.2.4 - Protocol 3: Sub-segments of Interest

The cost differentials for the time varying rates to which customers were being defaulted are not very extreme. So, most customers will not experience very large bill impacts as a result of the rate change. However, some customers with very large energy use and customers with very significant usage on-peak may experience very large bill impacts. Customers who were expected to experience large expected bill impacts received more intensive communications efforts. An effort was made by utility representatives to contact these customers personally to ensure they were informed of the impending rate change and the likely consequences for their electricity cost.

Since the awareness program is different depending on the expected impact of the rate change on the customers, and the fraction of customers who will experience significant bill impacts is relatively small (i.e., about 10%), it made sense to focus on these two different segments during the evaluation.

Table 8-9: Segments of Interest

Segments of Interest
Two different customer types are of concern during this awareness evaluation. They are:
· Customers who will experience relatively small bill impacts (i.e., < 5% changes) as a result of being defaulted on to time varying rates.
· Customers who will experience significant bill impacts (i.e., > 5% changes) as a result of being defaulted on to time varying rates.

8.2.5 - Protocol 4: The Proposed Research Design

Table 8-10 summarizes the situation leading to the proposed research design.

Table 8-10: Behavior and Energy Consumption Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?	NO	YES
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?	NO	YES
How long of a pre-treatment period of data collection is required?	N/A	1 Year
Is a control group (or groups) required for the experiment?	YES	YES
Is it possible to randomly assign observations to treatment and control groups?	NO*	NO*

While it is not possible to control the assignment of customers to treatment and control groups in this case; as explained above, customers were randomly assigned to three cohorts for purposes of defaulting them to the new time varying rates. One of the randomly chosen groups was defaulted onto time of use rates in year one. Another was defaulted in year two and the final group was defaulted in year three. Because of random assignment, the year two and year three groups were identical in all respects save the fact that the year two group received the awareness campaign in the fall of of year two.

In effect, this program design produced a randomized controlled trial (RCT) with a delayed treatment (for the parties who will experience the awareness campaign in year three).

The effects of the awareness campaign on customer knowledge and awareness of the impending rate change were measured by surveying the following groups of customers:

- those who were exposed to the awareness campaign in fall of year one, were subsequently defaulted on to time varying rates and experienced those rates for a period of approximately 14 months;
- those who were exposed to the awareness campaign in year two and were subsequently defaulted on to time varying rates in November of year two (i.e., those who experience the awareness campaign in the fall of year two); and
- those who have not yet been exposed to the awareness campaign.

The questions on the surveys concerning the customers' knowledge of time varying rates, the likely impacts of those rates on their electricity cost, the actions they can take to minimize their costs and their awareness that they are about to be defaulted on to those rates were basically identical for all three surveys. However, customers who were defaulted in year one will also be asked about their experience with the new rates and whether they have made any changes in their operation in response to the price changes. Those who were defaulted in year two will also be asked about their plans or intentions to change their operations in anticipation of the need to lower the impacts of time varying rates on their electricity costs.

Customers who did not experience the awareness campaign until year three provided measurements of the levels of knowledge and awareness that were present absent the information campaign.

8.2.6 - Protocol 5: The Sampling Plan

As explained above, to assess the effectiveness of the awareness campaign customers who do and do not experience the awareness campaign will be surveyed. The population receiving the awareness campaign each year is relatively large (i.e., >150,000) and survey measurements of the kind required to observe the impacts of the awareness campaign are expensive. In order to compare the survey responses of parties who are exposed to the awareness campaigns in the various years, it will be necessary to intensively follow up survey efforts with all parties to ensure that response rates are nearly identical for all populations under study. This is necessary because even small differences in response rates might be responsible for subtle differences in survey results between the study groups and thus invalidate the comparisons that are sought. Intensive follow up efforts may require repeated contacts with survey respondents and significant economic incentives. Such intensive survey efforts will lead to relatively expensive survey costs. For these reasons it will be necessary to sample customers for purposes of surveying.

The sample sizes selected for this evaluation are sufficient to measure the prevalence of knowledge, opinions and reactions to rate changes to within plus or minus 5% precision with 95% confidence. The sample sizes required for each of the study elements are shown in Table 8-11.

Table 8-11: Study Element and Sample Size (Example)

Study Element	Sample Size
Survey of customers receiving information in the 2012 awareness campaign	· 150 customers with high bill impacts · 250 customers with normal bill impacts
Survey of customers receiving information in the 2013 awareness campaign	· 150 customers with high bill impacts · 250 customers with normal bill impacts
Survey of customers who have not experienced awareness campaign	· 150 customers with high bill impacts · 250 customers with normal bill impacts

8.2.7 - Protocol 6: The Program Recruitment Strategy

Lists of parties who experienced either the normal or enhanced awareness campaigns during year one or year two will be obtained from the investor owned utilities, along with lists of customers who have not yet been exposed. These lists will be used for sampling customers into the required surveys.

To ensure the cooperation of customers selected for the study, surveyors will provide a nominal incentive to customers who complete the survey forms on the internet, in the mail or on the telephone. The incen- tive will be $40.

8.2.8 - Protocol 7: The Length of the Study

The awareness campaign is taking place over a three year interval. The impacts of the information campaign will be assessed during the second year.

8.2.9 - Protocol 8: Data Collection Requirements

Table 8-12 describes the data collection requirements for the evaluation. It outlines two types of data that will be collected during the study hourly electricity load data measured for parties who were and were not exposed to the awareness campaigns before and after exposure and survey measurements indicating the impacts of the awareness campaigns on knowledge, awareness and planned actions related to electricity consumption. The same survey instrument is used on all three treatment populations and for most of the questions on the survey it is possible to compare the responses from the different populations to discern the impacts of the awareness program.

Table 8-12: Data Collection Requirements

Energy Consumption		
Variable	Definition	Method
Electricity Consumption	Hourly electricity loads for one year	IOU hourly load measurements
	before and one year after exposure to	
Knowledge of Time of Use Rates		
Variable	Definition	Method
Understanding of current rate	What kind of rate do they think they have	Survey Response
Heard of TOU	Have they heard of TOU	Survey Response
How did they hear	How did they hear about TOU	Survey Response
Understanding of timing	Whether they understand summer time periods	Survey Response
Understanding of summer peak time	Whether they understand summer peak period pricing	Survey Response
Understanding of timing	Whether they understand winter time periods	Survey Response
Understanding of summer peak time	Whether they understand winter peak period pricing	Survey Response
Awareness of 2013 transition messages		
Variable	Definition	Method
Do they recall	Do they recall being informed of upcoming change	Survey Response
When	Do they recall when the change is to take place	Survey Response
How they received notice	How did they receive the notice	Survey Response
Awareness that flat rates are phased out	Do they understand they can't go back to flat	Survey Response
Awareness of possible bill change	Do they understand this may affect their bill	Survey Response
Understanding of how to control cost	Do they understand they can reduce their cost	Survey Response
Perceived ability to respond	Do they believe they can lower their bill	Survey Response
How they think their bill will change	Whether their bill will increase, decrease or same	Survey Response
Have you been advised	Has the utility advised them how to lower their bill	Survey Response
Have you been advised	Has the utility advised them to go to the website	Survey Response
Taken any actions	Have they taken any actions to to lower costs	Survey Response
Plan to take action	Do they have any actions planned	Survey Response
What actions	What actions do they plan to take	Survey Response
Awareness of 2012 transition messages		
Variable	Definition	Method
Do they recall	Do they recall being informed of upcoming change	Survey Response
When	Do they recall when the change is to take place	Survey Response
How they received notice	How did they receive the notice	Survey Response
Awareness that flat rates are phased out	Do they understand they can't go back to flat	Survey Response
Awareness of possible bill change	Do they understand this may affect their bill	Survey Response
Understanding of how to control cost	Do they understand they can reduce their cost	Survey Response
Reduced load during summer peak	Has your rirm reduced load during summer peak	Survey Response
Bill Change	How has your bill changed since default	Survey Response
Did information help	Did the information provided by utility help control cost	Survey Response
What steps were taken	What steps were taken to try to control cost	Survey Response
Plan to take action	Do they have any actions planned	Survey Response
What actions	What actions do they plan to take	Survey Response
Awareness by Control Customers		
Variable	Definition	Method
Do they recall	Do they recall being informed of upcoming change	Survey Response
When	Do they recall when the change is to take place	Survey Response
How they received notice	How did they receive the notice	Survey Response
Awareness that flat rates are phased out	Do they understand they can't go back to flat	Survey Response
Awareness of possible bill change	Do they understand this may affect their bill	Survey Response
How they think their bill will change	Whether their bill will increase, decrease or same	Survey Response
Have you been advised	Has the utility advised them how to lower their bill	Survey Response
Taken any actions	Have they taken any actions to to lower costs	Survey Response
Plan to take action	Do they have any actions planned	Survey Response
What actions	What actions do they plan to take	Survey Response

8. Example Applications of the Protocols for Specific Behavioral Interventions

8.3 Information Feedback Programs

In this section, an example of the application of the evaluation protocols to an information feedback campaign is presented. It is intended to show the level of depth that is required to meet the requirements of the protocols and to illustrate the types of information that are required to answer the questions in the protocols.

8.3.1 Introduction

The following is an example of a pilot information feedback program. The pilot includes a combination of feedback mechanisms including:

- A welcome package explaining the purpose of the Home Energy Reports (HER);
- Printed Energy Reports (ER)s delivered 5 times per year comparing selected consumers with neighbors and efficient neighbors and occasionally providing information promoting utility sponsored energy efficiency offerings; and
- A website portal allowing customers to access detailed information about their energy consumption along with the ability to set energy savings goals, track progress and obtain energy saving recommendations.

8.3.2 - Protocol 1: Definition of the Situation

Type of Program

The pilot is designed to evaluate the behavior change and energy savings resulting from providing a combination of information feedback techniques to selected customers. The core of the pilot program is a printed direct mail report that is periodically sent to households that contains a graphical comparison of the electricity (and sometimes gas) consumption of the subject household with that of "neighbours" and efficient "neighbours". The neighbours and efficient neighbors are households located nearby with homes of similar size and age (if known). In addition these reports sometimes contain recommended energy savings tips and promotions of utility sponsored energy efficiency programs. In addition to printed reports the Pilot will provide a web portal to customers allowing them to observe their electricity consumption; to set energy saving goals; to track their progress toward goals and to receive and process energy savings recommendations.

The Target Population

The target population includes residential customers.

The Behaviors Targeted for Modification

Residential customers engage in a wide range of behaviors that can be affected by the information in ERs. They control the utilization of lighting, the temperature of the thermostat in the home, the use of office and home entertainment equipment, water temperatures used in showering, clothes and dish washing, the length of dish and cloths washing cycles and the purchase of energy using equipment from light bulbs to major appliances. All of these choices are behaviors that are subject to modification by HER feedback. Changes in these behaviors are expected to produce changes in energy consumption.

The Mechanisms that Are Expected to Change Behavior

ERs are designed to modify consumer behavior by providing consumers with a normative comparison to other "similar" households. According to normative theory, in situations in which humans are uncertain about how to behave or how the world should appear, they often formulate their intentions and opinions by referring to the experience of others who they respect. In the case of energy consumption, consumers have no basis for determining whether the amount of energy they are using is normal compared to the behavior of others. In theory, providing high users with information that indicates that they are using a large amount of energy should cause them to investigate their energy use in an effort to identify whether they are engaging in wasteful practices that are leading their energy use to be abnormally high. As a result of these investigations consumers are likely to modify energy use related behaviors in order to lower their energy consumption.

Whether the Exposure to the Feedback Can Be Controlled

It is possible to control the presentation of feedback in the ERs and the proposed website. An RCT is the most powerful research design available for studying behavior. It should be used in this study.

Table 8-13: Ability to Control and Appropriate Experimental Design

Ability to Control	Appropriate Experimental Design
Able to randomize presentation of treatment – mandatory assignment of subjects to treatment and control conditions – **YES**	**Randomized Controlled Trial (RCT)**
Able to deny treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **YES**	RCT using recruit and deny tactic
Able to delay treatment to volunteers – mandatory assignment of volunteers to treatment and control conditions – **YES**	RCT using recruit and delay tactic
Able to randomly encourage subjects to accept treatment – YES	Randomized Encouragement Design (RED)
Able to assign subjects to treatment based on qualifying interval measurement (e.g., income, usage, building size, etc.) – YES	Regression Discontinuity Design (RDD)
Unable to assign subjects to treatments – NO	Quasi-experimental designs

The Outcomes that Will Be Observed

The outcomes of interest for this program are the customers' awareness of the ERs, their acceptance of the characterization of their energy use provided in the ERs (i.e., whether it is abnormally high or low), their use of the website and their energy use.

8.3.3 - Protocol 2: Description of the Outcome Variables to Be Observed

Table 8-14: Behavioral Outcome and Operational Definition

Behavioral Outcome	Operational Definition
Feedback · Awareness of the ERs · Reported website access · Whether they find the information contained in ERs credible · Whether they believe they are using an relatively large amount of energy · Whether they believe it is important to control their energy use · Whether they have identified changes in their energy use to lower their energy consumption · What actions they have taken to lower their energy use	**Behavior Measures** · Representative samples of treatment and control group customers will be surveyed to observe their answers to questions designed to measure the behavioral outcomes described on the left side of the table. · The frequency and extent of website access by parties in the treatment and control groups will be observed and compared.
Feedback · Change in energy consumption	**Energy Consumption** · Energy consumption for the treatment and control groups will be measured for one year before the onset of the feedback treatment, during the feedback period and after the feedback is removed. Monthly usage information will be used to compare the change in energy consumption

8.3.4 - Protocol 3: Sub-segments of Interest

Past implementations of neighbor based comparison programs have shown that the magnitude of savings varies with the magnitude of the customer energy use. Accordingly, customers in the top two quartiles of energy use display the highest relative response to the ERs. However, since approximately 25% of customers in a random sample will naturally fall into each usage segment, there is no need to stratify by this variable.

Table 8-15: Segments of Interest

Segments of Interest
· None required/

8.3.5 - Protocol 4: The Proposed Research Design

Table 8-16 summarizes the situation leading to the proposed research design.

Table 8-16: Behavior and Energy Consumption Measures

Question	Behavior Measures	Energy Consumption Measures
Will pre-treatment data be available?	NO	YES
Does the appropriate data already exist on all subjects, or do measurements need to taken in order to gather pre-treatment data?	NO	YES
How long of a pre-treatment period of data collection is required?	N/A	1 Year
Is a control group (or groups) required for the experiment?	YES	YES
Is it possible to randomly assign observations to treatment and control groups?	YES	YES

The research design for this project was a randomized controlled trial (RCT) in which a random sample of 100,000 qualifying residential customers of the utility were randomly divided into two equal sized groups treatment and control. The treatment group was exposed to the feedback contained in the pilot. The experiment ookplace over a two year interval with treatment group customers receiving 5 ERs per year. Treatment group customers received periodic promotional messages in their ERs and have access to a website in which they can study their energy use, set goals, track progress and view their neighbor comparisons. The control group did will not receive ERs and did not have access to the website.

At the conclusion of the first year, treatment and control group customers were surveyed to observe difference in awareness of the messages in the ERs, customers' perceptions of their energy use, their interest in saving energy, the extent to which they think it is important to save energy, and behaviors they are engaging in to save energy.

Energy savings were observed by comparing the energy use of treatment and control households before and after the onset of treatment.

8.3.6 - Protocol 5: The Sampling Plan

Despite the fact that only 25,000 total customers are required to detect a 1% change in energy consumption, the proposed treatment will be provided to 50,000 customers (to realize energy savings from the pilot). Because the pilot serves hundreds of thousands of customers, it will be necessary to select a sample of participating customers.

To select customers to participate in the pilot a random sample of 150,000 residential customer records will be randomly sampled from the customer information system and delivered to the energy report vendor. The vendor then used these records to identify customers who are eligible to receive the treatment. Typically, this involves removing customers for which it is impossible to define neighboring groups. This file will then be returned to the evaluator who will randomly select 50,000 customers to be provided the treatment and 15,000 customers to be designated as control group members. The records for the 50,000 treatment customers will be provided to the report provider for use in preparing an sending reports.

As explained in Protocol 4, samples of treatment and control group customers will be surveyed to collect information regarding their awareness of the ER, their assessment of its relevance to their household, their opinions about the importance of saving energy, and their reports of behaviors that influence energy consumption. It is extremely important that these surveys obtain relatively high response rates and that non-response adjustments are made in the event that significant non-response occurs (i.e., more than 20%). In the ideal case, the surveys will be carried out in person using a cluster sampling technique. Alternatively, the surveyors might employ a combination of direct mail and internet surveying. Telephone surveying should not be used because of the low response rates that are obtained with this method and the known sampling biases that exist in telephone sample frames.

The sample sizes selected for the overall treatment and control groups are sufficient to measure the difference in energy consumption between treatment and control customers to within plus or minus 1% with 95% confidence. The sample sizes for the proposed surveys are sufficient to measure the behavioral measurements to within plus or minus 5% precision with 95% confidence.

The sample sizes required for each of the study elements are summarized as shown in Table 8-17.

Table 8-17: Study Element and Sample Size (Example)

Study Element	Sample Size
Treatment	· 50,000
Control	· 15,000
Survey of treatment group customers	· 450
Survey of control group customers	· 450

8.3.7 - Protocol 6: The Program Recruitment Strategy

As explained in Protocol 5 the list of customers who participate in the treatment and control groups in the pilot will be obtained from a customer information system. Customers who are assigned to the treatment group will receive the treatment by default. That is, unless they opt out of the treatment it will be delivered to them. There is no need, therefore to recruit them.

However, the customers who will be surveyed as part of the study must voluntarily answer the questions that will be posed concerning behavior change. To ensure the cooperation of customers selected for the study, surveyors will provide a nominal incentive to be determined in consultation with EM&V staff at the IESO.

8.3.8 - Protocol 7: The Length of the Study

Evidence from prior studies of similar information feedback applications shows that impacts of ERs on energy consumption continue to grow for at least 18 months and have been observed to occur as long as 24 months after the start of the program. Therefore, it is recommended that the duration of the treatment be at least 24 months.

8.3.9 - Protocol 8: Data Collection Requirements

Table 8-18 describes the data collection requirements for the evaluation. It outlines two types of data that will be collected during the study monthly electricity usage measured for parties who were and were not exposed to the treatment before and after exposure; and survey measurements indicating the impacts of the feedback mechanism on knowledge, awareness and planned actions related to electricity consumption. The same survey instrument will be used on the treatment and control groups for most of the questions on the survey making it is possible to compare the responses from the different populations to discern the impacts of the treatment.

Table 8-18: Data Collection Requirements

Energy Consumption		
Variable	Definition	Method
Electricity Consumption	Electricity consumption before, during and after the treatment	Monthly electricity consumption measurements for the 12 months preceding, during and 12 months following start of the treatment
Awareness of HER and Website		
Variable	Definition	Method
Recall seeing HER	Customer reports whether they recognize report	Survey Response
Fate of HER	Customer reports what they do with HER	Survey Response
Awareness of website	Customer reports whether they have visited we3gsite	Survey Response
Recall of last HER	Customer reports the last month they they received HER	Survey Response
Reported last visit to website	Customer reports the month of last visit to website	Survey Response
Reported use of the website	Customer reports how often they use the website	Survey Response
Rated Helpfulness of the HER	Customer rates Helpfulness of HER	Survey Response
Rated Helpfulness of the web site	Customer rates Helpfulness of website	Survey Response
Reactions to HER Content (only for Treatment Customers)		
Variable	Definition	Method
Recall of HER comparison	Do they recall whether they are high or low users	Survey Response
Acceptance of HER comparison	Do they believe the comparison	Survey Response
Credibility of HER comparison	Do they think the comparison is credible	Survey Response
if not -- why not	Why is it not credible	Survey Response
Like	Customer reports whether they "like" the HER	
How important is it to save energy	Customer reports how important to save energy	Survey Response
Have you made any changes	Customer reports whether they have made changes	Survey Response
What changes were made	Customer reports changes	Survey Response
Do they think they have saved money	Customer reports whether they have saved money	Survey Response
How much saved	Custmer reports how much they have saved	Survey Response
Have you been advised	Has the utility advised them to go to the website	Survey Response
Helpful	Customer reports whether they find the report helpful	Survey Response
Discussed -- Family and friends	Customer reports whether they have discussed report	Survey Response
Energy Related Behaviors -- two poles -- unfettered energy use vs. conservation		
Variable	Definition	Method
Lighting	Which of two polar opposites describes customer	Survey Response
Entertainment	Which of two polar opposites describes customer	Survey Response
Thermostat Heating	Which of two polar opposites describes customer	Survey Response
Thermostat Air Conditioning	Which of two polar opposites describes customer	Survey Response
Showers	Which of two polar opposites describes customer	Survey Response
Clothes washing	Which of two polar opposites describes customer	Survey Response
Clothes drying	Which of two polar opposites describes customer	Survey Response
Office equipment	Which of two polar opposites describes customer	Survey Response
Vampire loads	Which of two polar opposites describes customer	Survey Response
Demographics		
Variable	Definition	Method
Gender	Respondent indicates	Survey Response
Age	Respondent indicates	Survey Response
Education	Respondent indicates	Survey Response
Household Income	Respondent indicates	Survey Response

Conservation Voltage Reduction Impact Evaluation Protocols

April 1, 2019

ieso
Connecting Today.
Powering Tomorrow.

Acknowledgement:
The IESO would like to acknowledge the work of Cadmus in the development of this Conservation Voltage Reduction Impact Evaluation protocols.

Table of Contents

Acronyms and Abbreviations

Acronym or Abbreviation	Definition
AMI	Advanced metering infrastructure
CSA	Canadian Standards Association
CSL	Customer service lines
CVR	Conservation voltage reduction
CV(RMSE)	Coefficient of variation of root mean squared error
EM&V	Evaluation, measurement and verification
IESO	The Independent Electric System Operator
LDC	Local distribution company
LTC	Load tap changer
NMBE	Normal mean bias error
OLS	Ordinary least squares
SCADA	Supervisory control and data acquisition
VAR	Volt-ampere reactive
VVO	Volt-VAR optimization

1. Introduction

Conservation voltage reduction (CVR) makes use of a basic phenomenon in electricity: that power scales with voltage for an ideal resistor ($P=V^2/R$, where P is power, V is voltage and R is electrical resistance). Reducing the voltage on a distribution circuit can reduce the power demand and energy usage on the circuit.

The protocols presented in this document specify which data are required to evaluate an implementation of CVR operated continuously (i.e. not triggered as an emergency response) and how to determine impacts. These impacts will include energy savings and reductions to real and reactive power. Reactive power is a means of quantifying the amount by which the alternating current is not in phase with the alternating voltage. Managing reactive power along a feeder is valuable for controlling voltages downstream from the substation. However, very high reactive power signifies current flowing through the distribution system that isn't delivering useful energy to end users, but that does result in line losses. Lower voltages can also lead to higher efficiencies in power transformers by reducing core and copper losses. The efficacy of CVR is typically expressed as set of a CVR factors, the ratio of the resulting percentage reduction in end-use energy or power demand to the percentage reduction in voltage.

End users will likely never be aware that CVR is occurring unless they have been notified of its implementation. For this reason, evaluation of CVR is primarily focused on verifying the resulting impact on energy and demand savings. Unlike most programs that lead to energy reduction for customers, process evaluations for CVR are generally not a primary evaluation focus. However, it may be valuable for local distribution companies (LDCs) to document the experience and learnings of the distribution engineers working with a CVR technology provider or following the funding processes for the procurement of CVR systems. The protocols presented in this document do not directly address process evaluation.

1.1 Intended Audience

The protocols presented in this paper describe best practices for evaluating CVR impacts and the information required to document methods and results.

While the protocols are written for evaluators calculating impacts, the protocols also may be of interest to program design and implementation staff to ensure future program designs can accommodate evaluation as well as utility system planners interested in understanding how CVR can be incorporated into the planning process. This introduction and Section 2 are directed to all audiences, while sections 3, 4 and 5 are intended for evaluation contractors. Although the final three sections contain detailed instructions that will be valuable for program managers and system planners to review, the first two sections should cover all necessary information for managing the evaluation process and coordinating between all parties. This document is also a resource for designing CVR programs. Understanding the full evaluation process is critical to developing and implementing CVR programs in such a way that impacts can be quantified.

1.2 Overview of CVR

Canadian Standards Association (CSA) C235-83 defines the allowable voltage ranges for

customers' electrical service.[1] LDCs' compliance with this standard gives customers assurance that their equipment can be safely powered from the grid. To monitor and control voltage with CVR, LDCs may deploy and operate specific distribution equipment to precisely manage voltage along the distribution feeders.

Figure 1 depicts the various components of a distribution feeder and the location of the CVR-related components.

Distribution substations distribute power from high-voltage transmission lines to industrial, commercial and residential customers. Transmission voltage is stepped down at the substation with transformers to serve distribution feeders. A distribution feeder transports electricity to customer service lines (CSL). The distribution feeder relies on a variety of equipment to help safely manage the power flow. Equipment examples include circuit breakers, protection relays, fuses, reclosers, capacitor banks, and transformers (sometimes called voltage regulators when not installed at the substation; used to step-up or step-down voltages as necessary).

Substation transformers have a primary side (where the transmission lines are connected) and a secondary side (where the distribution lines are connected). Voltage is generally managed at substations with load tap changer (LTC) transformers, often located in the substation. An LTC transformer regulates the voltage of the distribution feeder by adding or subtracting the number of wire coils on the secondary side of the transformer. These coils are "tapped" into with a mechanical connection to raise voltage with more coils connected and to lower voltage with fewer coils connected. Voltage on distribution lines must be regulated because power transported over long distances will drop in voltage due to various loads, line losses, and transformer losses.

Voltage drop from the substation is generally reduced by specifying larger conductor sizes, but even optimally sized conductors will see voltage drop over the length of the line. Line losses are also managed by utilizing higher voltages—often 12,000 to 20,000 volts—on distribution feeders. However, these high-voltage lines are very hazardous, so transformers are used again to step down voltage for safe operation by end-use appliances, generally 240 volts to 120 volts for a residence. Customer service lines are connected to the low-voltage side of step-down transformers to convert the higher distribution-level voltage down to a usable level for household electrical outlets and other customer circuits.

Power to these circuits is measured with electric meters connected to the service lines. Modern metering systems, which employ advanced metering infrastructure (AMI), provide for remote monitoring of voltage, power and energy usage. AMI meters provide meter information over mesh radio or cellular networks that can be processed for billing and analysis purposes.

[1] CAN3-C235-83 (R2015) - Preferred Voltage Levels for AC Systems, 0 to 50 000 V, 2nd Edition, CSA Group

Figure 1. Example of Distribution Feeder and its Various Components

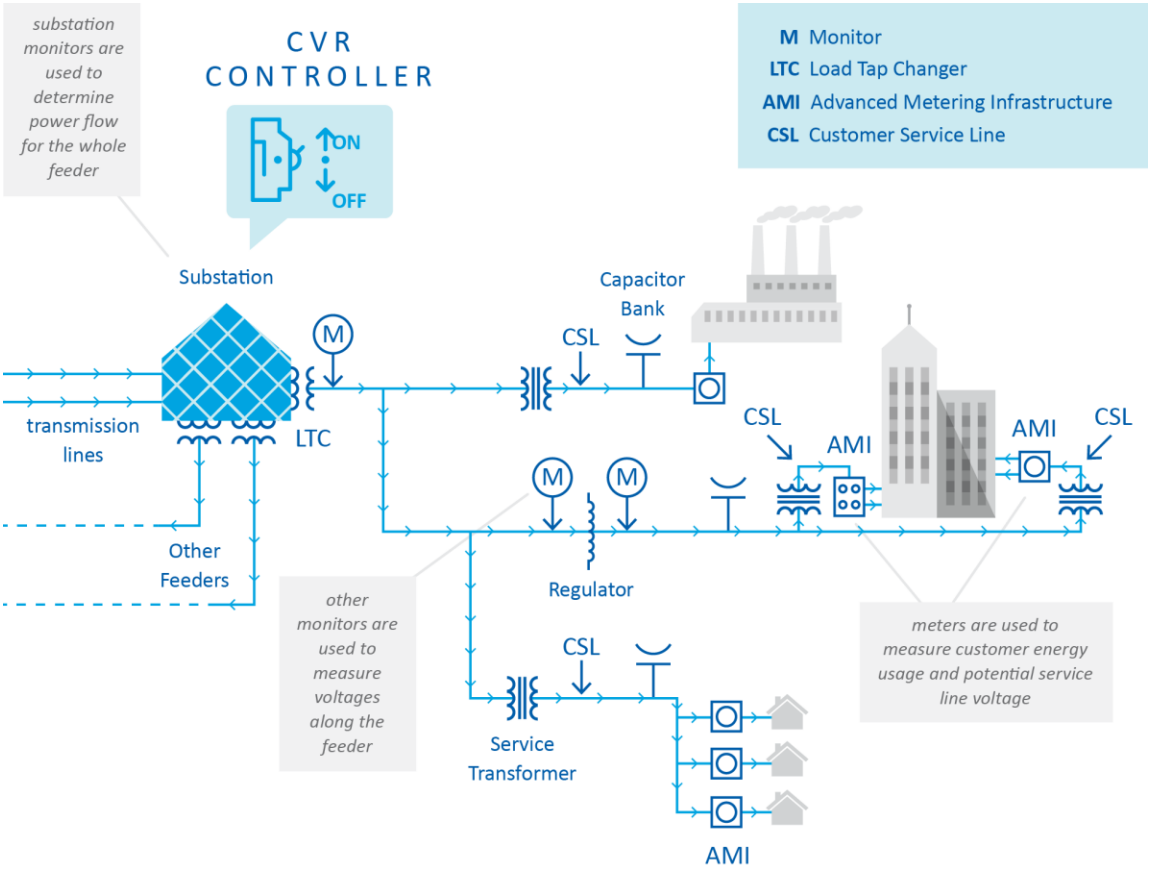
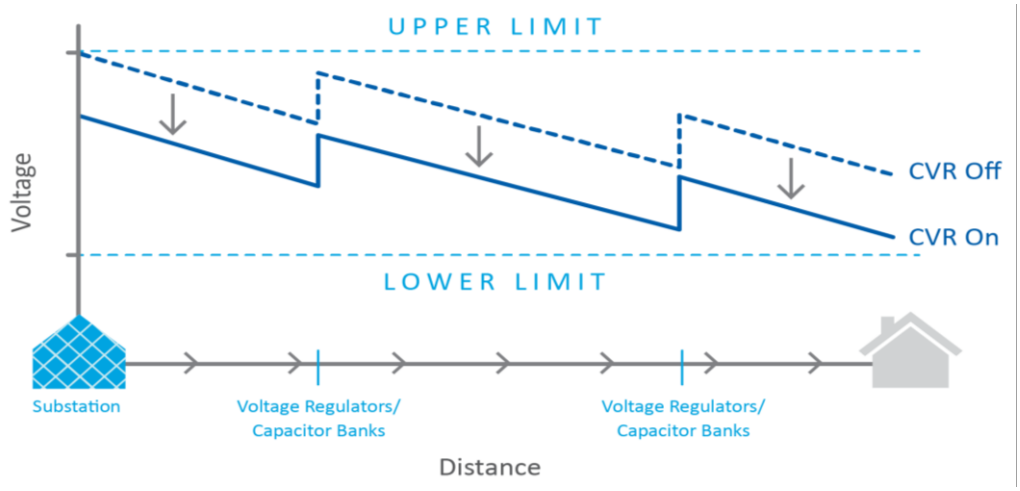


Figure 2. Example of Primary Voltage Profile (not Customer Service Lines) Drop along a Feeder*



*Voltages vary along the distribution feeder and generally decrease farther away from the substation. The two lines show how CVR implementation optimizes equipment to keep customer service lines in the lower range of the allowable band.

Distribution management systems use computer systems to monitor and control equipment connected to substations and the distribution system. Distribution management systems make optimal use of system control and data acquisition (SCADA) to monitor and control substation and distribution system equipment. This includes the collection of voltage readings from monitors or AMI meters and control of a feeder's LTC transformer or feeder-head regulator. A relatively new distribution management capability includes volt/VAR optimization (VVO) systems that control (in addition to a substation's voltage control) voltage regulators and capacitor banks installed along distribution feeders. Capacitor banks on industrial customer service lines have been used for many years to manage reactive power caused by large inductive loads, such as the start-up of large motors. VVO allows finer control of voltage and reactive power on a distribution feeder and more precise voltage reduction across all customers.

Distribution management systems, including VVO and non-VVO systems, can be used to implement CVR by changing the default settings on distribution equipment to operate at a lower base voltage while keeping customer voltages within required ranges. Without CVR, general default settings would be maintained at higher-than-minimum-allowable voltages for customer service lines. The key benefit of systems that implement CVR is that they can effectively reduce power and energy consumption by more precisely controlling voltages near the lower allowable bound, without the risk of dropping out of allowable voltage ranges near the end of the distribution line.

The reduction of distribution feeder voltage and, as a result, the reduction of customer service line voltage can have a complicated impact on power demand and energy usage. In this document, a reduction in power demand has been defined specifically as a reduction in average usage during the peak hours, as defined by the IESO. Additional definitions of peak periods can be similarly applied to account for reductions in demand on distribution equipment. A reduction in peak power demand is more beneficial then reduction at other periods because the degradation of power distribution equipment can be accelerated under higher loading. For non-cycling resistive loads (e.g., incandescent lamps), the reduction of voltage reduces the current, leading to a reduction in power demand. Distribution line losses and transformer losses behave in this manner as well, contributing to the benefits of CVR.

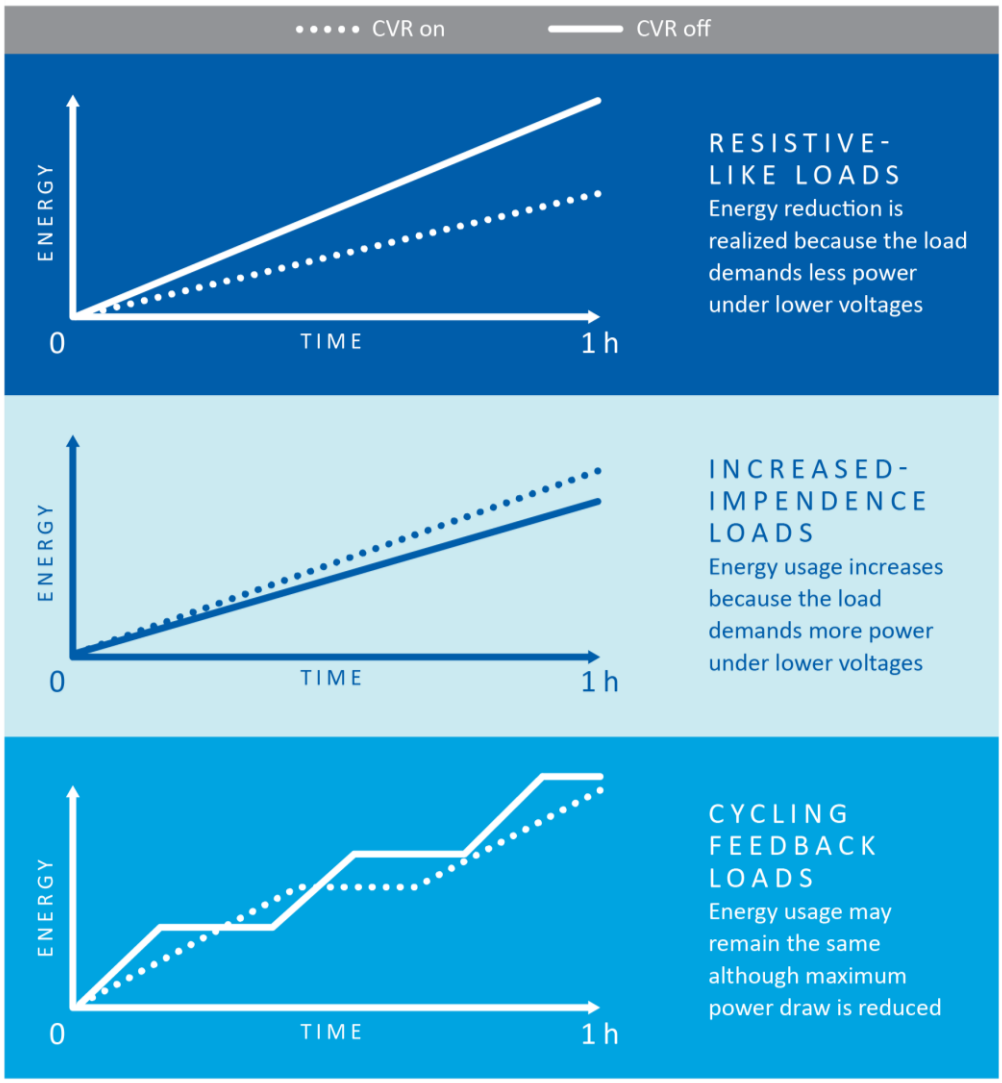
Some non-resistive loads—particularly those with digital circuits (e.g., televisions and computers) or loads with variable frequency drives (e.g., industrial motors)—may, depending on their operating conditions, use *more* energy at lower service voltages. Additionally, loads with thermal cycles or other feedback controls (e.g., electric heaters or dehumidifiers) may compensate for lower voltages by running longer. While the maximum instantaneous power demand may be reduced at lower voltages, compensation by control algorithms may mean that the same total energy would be used over an extended period. (Such compensation can occur, for example, when baseboard heaters run longer when turned on.) Figure 3 shows some examples of how various loads can react to lower voltage with modified energy profiles and changes in power demand. When all customer loads are aggregated for a

distribution feeder, most evaluations programs have found that energy and demand savings were realized by implementing CVR. However, given the wide range of load types across residential, commercial and industrial sectors, evaluations of CVR are always needed to determine and claim impact.

1.3 Evaluation Overview

Utilities and research organizations currently use a wide range of methods and tools to measure and verify CVR performance and savings. Many use traditional power-flow or feeder simulation models, preferably calibrated to measured feeder performance and equipment. Others conduct before-and-after studies under a set of assumptions accounting for seasonal and time-of-day variations. In all cases, it is important that the method recognizes the relationship between voltage and load characteristics (i.e., resistive, inductive, capacitive and cycling loads).

Figure 3. Energy Use Over Time for Load Types with CVR On and CVR Off



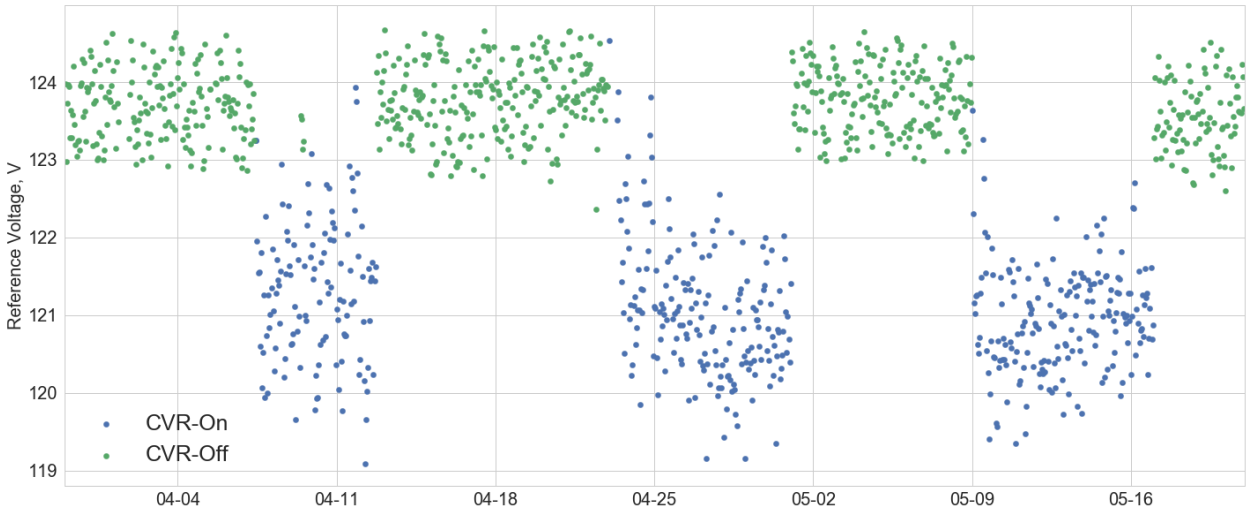
Given the complexity and variability of all the energy-consuming devices served by a distribution feeder, it would be impractical to determine the impacts of CVR using just the principles of power flow and modelled loads. This is because of the extreme cost and complexity of testing necessary to determine how every load would react to a reduction in voltage under all conditions. While assumptions concerning loading characteristics could be made to determine theoretical impacts, this is not preferred when trying to determine the performance of CVR implementation on a specific feeder or customer, when assumptions are nearly impossible to validate. Instead, the best practice approach, as described in these protocols, is to develop statistical models to predict the power demands for the feeder or for customers when the CVR system is on versus off. One of the common methods for evaluating CVR energy savings and demand impact performance is the alternating-periods method. By alternating between days when CVR is active and days when it is not, it is possible to isolate the effect of CVR and to measure its impact on energy usage and power demand. This method includes the following general steps:

- Step 1. Run CVR for a full year with the system alternating between on and off periods to collect necessary system-state data for both cases.
- Step 2. Gather the data and create system-state regression models of the on and off periods with independent variables (weather, weekday/weekend, time of day, etc.).
- Step 3. Apply models to the same dataset describing a typical meteorological year.
- Step 4. Calculate impacts and CVR factors and report the expected savings.
- Step 5. Run any second-stage analysis to compare CVR efficacy across feeders or customers.

This approach requires the CVR implementer to follow a predetermined schedule of repeatedly turning the CVR system on (for three to eight days) and off (for three to eight days) for at least a year so that sufficiently large datasets can be produced for both cases. An example of operating the voltages in this manner is shown below in Figure 4. A full year is necessary to determine annual expected impacts, given the variability in load types between seasons. The full year of usable data can only begin after the system has been fully configured, and when no further modifications are made to control settings and monitoring locations.

As mentioned above, operating the CVR system in this experimental fashion and then producing models from the two datasets is known as the alternating-periods method for evaluating CVR. The method involves specifying regression models of system-state variables (i.e., voltage, real power, reactive power) based on scenario features (i.e., meteorological and other time-series data). Because one set of regression models is fit to CVR-on days, and another set is fit to CVR-off days, two sets of predicted values for the voltages and power can be produced for a typical year.

Figure 4. Example of Voltages Varying as CVR is Turned On and Off

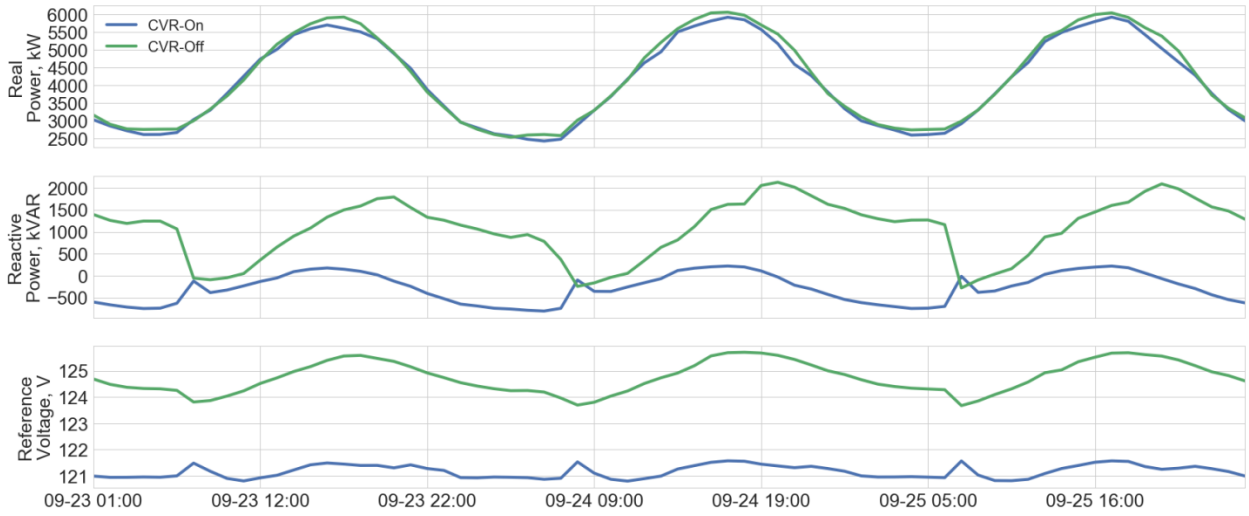


By taking the difference between results for CVR-on and CVR-off system state (i.e., voltage and power) regression models when applied to the same typical meteorological year, expected impacts are determined. If a full year of data is not available, then it is only possible to estimate impacts for the months for which data are available. Models can be developed for each feeder and, optionally, for all customers or only specific customers, depending on the intent of the evaluation. The savings benefits in operating the CVR system can be estimated by subtracting the differences in power for CVR-off days from those of CVR-on days during the same meteorological and similar time periods. The estimated impact of CVR on feeders or aggregated customer groups can be combined in a cross-sectional analysis to try to determine which characteristics are correlated with CVR factors. Figure 5 depicts how the system state estimates for voltage and power will differ between CVR-on and CVR-off models when applied to the same day.

The evaluation protocols for determining expected impacts for a typical year or season are contained in sections 3 through 5. These protocols should not be used for determining demand response implementation of CVR (i.e., CVR that is only implemented for a few hours). The models produced from data taken from CVR running continuously (for at least one day) would not reflect the expected response of a feeder or customer that only had CVR implemented for several hours.

A significant body of literature is available covering the theoretical potential and the evaluated impact of CVR implementation across North America. (Relevant literature is listed in Appendix A Literature Review.) These protocols build upon this previous work to define a standardized—yet generalizable—set of procedures for evaluating the impact of CVR implementations, and for building upon those evaluation results to estimate the potential impacts of CVR in guiding future deployments.

Figure 5. Example of Estimating Hourly Real Power, Reactive Power and Voltage for Three Days from a Typical Year with Both CVR-On and CVR-Off Models



2. Program Management for CVR Evaluation

This section introduces the CVR evaluation process and offers additional guidance for managers overseeing the program, deciding the goals of the evaluation and coordinating the collection of data. Every implementation of CVR will have different data availability given the high variability between distribution companies, feeder configurations, SCADA data and vendor equipment monitored data availability, customer types and implementer technologies. These protocols have been developed to handle this variability and give evaluators guidance on what may be required to evaluate CVR and how to combine data from these varied sources. It will be necessary for evaluation managers to maintain open communication among all parties involved and make sure that the evaluation goals and requirements are communicated and well understood. This section will introduce the metrics used to quantify impacts and give an overview of how they are produced.

2.1 CVR Impact Metrics

The result of each evaluation will be a standard set of impact metrics, outlined below, describing the performance of the CVR implementation. These metrics can be used to describe CVR performance for the whole distribution feeder, for all customers served by a feeder or for individual customers under specific meteorological scenarios, such as a typical summer or during a typical year. These impact metrics are ultimately what will be used to claim impact after the on/off alternating-periods data collection and CVR begins to be on continuously. The metrics are defined below.

Average Voltage Reduction: The average difference between the voltage from the CVR-on model and CVR-off model during the typical meteorological year. For a feeder-level analysis, this voltage value ideally will be an aggregated average from multiple locations on the feeder. Using and aggregating multiple voltage measurements from all CVR-controlled equipment is preferred because voltages can vary significantly across the whole feeder. This feeder-level voltage value should be used if customer impacts are required as a summed group. If the goal is to perform a cross-sectional analysis between individual customers, then average voltage reduction will need to be determined for each customer using measurements from voltage monitors or AMI on customer service lines.

Energy Savings: The quantity of energy that the CVR-off model predicts would have been used beyond what was predicted for the CVR-on model during the typical meteorological year. This could be for a whole year or at least three consecutive calendar months depending on the data available.

Average Demand Reduction: The average difference in demand during a defined period between the CVR-off and CVR-on models requires weighted averages for winter and summer months, as outlined in Table 2 of Section 3.1.2 of this document, to account of transmission level impacts. Additional definitions of peak periods can be developed and applied in a consistent manner to account for reductions in demand on distribution equipment.

Reactive Power Reduction: The average difference in reactive power during a defined period between the CVR-off and CVR-on models. Measuring reactive power requires monitors to take voltage and current readings at

very high frequencies to determine how much these two parameters are in phase. The more that current and voltage are out of phase, the higher the reactive power. Reactive power reflects unnecessary current flowing through the distribution lines, leading to real energy loses. For impact metrics, the difference in reactive power for the CVR-on and CVR-off case should be determined both on average during the whole typical meteorological year as well as during the system peaks, as defined by the IESO. Additional definitions of peak periods can be applied in a consistent manner to account for impacts on distribution equipment.

Isolated Impacts Realized on Distribution System: The impact metrics listed above can be determined for a whole feeder or for customers served by a feeder using the alternating-periods method. If feeder-level impacts and customer-level impacts for all customers on the feeder have been determined, it is possible to take the difference between these two sets to determine impacts that were realized on the distribution equipment. The impacts as determined at for the whole feeder include impacts realized by both the customers and the distribution equipment. Therefore, by subtracting the customer impacts from the feeder impacts, the impacts realized on the distribution system alone can be determined.

CVR Factors: The ratio of energy and demand impacts to voltage reduction. CVR factors should be calculated by normalizing the modelled reductions in energy, demand and reactive power by the corresponding reductions in voltage during the same period. A CVR factor for energy savings equal to 1.0 would signify that for every percentage reduction in voltage there was a percentage reduction in energy usage. CVR factors can be calculated by substation, feeder, or customer group.

In addition to the metrics listed above—which would be produced for specific feeders, distribution equipment, customer groups or customers—a second-stage analysis may be conducted to understand why the efficacy of CVR may have differed between various systems. This would be called a cross-sectional analysis because the intent would be to determine which feeder or customer characteristics are strong predictors of CVR factors. This analysis builds upon both impact metrics determined using the alternating-periods method as well as descriptive characteristic data. Descriptive characteristic data are a core component of this analysis and may require a significant effort to collect if not already tracked by the LDC.

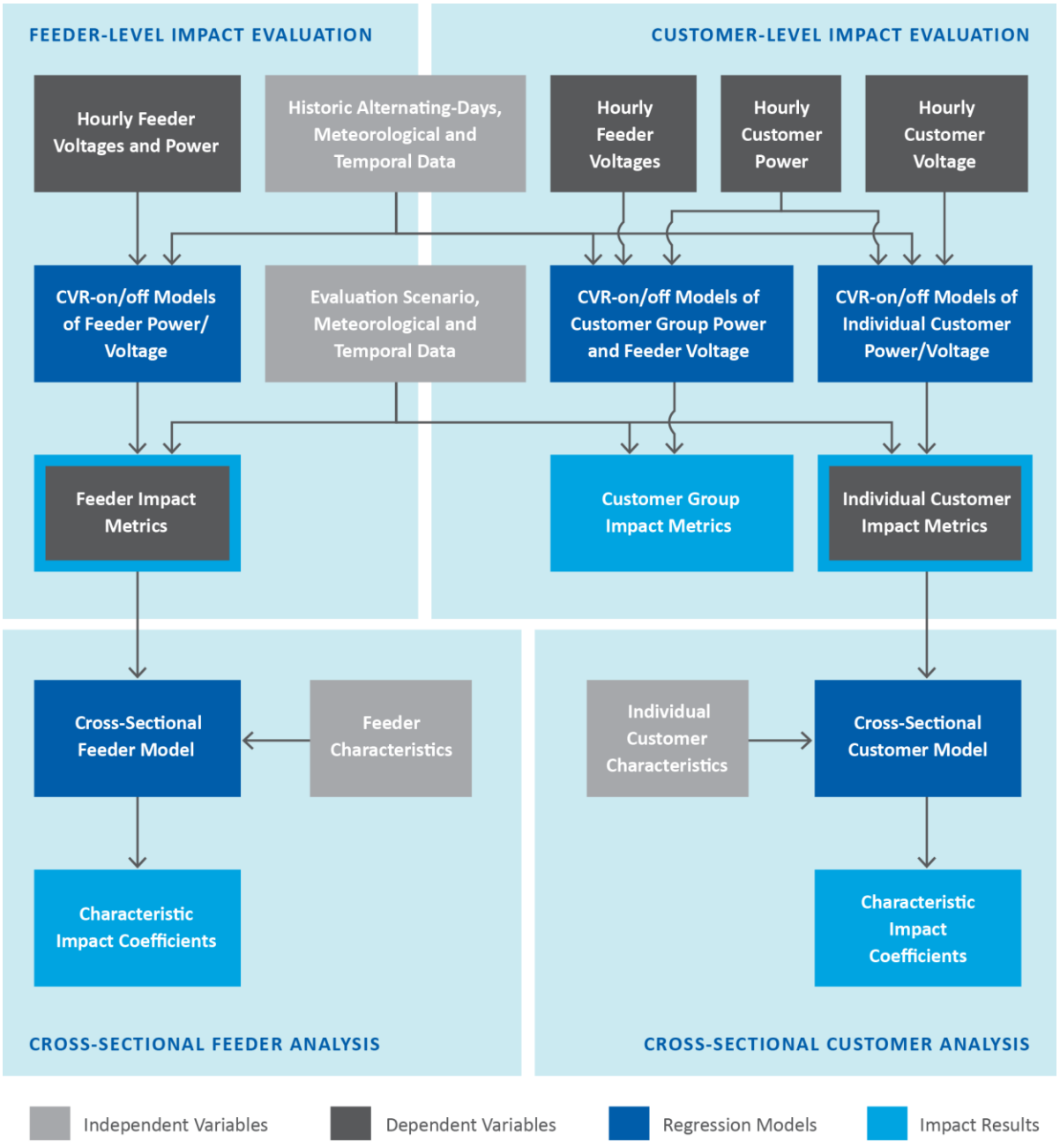
Cross-Sectional Coefficients: The weights of feeders or customer characteristics associated with the efficacy of CVR as quantified by CVR factors. The coefficients of a cross-section analysis can be used to predict the impact metrics for systems not included in the sample. For example, if cross-section analysis was conducted on a group of feeders using a set of standardized characteristics tracked by the LDC, then the cross-sectional coefficients could be used to predict the efficacy of CVR for other feeders with the same set of known characteristics.

2.2 Data Requirements and Defining Goals

Depending on data availability and the program and LDC objectives, different evaluation goals should be considered.

Figure 6 shows how data types are used in the different types of analyses and what the results of those analyses would be. A summary of these analysis types follows.

Figure 6. Flow from Data Sources to Statistical Model and Impact Metrics for All Types of Analysis*



*The results from feeder-level and customer-level impact evaluations are used as the dependent variables for cross-sectional analysis.

Feeder-Level Impact Evaluation: Every implementation of CVR should be accompanied by a feeder-level impact evaluation with the goal of determining impact metrics associated with the power measured at the substation. One set of models will be developed using data only from CVR-on days, while the other set will be fit to data from CVR-off days. If hourly voltage

measurements are not available along the feeder, then a measurement taken at the substation should be used. Both sets of models (i.e., CVR-on and CVR-off) will then be applied to the same typical meteorological year. The differences between these models are used to determine the CVR impact metrics, introduced in Section 2.1. These metrics are the definitive values for claiming impact after CVR begins operating continuously.

Customer-Level Impact Evaluation: If hourly usage is available for all customers, these values should be aggregated, and the CVR impact should be determined for the whole group. In this case, the same models used to estimate feeder-level voltage (as described above) should be used to determine average voltage reduction for the group. If hourly voltage measurements are available for individual customers, then there is an opportunity to determine CVR impacts for each customer, which is a requirement for customer-level cross-sectional analysis. Optionally, several large customers could be specifically modelled to determine the CVR impact on their usage.

Cross-Sectional with Customers and Feeders: The CVR factors can be compared among individual feeders or individual customers such that the characteristics known for all members of each group can be shown to be positive or negative indicators of performance. Every characteristic will be assigned a weight corresponding to its correlation with the CVR factors. These weighted coefficients can then be used to select feeders and customers that have promising characteristics for future CVR deployment. Therefore, the value of this analysis is highly dependent upon making the same set of characteristics used to determine the coefficients available more generally for other feeders or customers.

2.3 Customer Considerations

While the implementation of CVR likely would go undetected by most customers, it would be worthwhile to consider if any facilities served by the distribution feeders have critical loads with irregular voltage tolerances. Such facilities may be data centres, industrial processing and manufacturing plants and other buildings with high precision and electrically sensitive equipment. Generally, these facilities should already have systems in place to handle voltage sags within the allowable service range, but it would be good practice to let these facility managers know that CVR will be implemented. For example, some equipment vendors or technicians may set strict voltage bands on safety equipment in such a way that circuit breakers and fuses may trip near the lower voltage bound. Letting facility managers know that CVR is being implemented would give them important information for troubleshooting unintended consequences. To alleviate concerns, if the CVR control system is collecting voltage information from a set of *bellwether customer* interconnections via a utility's SCADA or other telemetry system, sensitive customers could have their voltages directly monitored and kept within allowed bounds.

2.4 Managing an Evaluation

Evaluation managers overseeing the CVR evaluation are responsible for determining which types of analyses should be conducted given organizational priorities, budgets and the availability of measurement devices. Program managers will need to work with the LDC, the CVR implementer and the contracted evaluator to make sure that necessary data are collected and transferred to the evaluator. Finally,

program managers are responsible for reviewing the final impact evaluation results and overseeing the publication of reports. This section offers some guidance for these activities.

2.4.1 Hire Qualified Evaluators

Any contractor hired to follow these protocols for evaluating CVR will need to be proficient in weather-normalized regression modelling and knowledgeable of power distribution systems at a sufficient level to process the data and produce impact metrics. For instance, the evaluators should be able to interpret electrical single-line diagrams from LDCs depicting the feeders on which CVR is being implemented and to associate shared measurements with specific equipment locations. Evaluators will need to be able to identify data anomalies, perform unit conversion (including from power factor to real and reactive components) and normalize voltage measurements (potentially line-to-line or line-to-neutral) for the system's various bases. Evaluators will need the capability to produce complex weather-normalized statistical regression models that are optimized through multi-round cross-validation and brute-force optimization using numerical analysis software for training and testing regression models fit to numerical and categorical features.

2.4.2 Set and Communicate Evaluation Goals

As outlined in Section 2.2, data requirements depend on the specific analyses to be performed. Options for setting evaluation goals will be limited by the available measurement devices and the frequency of data captured. At a minimum, the data requirements for conducting a feeder-level impact analysis must be met (hourly values for feeder voltages and power flows, and reported states of the CVR control system). The CVR-on/CVR-off schedule needs to be defined, while considering any scheduled maintenance that may compromise data that can be used to create the models. These on/off periods should be no shorter than three days and no longer then eight days.

It is recommended that at least three days be used as the minimum for most cases because it can take a few hours for voltages to transition under the differing control schemes. If several feeders are being evaluated and tracking information shows major differences in customer characteristics served by that feeder (e.g., some feeders serve mostly industrial customers, while others serve mostly residencies), then consider conducting a cross-sectional analysis across the feeder group. A qualified evaluator will find that only a minimal amount of additional effort is required to perform a cross-sectional analysis across feeders in addition to the feeder-level impacts if characteristic information on the feeders is available.

If hourly energy usage or average power measurements are available from customer meters, then it is highly recommended that all meters be aggregated as a group and that the total combined impact for all customers be determined. The same meteorological and temporal data should be used for the typical meteorological year. Use the same feeder voltage model that was used for the feeder-level evaluation to determine average voltage reduction and, in extension, CVR factors.

The benefits of conducting cross-sectional analysis across customers (e.g., metered building usage) would be the ability to estimate how CVR would impact customers on feeders without CVR. However, a significant level of

effort is required to perform a cross-sectional analysis across customers because impact metrics need to be determined for each customer individually. Hourly voltage needs to be measured at individual customer service lines to determine individual customer CVR factors. Additionally, customer characteristics must be available to be used as the independent variables.

As the data availability and evaluation goals are identified, it will be necessary to communicate the data requirements, the alternating-periods procedure and the ultimate impact metric types to the LDC, the implementer and the evaluator. All parties should agree to the evaluation plan, the predetermined on-day/off-day schedule and the timeline for transferring data and reporting results.

2.4.3 Coordinate Data Collection

As discussed above, data are collected from multiple locations. The evaluator should be responsible for collecting meteorological data and day-type data, but measurements from all the devices and the reported state of the CVR system (on/off/transition/monitoring) will need to be the shared responsibility of the implementer and the LDC. A full year of data is required to determine annual expected impacts. If all calendar months cannot be included in the evaluation, as defined in Section 3.1, then annual savings cannot be determined. Results for less than a year can be prepared for at least three consecutive months, but these results would not describe expected impacts for all months of the year. The implementer, the LDC and the evaluator should agree upon the schedule that makes the most sense, given the program goals. The program manager and LDC should confirm that the CVR implementer is following the agreed-upon CVR-on/off schedule. Given the complexity of data types and the number of sources, expect necessary back and forth as evaluators review and process the datasets.

2.4.4 Oversee Reporting

The remainder of this document covers in detail the requirements for reporting the impact metrics introduced in Section 2.1. In addition to these metrics, the normal mean bias error (NMBE) and the coefficient of variation of the root mean square error [CV(RMSE)] should be reported for the models used to estimate voltage and power flow for both CVR-on and CVR-off. These are industry standard metrics for reporting statistical model quality. If customers are modelled individually with the intent of performing a cross-sectional analysis, then the percentiles of NMBE and CV(RMSE) should be reported for each model type for all customers. Ideally, NMBE is under 0.5% and CV(RMSE) is under 5% for all feeder-level system state model of voltage and real power when evaluated over several thousand hours. Impact values, such as energy savings and peak demand reduction, should be clearly identified to be statistically significant and be accompanied by confidence intervals and relative precision values.

3. Protocols for Alternating-Periods Feeder-Level Impact Evaluation

This section presents protocols for evaluating the energy and demand impacts of implementing CVR using an alternating-periods analysis at the feeder level. Alternating periods means that the CVR system controller operates the CVR software to cycle between active (CVR-on) and inactive (CVR-off) days (defined in Section 3.1). This set of protocols is structured as the following four steps:

- Step 1. Collect historical and scenario data
- Step 2. Develop models to describe the voltage, reactive power and real power flows during the alternating CVR-on and CVR-off days
- Step 3. Derive impact metrics (energy savings, demand reduction, reactive power impacts and CVR factors) from the difference in voltage, reactive power and real power between CVR-on and CVR-off days for different scenarios
- Step 4. Report results

In general, with the addition of new data, all steps should be repeated from the beginning.

Plan to produce expected savings for running CVR continuously (minimum of three consecutive months up to one calendar year of continuous operation), which will require development of a dataset of the models' independent variables for a typical meteorological year. The model will also need to produce savings estimates during peak demand periods (as defined by the IESO and determined by local loading conditions) and any other weather scenarios of interest for demand reduction and reactive power impacts. The models need to be calibrated, using historical data from either CVR-on and CVR-off days, and will then be applied to a typical meteorological year. The difference in the estimates between CVR-on and CVR-off models is used to determine *expected savings*.

This feeder-level protocol relies on power measurements taken from substation monitoring equipment. The results provide aggregated impacts realized along the power distribution equipment and behind customer billing meters. If customer-level impacts are also going to be analysed, as discussed in Section 4, then it may be valuable to also conduct additional feeder-level impact analyses at other points along the feeder to isolate impacts to specific distribution zones. In this case, it will be necessary to collect additional power flow measurements from other monitors downstream and to determine appropriate customer reference voltages for these various zones.

3.1 Compile and Prepare Dataset

These protocols assume that the CVR system controller operates the software to cycle between active (CVR-on) and inactive (CVR off) over similar durations. These period in either the on or off state should be no shorter than three days and no greater than eight days. Each set of days, depending on the operation of the CVR system, are called "CVR-on days" or "CVR-off days" to denote whether the CVR system is engaged. Ideally, periods would alternate three days on and three days off following a predetermined schedule that was set before the evaluation data measurement begins. In this way, there is no concern that certain days were selected to demonstrate CVR under favourable circumstances. The alternating-periods method

allows the ability to determine the impact of CVR while controlling for prolonged events (e.g., seasonal weather events such as a heat wave, school vacations, scheduled downtime for industrial complexes).

A full year of data is necessary to determine annual expected savings, although impact metrics can be determined for a specific season if data are only available for the months of the corresponding season. Preferably, the evaluation dataset should span at least one year with all calendar months included. A full year of data requires that every calendar month have at least 37.5% of the hours included in the CVR-on set and 37.5% of the hours included in the CVR-off set, accounting for a coverage of at least 75% of hours each month. For instance, January has 744 hours, so for January to be included in the analysis then there must be at least 279 hours with CVR-on and 279 hours with CVR-off included in the dataset. If not all twelve months meet this criteria then impacts of CVR cannot be determined for the entire year, although impacts can be quantified for a subset of at least three consecutive months.

3.1.1 Collect Data

Data requirements should be communicated to—and tracked from—the various source providers, including an inventory of the data and when they were received. Table 1 below depicts the various data types, sources and use in the analysis. Measurements should be made in at least hourly increments or more frequently so that a processed dataset containing all the data sources for every hour can be prepared, as discussed in the section 3.1.2. These data needs should be communicated to the CVR control vendor and the local distribution company before the evaluation period begins, and data should be shared at regular intervals so that the evaluator can check and raise any concerns in advance, such as missing data or measurements at unexpected scales.

The evaluation will require a dataset of independent variables for the typical meteorological year and any other weather scenarios of interest. It is important to keep a copy of the original, unprocessed data used in the analysis with the goal that another evaluator could replicate the results if using the same initial dataset and following the same methodology.

3.1.2 Compile and Clean Data

The original data will need to be processed to ensure all independent and dependent variables are formatted consistently. This will require formatting date and time strings, converting values to different units and mapping metadata from dictionaries that can include equipment and measurement metadata. Store the processed data separately from the original data, and document the steps involved to transform the original data into the processed data.

Perform the following data-processing steps as necessary:

- Standardize set index to local time stamps
- Convert or encode values as numeric data types
- Resample the data at a specified interval (such as 15-minute, hourly or daily frequencies) to create consistent intervals at one hour
- Standardize labeling schemes

Table 1. CVR Evaluation Data Types, Sources and Uses for Feeder Level-Analysis

Data Type	Data Source	Use in Analysis
Total real and reactive power flow for feeders at substation interconnection	Substation SCADA of transformers or other monitoring systems at the head of the feeder	Dependent variables for fitting models
CVR-controlled equipment voltages	SCADA monitors of all CVR-controlled equipment (load tap changers at substation, along feeder voltage regulators and capacitor banks), power terminals from LDC or the CVR implementer	Dependent variables for fitting models
CVR state (e.g., on, off, transition, monitor) time stamps and control system state definition information	CVR software controller system and operational tracking logs	Split dataset on for on/off periods so that impact from the CVR system being "on" can be determined
Schedule of CVR system installation, commissioning and reconfiguration of hardware or settings	CVR implementer's project documentation	Identify the start time from when the CVR impacts are representative of ongoing performance
Time periods when external impacts led to compromised CVR activity or the CVR controller's schedule was modified	LDC's line/relay/equipment maintenance schedule, customer-blackout and voltage-issue records and CVR implementer's documentation of system errors caused by software or hardware malfunction	Identify periods when irregular circumstances and external impacts would not make the power flow and voltage data representative of CVR being either on or off
Voltages of customer meters	AMI measurements stored by LDC or CVR implementer	Inspect that substation voltage drops correspond in timing to customer voltage drops
Historical hourly weather of at least temperature and humidity	Nearest weather station with available data[1]	Independent variables for model fitting
Scenario hourly weather data (such as typical meteorological data or data from the hottest week from the last decade)	Canadian Weather Energy and Engineering Datasets (CWEEDS)[2]	Independent variables for running estimates during a typical meteorological year
Dates of local holidays and hour- and day-type definitions from the local distribution company	The distribution company and from Ontario's Ministry of Labour[3]	Prepare independent variables with additional temporal information

1. http://climate.weather.gc.ca/historical_data/search_historic_data_e.html
2. http://climate.weather.gc.ca/prods_servs/engineering_e.html
3. https://www.labour.gov.on.ca/english/es/tools/esworkbook/publicholiday.php

- Identify the need for (and limits of) interpolation for possible missing data
- Clean missing readings, such as dropping non-numerical data

After the data are compiled into a set, perform these tasks:

- Identify the starting date of usable data given the information about the installation and commissioning of the CVR equipment and control settings. Data collected before the CVR system configuration is finalized should not be used for modelling because any CVR activity before this starting data would not be representative of expected performance.
- Determine if critical events at the feeder level may reduce the total amount of historical data that should be included in the evaluation model (e.g., a power outage occurred for an extended period on a circuit, a voltage regulator or capacitor bank was replaced, a software update interrupted CVR system operation). A temporary change to the system may mean that only the days in question should be removed.
- Determine the total number of hours that the system operated in either a CVR-on or CVR-off state for each consecutive period. Remove any periods less than 24 hours in duration, because these shorter periods are not representative of expected behaviour of the system. Periods shorter than 24 hours may be caused by unexpected system failures or emergency maintenance.
- Identify data anomalies and outliers that are physically unreasonable, such as voltage values being far from the expected nominal values for the system or power flow dropping to exactly zero for several hours. Flag and remove these values.
- Establish criteria for dropping or filling in missing data, and plan to report and justify these decisions. Data should be dropped when there is evidence that a reading is unreliable, typically when values fall outside of known possible ranges for temperature, humidity, voltage and power flow. These ranges need to be determined on a case-by-case basis, but generally can be defined as values outside several standard deviations of the mean. Values outside of acceptable ranges can occur when sensors record error messages or are not positioned or calibrated correctly. Data should also be dropped if the same measurement is recorded for several hours in a row (for example, if the temperature remains the same for 36 hours in a row). This would be unrealistic, given the precision of the measurement and the characteristics of the system. Filling in missing data using interpolation is acceptable in some situations (such as a missing hourly reading of outdoor air temperature), but the limits of this interpolation need to be established in the context of the quantity being measured. Do not use interpolation to fill in data gaps greater than five hours.
- Produce histograms and violin plots of temperatures from CVR-on and CVR-off days by month to determine that

comparable ranges of measurements were taken for both cases. If not, then additional data should be collected to correct for this discrepancy.

Sum the measures of power flow by phase for each feeder separately to determine total real and reactive power values for the feeder, and then use these as the dependent variables for fitting power models. Average voltage for each feeder should be calculated as the average root-mean-square voltage for all phases to neutral for all buses of CVR-controlled equipment sharing the voltage base of the distribution line. This would include the low-side bus of the substation LTC and all line buses for applicable voltage regulators and capacitor banks. Use this average voltage as the dependent variable for fitting the voltage models. If hourly voltage measurements are only known for a bus located at the substation, use this bus as the feeder reference voltage, although this not preferred. Voltages vary significantly along a distribution feeder, so an average of multiple locations is more representative of a feeder reference voltage. In the very least, plan to take the mean of a measurement taken at the substation and another taken at the "end-of-line," at the end of the feeder.

With the historical dataset cleaned, split the processed data into CVR-on and CVR-off groups. Grouping data in this manner allows for building the statistical models to fit the appropriate CVR activity.

Next, prepare and save subsets of the typical meteorological year data that include only hours from the IESO-defined winter and summer peak periods, as shown in Table 2. These subsets will be used to prepare independent variables for modelling voltage, real power and reactive power during the defined seasonal system peaks. Additional definitions of peak periods can be developed and applied in a consistent manner to account for reductions in demand on distribution equipment. It is good practice to determine average power from CVR-on and CVR-off days during the all hours and during peak periods. These averages, which occur on the measured values and not on weather-normalized modeled results, should give an initial impression of how effective the CVR system was at reducing energy on the feeder's analyzed.

Table 2. The IESO EM&V Standard Definitions of Peak for Calculating Demand Savings[1]

Time		Month	Weight
Summer (weekdays)	1 p.m. to 7 p.m.[2]	June	30%
		July	39%
		August	31%
Winter (weekdays)	6 p.m. to 8 p.m.	December	65%
		January	16%
		February	19%

1. The defined summer and winter peak blocks for 2015–2020, based on analysis of Ontario System Hourly Load data from 2003–2010. Average peak reduction will need to be first averaged by month and then taken as a weighted average across months using the weights defined in this table.
2. Adjusted for Daylight Savings Time.

3.2 *Create System-State Models*

This evaluation protocol relies on statistical modelling to evaluate energy and demand savings resulting from the CVR implementation. The approach quantifies savings using processed datasets described earlier. The model predicts a feeder's voltage, reactive power and real power (dependent variables or state variables) from a set of temporal and meteorological independent variables. Every state variable for every feeder from each substation controlled for CVR will have two models: one for CVR-on days and one for CVR-off days. Fitting models to sets of historical records allows for predicting future results under different scenarios. This section describes some best practices for developing the models.

3.2.1 Specify the Regression Model

The models will define real power, reactive power and average distribution voltage for each distribution feeder as the dependent variables. These state variables will be modelled on an hourly basis and will require that the set of regression models (for real power, reactive power and voltage) are fit to data describing the weather and other temporal effects. Regression models are required as the system-state variables being modelled are numeric quantities and not categories. The selection of features used as independent variables is discussed the following section.

Table 3 describes the different types of regression models that can be considered for this analysis.

Some of the modelling types in the time series and machine learning classes may require a scripting language to be developed because they are not available in Excel. These nonlinear and discrete models should have superior prediction ability compared to linear models for hourly predictions. Multiple types should be tried under various conditions, and the best-performing model should be selected, as discussed in Section 3.2.3.

Table 3. Regression Model Types

Model Class	Model Type	Common Use Case
Linear	Single and multiple linear regression, ridge regression, Lasso regression	Low temporal resolution usage data, known physical relationships, observed linear trends
Time series	Autoregressive integrated moving average (ARIMA), error term models, transfer functions	High temporal periodicity and seasonality
Machine learning	Decision trees, random forests, neural networks, gaussian process	Nonlinear relationships, complex systems, large amounts of data

3.2.2 Produce Additional Independent Variables

Additional independent variables that can be created from independent variable data and which improve modeling performance should be added to the dataset. One example is converting the weather data into a computation of cooling degree days or heating degree days. These additional independent variables are also called "engineered features" and are helpful because they are repeatable and standardized transformations of processed datasets that can dramatically improve some models' ability to predict dependent variables of interest.

Because the models for the CVR evaluation are structured to predict average hourly values for voltage and power flows, each hourly observation point should include all information of interest for that hour, even if occurring at a different hour. For instance, the average 3 a.m. temperature may be valuable for predicting average power flows for 5 a.m. To associate dependent variables with independent variables measured at other than the hour being predicted, other data can be included as a panel or trailed, with additional columns to store preceding data on every row (as shown in Figure 7). In this way, additional information is engineered into the dataset, which provides more information for fitting the model. At a minimum, the following independent variables should be prepared for each hour:

- Twenty-four-hour trailing of weather (at least temperature and relative humidity) for every hour, or heat build-up variables (such as heating degree days and cooling degree days) for the last 12 hours
- Time-of-day tags for morning/afternoon/evening/night designations
- Holiday and other day-type (weekday/weekend) designations

Figure 7. Example of Trailing Weather to Build In Preceding Hours' Weather Data

	t-0	t-1	t-2	t-3	t-4	t-5
1:00	65	n/a	n/a	n/a	n/a	n/a
2:00	66	65	n/a	n/a	n/a	n/a
3:00	67	66	65	n/a	n/a	n/a
4:00	68	67	66	65	n/a	n/a
5:00	69	68	67	66	65	n/a
6:00	70	69	68	67	66	65
7:00	71	70	69	68	67	66
8:00	72	71	70	69	68	67
9:00	73	72	71	70	69	68
10:00	74	73	72	71	70	69
11:00	75	74	73	72	71	70
12:00	75	75	74	73	72	71
13:00	76	75	75	74	73	72
14:00	76	76	75	75	74	73
15:00	77	76	76	75	75	74
16:00	76	77	76	76	75	75
17:00	75	76	77	76	76	75
18:00	74	75	76	77	76	76
19:00	73	74	75	76	77	76
20:00	72	73	74	75	76	77
21:00	71	72	73	74	75	76
22:00	70	71	72	73	74	75
23:00	69	70	71	72	73	74
0:00	68	69	70	71	72	73

Additional information, if it leads to improvements in model quality, can be included such as the days until and days since a government or school holiday.

3.2.3 Optimize the Models

Optimizing the models requires both training (i.e., fitting the model to the dataset with a portion held-out) and testing (i.e., scoring the model on its ability to accurately predict the held-out data). Plan to perform this training and testing sequence multiple times with different subsets of the historical data. This process, called "cross-validation," measures prediction quality and is the most robust method to

determine that a model is fitting the data well. The analysis framework used should have cross-validation functionality built directly into the modelling toolkit. At the end of the cross-validation process, the validation scores for each held-out dataset will illustrate how well the model fits the dataset. Validation scores to use for evaluating regression models should include mean or median absolute error, mean or median squared error, root-mean-squared error and r-squared scores.

Randomly splitting the data into training and testing sets for cross-validation can introduce bias. For instance, a random split has a certain probability of holding out all the hottest days from the training set, which means that the trained model would not be well prepared to predict for those conditions. To fully understand model quality, this splitting process should be repeated many times—typically hundreds or thousands of times—for each model. These simulations build distributions of validation scores for each model to inform the selection of final models.

When splitting the data for training and testing sets, it is recommended that whole days and whole weeks be grouped together to limit predictions based on autocorrelation between two similar hours occurring in sequence. This would be required for modelling techniques that use continuous timeseries data for fitting and prediction. Optimizing a model in this manner prevents *overfitting* the model to the training dataset. An overfit model would not be able to generalize well to new data and would not be well suited for making estimates of system states during a typical meteorological year.

With the ability to check model fit via multi-round cross-validation, it is now possible to optimize the models by scoring many different modelling settings. The entire modelling process can be iterated to determine the optimal set given the scoring types listed above. Use a platform with standard tools for performing model optimization given the desired scoring criteria, such as r-squared scores. Always start with a simple model with a small set of independent variables, such as ordinary least squares (OLS) using each hour's corresponding temperature and humidity. Additional model and independent variable complexity should only be justified if they improve the quality of the cross-validation scores.

Finally, when the model is optimized, determine the normal mean bias error (NMBE) and the coefficient of variation of root mean square error [CV(RMSE)]. NMBE would reflect cumulative errors introduced by the model during the typical meteorological year. In contrast, CV(RMSE) would quantify the spread of errors from individual predictions that may cancel out in a cumulative metric. These values are defined as follows:

$$NMBE = \frac{\frac{1}{N}\sum_i^N (y_i - \hat{y}_i)}{\bar{y}} \times 100 \quad (1)$$

$$CV(RMSE) = \frac{\sqrt{\frac{1}{N}\sum_i^N (y_i - \hat{y}_i)^2}}{\bar{y}} \times 100 \quad (2)$$

Where y_i is the actual measured quantity, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of all y_i and N is the total number hours. For this step, do not use a subset of data, as in the case of cross-validation, but rather use the complete available set of all applicable measurements. NMBE and CV(RMSE) metrics should be reported for every system-state model to convey a standardized measure of the model performance.

3.3 Determine Feeder-Level Impacts

With models that have been validated and optimized, the next step is to produce and compare predictions for CVR-on and CVR-off (baseline) scenarios.

3.3.1 Finalize Evaluation Metrics

The following evaluation metrics, as introduced in Section 2, are the key metrics for determining the impacts of implementing CVR.

Average Voltage Reduction: The average difference between the feeder-level voltage from the CVR-off models and CVR-on models is $\Delta\bar{V}$:

$$\Delta\bar{V} = \bar{V}_{off} - \bar{V}_{On} = \frac{\sum_i^N V_{off,i}}{N} - \frac{\sum_i^N V_{on,i}}{N} \quad (3)$$

Where N is number of hours in the typical meteorological year and $V_{on,i}$ and $V_{off,i}$ are customer reference voltages during hour i for the cases CVR is either on or off. The relative percentage difference in voltage would be $\%\Delta\bar{V}$:

$$\%\Delta\bar{V} = \frac{\bar{V}_{off} - \bar{V}_{On}}{\bar{V}_{off}} * 100 \quad (4)$$

Voltage reduction should be determined on average during the typical meteorological year and during the IESO peak periods.

Total Energy Savings: The quantity of energy that the CVR-off model predicts would have been used beyond what was predicted for the CVR-on case is ΔE:

$$\Delta E = E_{off} - E_{on} = \sum_i^N (P_{off,i} * 1hour) - \sum_i^N (P_{on,i} * 1hour) \quad (5)$$

Where N is the number of hours in the typical meteorological year and $P_{on,i}$ and $P_{off,i}$ are the average hourly real power flow for hour i for the cases CVR is either on or off. Energy savings should be determined for the typical meteorological.

Average Demand Reduction: The average difference in real demand between the CVR-on and CVR-off models is $\Delta\bar{P}$:

$$\Delta\bar{P} = \bar{P}_{off} - \bar{P}_{On} = \frac{\sum_i^N P_{off,i}}{N} - \frac{\sum_i^N P_{on,i}}{N} \quad (6)$$

Where N is the number of hours in demand reduction period of interest and $P_{on,i}$ and $P_{off,i}$ are the average hourly real power flow for hour i. Demand reduction should be determined on average during the typical meteorological year (i.e. 8760 hours), during the IESO peak periods, and during any other peak period definitions as required.

Reactive Power Impacts: The average difference in reactive power between the CVR-off and CVR-on models is $\Delta\bar{Q}$:

$$\Delta\bar{Q} = \bar{Q}_{off} - \bar{Q}_{On} = \frac{\sum_i^N Q_{off,i}}{N} - \frac{\sum_i^N Q_{on,i}}{N} \quad (7)$$

Where N is the number of hours in the typical meteorological year and $Q_{on,i}$ and $Q_{off,i}$ are the average hourly reactive power flow for hour i for the cases CVR is either on or off. Reactive power reduction should be determined on average during the typical meteorological year and during the IESO peak periods.

CVR Factors: CVR factors should be calculated for the various periods of interest by normalizing the modelled relative reductions in energy, demand and reactive power by the corresponding relative reductions in voltage. This provides a standard method to determine how much impact decreasing the voltage of the feeder has on other quantities of interest. The CVR factor of variable X (which can be energy savings, demand reduction or reactive power

impacts) for a feeder during a specified typical meteorological year would be $f_{CVR,X}$:

$$f_{CVR,X} = \frac{\%\Delta X}{\%\Delta \bar{V}} \quad (8)$$

Where $\%\Delta X$ is the relative difference of the quantity of interest and $\%\Delta \bar{V}$ is the relative difference of average feeder reference voltage.

CVR factors for energy savings should be determined for a typical meteorological year. CVR factors for demand and reactive impacts should be determined as the average value during the peak periods, as defined in Table 2 of Section 3.1.2, and for any other weather scenarios prepared for the analysis.

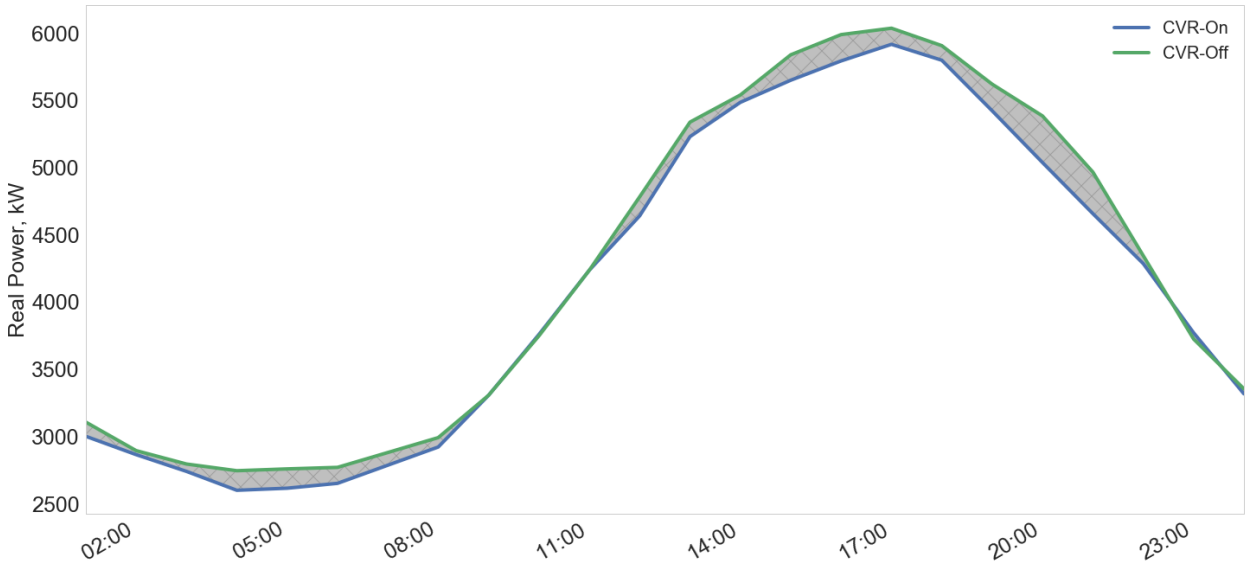
3.3.2 Predict Hourly System State for CVR-On and CVR-Off Cases

Every feeder should have six optimized models to predict power flows (real and reactive) and customer reference voltages during CVR-on and CVR-off days. With these models fitted to the complete sets of historical CVR-on and CVR-off data, the models can be used to predict power flows and voltages during the typical meteorological year.

3.3.3 Compute Impacts from Predictions

Using the model predictions, expected impact calculations can be performed, as outlined in Section 3.3.1. The power flow and voltage predictions for the CVR-on and CVR-off scenarios should rely on data with the same time stamps for computing differences. With the voltage reduction, energy savings, demand reduction and reactive power impacts determined for the various periods of interest, corresponding CVR factors can be calculated for each evaluated feeder.

Figure 8. Example of Determining Energy Savings as the Difference Between Usage from CVR-On and CVR-Off. The Greyed Area Depicts the of Energy Savings Over a 24-hour Period.



3.3.4 Determine Precision of Results

Three sources of uncertainty need to be accounted for when determining the precision of the impact results: measures of model quality for both the CVR-on and CVR-off conditions and a measure of variance in the difference between the models when they are used to infer usage during a typical meteorological year. These three terms should be combined into an aggregated standard error as follows:

$$SE = \sqrt{{SE_{diff}}^2 + {SE_{on\_model}}^2 + {SE_{off\_model}}^2} = \sqrt{\frac{{\sigma_{diff}}^2}{n_{diff}} + \frac{MSE_{on}}{n_{on}} + \frac{MSE_{off}}{n_{off}}}$$

SE_{diff}: Standard error for the difference for hourly impacts during evaluation scenario

$SE_{on\_model}, SE_{off\_model}$: Standard errors for the prediction of the CVR on/off model

σ_{diff}: Standard deviation of the difference between the models across all periods

MSE_{on}, MSE_{off}: Mean Square Error, a quantification of high variance in the model

$n_{diff,}\ n_{on,}\ n_{off}$: The number of samples in the evaluation scenario or model fit

The aggregated standard error, SE, accounts for not only the quality of the hourly models but also the extent to which the CVR impacts are consistent. The confidence interval for average impacts can be determined using a two-tailed z score at a specified alpha level. The results can be tested for statistical significance at 90% confidence by using a z score of 1.645 and determining the confidence interval. The upper and lower bounds of the confidence interval can be computed using the following equation:

$$\bar{x} \pm z * SE$$

SE: aggregated standard error for the impact metric as defined above

z: two-tailed z score determined as a function of desired confidence

$\bar{x}$: mean of the distribution for the impact metric across all hours

If the bounds of this interval crosses zero, then the CVR impacts cannot be considered statistically significant. The relative precision of the impact as a percentage can be determined using the same terms in following equation:

$$\frac{z * SE}{\bar{x}} * 100$$

3.4 Report Results

The evaluation report should present findings and provide context for the results. The report should include a narrative about how the CVR is implemented, a methods section covering evaluation assumptions and graphics to illustrate the different variables evaluated and the impact metrics including confidence intervals and relative precision as described in Section 3.3.4. The report should also reference these protocols to provide the reader with additional resources for understanding how the evaluation methods were informed.

3.4.1 Describe the CVR Program or Pilot

Summarize how the CVR program was implemented and include a description of the scale of the distribution system under CVR control. This could be counts and descriptions of

feeders, transformers, customers or other geographic metrics defining the scale of the operation. Name the CVR software vendor and any other products or companies involved with the ongoing operation of the system. List the dates of major milestones for the CVR system's testing and operation.

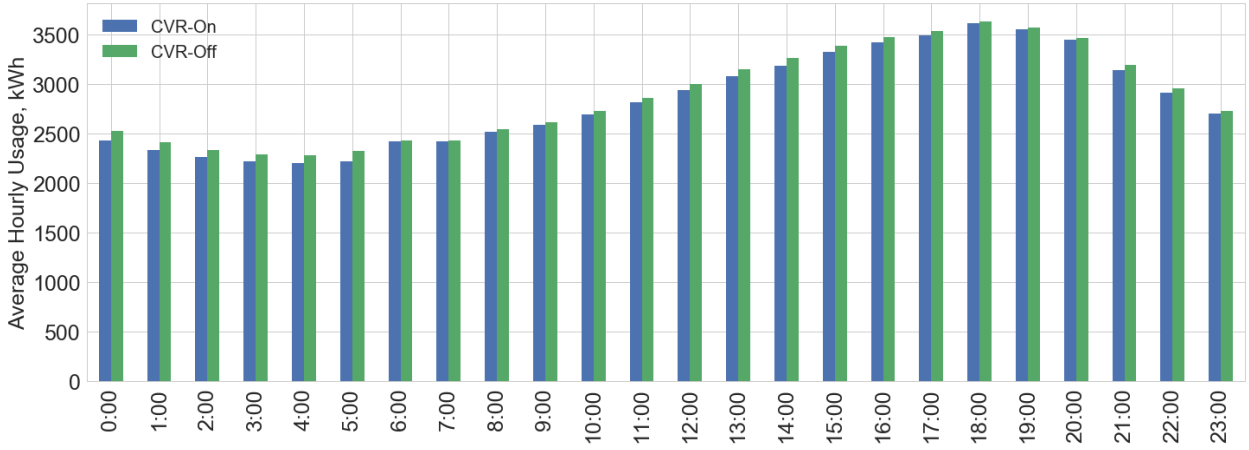
3.4.2 Describe Evaluation Process and Assumptions

Outline the series of steps taken to compile and process the dataset, fit and optimize the system-state models and calculate the evaluation metrics. Declare the data sources and the time periods of all historical data used. Specify how any weather scenarios beyond the standard IESO peak definitions were constructed or sourced. List filters used to cleanse the data, and remark on the quantity of data that needs to be dropped or modified because of quality concerns. State the regression model type used, identify the process used to produce additional independent variables and describe how the prediction quality of the models was validated and optimized.

3.4.3 Share Plots of Historical, Scenario and Predicted Data

Use graphics to reinforce the conclusions that will be drawn from the tabulated impact metrics. Identify the sources of all plotted data and make sure the reader understands how the presented data fit into the overall evaluation process. Historical data should be plotted first and should include power flow, voltage and weather data from actual measurements taken during the evaluation period. Scenario data would be shown next and include typical meteorological year data for expected energy impacts, subsets for the peak period definitions and other weather scenarios for expected demand impacts. Finally, plots for predicted power flows and voltages should be shown. It would also be appropriate to produce illustrative plots of underlying evaluation metrics, such as expected cumulative energy usage during a typical meteorological year or hourly voltage profiles by season from CVR-on and CVR-off cases. Figure 9 below shows an example of average hourly energy usage for when CVR is either on or off.

Figure 9. Example of Average Hourly Energy Usage Over a Year for CVR On vs Off



3.4.4 Present Expected Savings by Feeder

Prepare tables to present evaluation metrics for all the feeders in the study and, optionally, by substation, if multiple substations are under consideration. Evaluation metrics, as defined in Section 3.3.1, should be associated with the typical meteorological year. Report average voltage reduction, either total energy savings or average demand savings and CVR factors as applicable for each feeder. Additionally, NMBE and CV(RMSE) should be reported for every system-state model to convey a standardized measure of the model performance. Refer to Section 3.2.3 for these definitions. To present modelling fit to out-of-sample predictions, share percentiles of r-squared scores from cross-validating the optimized models as well as confidence intervals and relative precision as defined in Section 3.3.4. Summarize how this cross-validation was executed, including the number of rounds and the sizing criteria for the held-out sample.

4. Protocols for Alternating-Periods Customer-Level Impact Evaluations

This section presents protocols for evaluating the energy and demand impacts of implementing CVR using an alternating-periods analysis at the customer level. In contrast to a feeder-level analysis in which the power, energy and impacts are determined at the substation, a customer-level analysis determines how CVR affects usage as measured by the customer billing meters. An impact evaluation at the customer level should be done in concert with an impact evaluation at the feeder level, and it has similar data and analysis requirements. Read Section 3 first, and then return to this section, which expands upon the workflow and recommendations discussed at the feeder level. Many of the steps are similar for both protocols, and some text is repeated to provide sufficient context.

Alternating periods means that the CVR system controller operates the CVR software to cycle between active (CVR-on) and inactive (CVR-off) days. This set of protocols is structured as the following five steps:

- Step 1. Collect historical and typical-year data; sum up customer hourly energy usage for a group analysis and prepare individual customer usage on a daily basis.
- Step 2. Develop models to describe the voltage, reactive power and real power flows during the alternating CVR-on and CVR-off days
- Step 3. Derive impact metrics (energy savings, demand reduction, reactive power impacts and CVR factors) from the difference in voltage, reactive power and real power between CVR-on and CVR-off days for different scenarios at the customer level
- Step 4. Report results

Plan to produce annual expected energy savings, peak demand impacts and CVR factors for all customers as a group. Modelling every customer separately should only be done on a daily basis given the high variance in hourly usage. Models are developed from historical data from both CVR-on and CVR-off days. These models will then be applied to the typical meteorological year. Savings based on meteorological scenarios for a typical year and during defined IESO peak periods are called *expected savings*.

The customer-level impact results may be paired with results from a feeder-level impact assessment to isolate CVR impacts in front of the meter versus behind the meter. As desired, the feeder-level impacts can also be determined at nodes downstream from the substation. This allows for loading impacts to be determined for specific distribution equipment and line segments.

4.1 Compile and Prepare Dataset

If feasible, conduct data preparation efforts for a customer-level analysis at the same time as feeder-level analyses for the customers' circuits. This is most applicable to the time stamp event data from the CVR controller system, the feeder-level voltage and the independent variables defining the weather and other temporal features.

4.1.1 Collect Data

Data will be required from multiple sources, and these data may include sensitive information

about customers (e.g., names, addresses, account numbers and metering IDs). Data requirements should be communicated to—and tracked from—the various sources, including an inventory of the data and when they were received. Table 4 below depicts the additional data needed beyond those already specified in Table 1 of Section 3.1.1. Measurements should be collected in at least hourly increments or more frequently so that a processed dataset containing all the data sources for every hour can be prepared, as discussed in section 4.1.2. If hourly reactive power is available for large customers, these data can also be requested and used, but it is assumed that these measurements will not be available in most cases.

Table 4. Additional CVR Evaluation Data Types, Source and Use for Customer Level-Analysis

Data Type	Data Source	Use in Analysis
Total hourly energy usage for each customer	Billing advanced metering infrastructure (AMI) meters	Dependent variables for fitting models
Customer connection voltage	Billing AMI meters or SCADA of step-down transformers	Determine customer-specific CVR factors

4.1.2 Compile and Clean Data

The original data will need to be processed to ensure that all independent and dependent variables are formatted consistently. This may require formatting date and time strings, converting values to different units and mapping metadata from dictionaries that can include equipment and measurement metadata. Store the processed data separately from the original data, and document the steps involved to transform the original data into the processed data. Plan to perform the same data processing steps as outlined in Section 3.1.2. Ideally, if this customer-level analysis is done in tandem with the feeder-level analysis, then all the data describing the average feeder voltage and the independent variables should already be available for use.

Given the large number of customers who are likely on each feeder, it is recommended to group customers and model their aggregated energy consumption. This could reduce, from thousands to dozens, the number of models that need to be created. Aggregate all customers by feeder unless individual customer voltages are available and there are plans to conduct cross-sectional analysis of customers, as outlined in Section 5. Alternatively, if planning to develop daily energy usage models for every customer, then aggregate the hourly energy usage to daily values. If voltages are available for each customer, then customer-specific CVR factors can be calculated. In this case, a cross-sectional analysis of customers could be performed.

With the historical set cleaned, split the processed data into CVR-on and CVR-off groups. Grouping data in this manner allows for building the statistical models to fit the appropriate CVR activity.

Next, prepare subsets of the typical meteorological year data that include only hours from the IESO-defined winter and summer peak

periods, as specified in Table 2 of Section 3.1.2. These subsets will be used to prepare independent variables for modelling voltage and real power during the defined seasonal system peaks. Next, prepare additional subsets of data for additional weather scenarios. Refer to Section 3.1.2 for additional guidance.

4.2 *Create System-State Models*

This evaluation protocol relies on statistical modelling to evaluate energy and demand savings resulting from the CVR implementation. Follow instructions as specified in Section 3.2.

4.2.1 Select an Appropriate Regression Model

The models will define real power (and, optionally, customer voltage) as the dependent variables for each customer or customer group. These state variables will be modelled on an hourly or daily basis and will require that the set of regression models (for energy consumption of customers) are fit to data describing the weather and other temporal effects. Refer to Section 3.2.1 for further guidance on regression model selection.

4.2.2 Produce Additional Independent Variables

Additional independent variables that can be created from independent variable data should be added to the dataset. One example is converting the weather data into a computation of cooling degree days or heating degree days. These additional independent variables are also called "engineered features" and are helpful because they are repeatable and standardized transformations of processed datasets that can dramatically improve some models' ability to predict dependent variables of interest. Depending on data available and the type of regression model chosen, additional engineered features improve the models' accuracy for predicting voltage and power flow. Refer to Section 3.2.2 for further guidance on producing additional independent variables. The complete set of independent variables prepared for a feeder-level analysis (i.e., weather and temporal features) can also be used for the customer-level analysis.

4.2.3 Optimize the Models

Optimizing the models requires both training and testing. Train (i.e., fit) the model to a subset of the historical dataset and then test the model by seeing how well it predicts the held-out historical data (i.e., the subset of data not used to train the model). Plan to perform this training and testing sequence multiple times with different subsets of the historical data. This process, called "cross-validation," measures prediction quality and is the most robust method to determine that a model is fitting the data well. The analysis framework used should have cross-validation functionality built directly into the modelling toolkit. At the end of the cross-validation process, the validation scores for each held-out dataset will illustrate how well the model fits the dataset. Refer to Section 3.2.3 for guidance on model optimization.

4.3 *Determine Customer-Level Savings*

4.3.1 Finalize Evaluation Metrics

The following evaluation metrics, as discussed in Section 2, are the key metrics for determining the impacts of implementing CVR. For complete definitions and equations of the following impact metrics, refer to Section 3.3.1.

Average Voltage Reduction: The average difference between voltages from the CVR-off

models and CVR-on models. This can either be associated with the average voltage of the feeder (which should be used for customer groups) or with the individual customer voltage. Average voltage reduction should be calculated for both the whole typical meteorological year and for the specific IESO peak periods.

Total Energy Savings: The quantity of customer energy usage that the CVR-off model predicts would have been used beyond what was predicted for the CVR-on case for a typical meteorological year. To normalize across all customer, energy savings should be determined as a relative rather than absolute value.

Average Demand Reduction: The average difference in real demand between the CVR-on and CVR-off models during the IESO peak periods if modeling usage on an hourly basis. To normalize across all customers, demand reduction should be determined as a relative rather than absolute value.

CVR Factors: CVR factors should be calculated by normalizing the modelled relative reductions in energy demand and by the corresponding relative reductions in voltage. If the power flows of the customers are modelled individually, then CVR factors can be reported for every customer and the $\%\Delta\bar{V}$ should correspond to the modelled voltages for the customer's interconnection. CVR factors for energy savings should be determined for a typical meteorological year. CVR factors for demand should be determined as the average value for the typical year and during the peak periods, as defined in Table 2 of Section 3.1.2.

4.3.2 Predict System State for CVR-On and CVR-Off Cases

Every customer or customer group should have one optimized model to predict power flows during CVR-on and CVR-off days. With these models fitted to the complete sets of historical CVR-on and CVR-off data, the models can be used to predict power flows during the evaluation periods of interest (typical meteorological year, prepared peak periods, any additional weather scenarios). If modelling customers individually for a cross-sectional analysis, also predict hourly voltages for each customer.

4.3.3 Compute Impacts from Predictions

Using the model predictions, expected impact calculations can be performed, as outlined in Section 3.3.1. The power flow of customers and feeder-level voltage predictions for the CVR-on and CVR-off scenarios should rely on data with the same time stamps for computing differences. With the feeder-level voltage reduction, and customer energy savings and demand reduction impacts determined for the various periods of interest, corresponding CVR factors can be calculated.

4.3.4 Determine Precision of Results

The standard error for the average relative impacts among each group customers should be calculated using the following formula:

$$SE = \sqrt{\frac{\sigma^2}{n_{sites}}}$$

SE: standard error for the average impacts on the group of customers

σ^2: variance among customers' relative reduction values

n_{sites}: number sites included in the treatment group

The confidence interval for average impacts can be determined using a two-tailed z score at a specified alpha level. Determine statistical significance, confidence intervals and relative precision using the equations defined above in section 3.3.4

4.4 Report Results

An evaluation report of customer-level impacts should be presented along with the accompanying feeder-level impacts. Follow the reporting instructions specified in Section 3.4. Results for customer groups or isolated distribution equipment can be presented in the same fashion as feeder-level results, except that reactive power and demand impacts may not be available. If conducting individual customer impact evaluations, then it may not be possible to present modelling fits for every single customer as this level of granular detail is too exhaustive for a impact report. In such cases, report distribution statistics of modelling fits. Present the average impacts determined for each customers on each feeder and whether results were statistically significant along with confidence intervals as described in Section 4.3.4. A complete set of impacts for every customer can be shared in a workbook.

5. Protocols for Cross-Sectional Analysis with Feeders and Customers

The preceding protocols define the steps necessary to determine CVR impacts for specific feeders and customers. If multiple feeders have been included in the analysis, or if customers were modelled individually, there is an additional opportunity to quantify trends across the group to see if certain characteristics are strongly correlated with CVR factors. The results from this cross-sectional analysis could be used to estimate CVR impacts on other systems for which the same characteristics are also known. In this way, the results from piloted CVR implementations can be used to inform where the next best opportunities for CVR may be. The quality of these correlations will be highly dependent on which characteristics are available to describe the feeders and the customers.

5.1 Compile and Prepare Dataset

Sometimes called "second-stage" analyses, cross-sectional analyses build upon the results of previous analyses that have already determined impacts at the feeder or customer level. Therefore, it is assumed that protocols as outlined in sections 3 and 4 have already been followed, and that CVR factors for feeders and individual customers have been calculated. These CVR factors can be used from different pilots, programs, utilities and geographic regions, but the method for determining CVR factors should be the alternating-periods method across all cases.

5.1.1 Collect Data

The impact metrics for all feeders or customers must be compiled. Additional characteristics that will be used as features to describe these systems then need be requested and compiled. Table 5 provides examples of features that may be available for this type of analysis. Compile characteristic independent variables for a cross-sectional regression across all systems in the group.

5.1.2 Compile and Clean Data

Every feeder or customer will need a unique row in a data table that includes the impact results and all descriptive features. Numeric features can be included with no modification. Categorical features will need to be turned into several columns of true/false values (represented by 1 and 0, respectively) that represent whether that instance is part of the corresponding categories. These additional variables are sometimes called "dummy variables." In this way, categorical data can be included in the cross-sectional regression.

Methods employed to determine CVR factors need to be consistent for all instances of feeders and all instances of customers, as defined by the protocols in the previous sections. Only combine CVR factors produced from evaluations following the alternating-periods method in a cross-sectional analysis. Remove from the compiled dataset all instances that do not meet this criterion.

Table 5. Examples of Features that May be Available for Cross-Sectional Analysis

Numerical Features[1]	
Feeder	**Customer**
• *% Facility types* • *% of measure or program participation* • *Total square footage of all customers* • *Average annual load* • *Voltage base of feeder* • *Total length of feeder* • *Heating and cooling degree days of typical meteorological year (if different scenarios were used across the set of feeders)* • Available demographic information	• *Average annual consumption* • *Seasonal peak demands* • *Square footage of building(s)* • *Heating and cooling degree days of typical meteorological year (if different scenarios were used across the set of feeders)* • Number of residents • Number of workers • Capacity of distributed generation or energy storage
Categorical Features[1]	
Feeder	**Customer**
• *CVR control vendor* • *Local distribution company*	• *CVR control vendor* • *Facility type* • Historical participation in energy efficiency or demand management program

1. Essential features to include in this analysis have been italicized.

5.2 Correlate Features with Impacts

With all the data prepared, it is now possible to fit a multivariable linear regression to estimate CVR impacts based on either feeder or customer characteristics. Figure 10 below presents the complete equation showing how the compiled dataset will be used to determine optimal coefficients ($\boldsymbol{\beta_q}, \boldsymbol{\gamma_k}$) by minimizing the regression residual ($\boldsymbol{\varepsilon_j}$) for all systems in the sample. Coefficients should be optimized using ordinary least squares (OLS) methodology. The determined coefficients as a set represent the cross-sectional weights given to each characteristic representing how it would predict for CVR factors when fitting the regression. The coefficients can be used in Equation 11 (shown in Figure 10) to predict the impact metric for an out-of-sample system (without the residual error term that cannot be determined). For example, if annual energy usage, building square footage and facility type were used as features to determine cross-sectional coefficients for peak demand reduction CVR factors, then the peak demand reduction CVR factor could be estimated using those fit coefficients for a facility that has never received CVR if the annual energy usage, building square footage and facility type were known.

Figure 10. Linear Regression for Cross-Sectional Analysis

$$Y_j = \sum_q \beta_q x_{qj} + \sum_k \gamma_k I_{kj} + \varepsilon_j \quad (11)$$

j = identifier for each system (feeder/customer) in cross-sectional analysis

q = identifier for numeric characteristic

k = identifier for binary categorical characteristic

$\boldsymbol{Y_j}$ = CVR factor (for energy savings or real/reactive power reduction) for system j

$\boldsymbol{I_{kj}}$ = 0/1 hot-encoded dummy variable. Equal to 1 if for system j the characteristic k is true. Equal to 0 if for system j the characteristic k is false.

$\boldsymbol{x_{qj}}$ = value of the numerical characteristic q. Let x_{0j}, the first term of this vector, equal 1 for all premises so that β_o serves as an intercept term

$\boldsymbol{\beta_q, \gamma_k}$ = coefficients determined by the regression

$\boldsymbol{\varepsilon_j}$ = regression residual for system j

5.3 Report Results

The results from a cross-sectional analysis should either accompany a report for a specific CVR implementation or should reference the reports from where results are being used. If using CVR factors prepared by other evaluators, then provide context and references for those results. This would include a narrative about how the CVR was implemented, a summary of methods covering evaluation assumptions and the final impact metrics. Finally, reference these protocols to provide the reader with additional resources for understanding how the cross-sectional method was informed.

5.3.1 Describe Process and Assumptions

Outline the series of steps taken to compile the dataset and conduct the linear regression. Declare the data sources and produce tables summarizing the features used to characterize the feeders or customers. List filters used to cleanse the data, and remark on the prediction quality of the cross-sectional model. State any methods employed to optimize variables used for the regression.

Additionally, NMBE and CV(RMSE) should be reported for the cross-sectional models to convey a standardized measure of the model performance. Refer to Section 3.2.3 for these definitions.

5.3.2 Share Plots and Tables of Feature Data

Use tables and graphics to communicate the diversity and distribution of feature values and characteristics among the population of systems being analysed. This should include scatter plots and histograms. Also include histograms of

expected impact metrics being used to fit the regression.

5.3.3 Present Cross-Sectional Coefficients

Prepare tables to present the cross-sectional coefficients determined through the OLS linear regression for each expected impact metric. Be sure to make it clear when coefficients are positive or negative. Coefficients for numerical characteristics should include the corresponding units (e.g., kWh/square foot), while categorical coefficients for dummy variables should have units of the impact metric (e.g., kWh). Comment on why certain variables may be strongly correlated with CVR factors and why they may be having a positive or negative impact. Offer suggestions on how these coefficients can be used by utility system planners to identify other feeders that would be good candidates for CVR.

Appendix A. Literature Review

Relevant literature covering the theoretical potential of CVR implementation across North America:

- D. Pinney, "Costs and Benefits of Conservation Voltage Reduction," U.S. DOE/NETL, Arlington, Virginia, 2014.
- K. Schneider, "Evaluation of Conservation Voltage Reduction (CVR) on a National Level," U.S. Department of Energy, Oak Ridge, TN, 2010.
- R. Singh and F. Tuffner, "Effects of distributed energy resources on conservation voltage reduction (CVR)," IEEE, Detroit, MI, USA, 2011.
- R. Scalley and D. Kasten, "The Effects of Distribution Voltage Reduction on Power and Energy Consumption," *IEEE Transactions on Education* , vol. 24, no. 3, pp. 210-216, 1981.

Relevant literature covering the evaluated impact of CVR implementation across North America.

- M. Diaz-Aguiló, "Field-Validated Load Model for the Analysis of CVR in Distribution Secondary Networks: Energy Conservation," *IEEE Transactions on Power Delivery*, vol. 28, no. 4, pp. 2428-2436, 2013.
- R. Griffith, "Voltage and Reactive Power Optimization (Smart Grid Pilot Project, Final Report)," PG&E, 2016.
- K. Cooney, D. Greenberg, F. Stern and P. Higgins, "Avista Utilities' Conservation Voltage Reduction Program Impact Evaluation," Avista Utilities, 2014.
- NREL, "Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol," in *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, 2013, pp. 4.7.12, Page 8-19.
- M. Messenger, "Verification of the Energy Saving Impacts of the Potomac Edison CVR Program," Itron, Davis, CA, 2014.
- Regional Techincal Forum, "Automated CVR Protocol No. 1: Procotol Document v1.1," 2004
- S. Lefebvre and G. Gaba, "Measuring the Efficiency of Voltage Reduction at Hydro-Quebec Distribution," *IEEE Power and Energy Society General Meeting*, no. Conversion and Delivery of Electrical Energy in the 21st Century, 2008.
- US EPA "Conservation Voltage Reduction/Volt VAR Optimization EM&V Practices," 2017