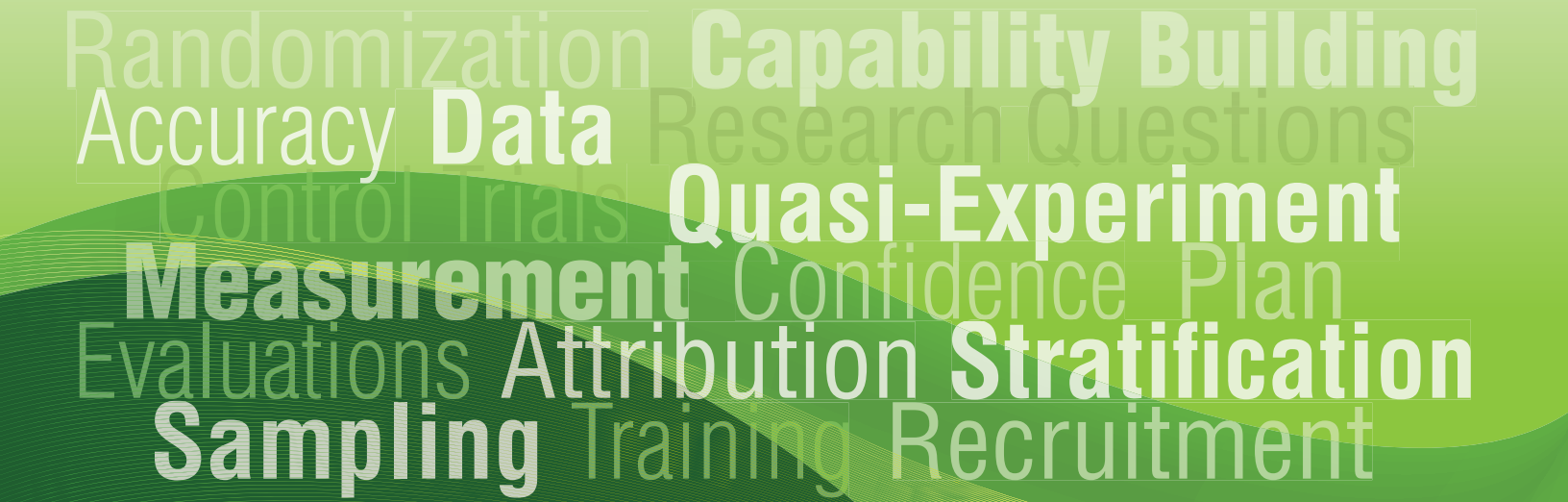
ONTARIO POWER AUTHORITY

Protocols for Evaluating Behavioral Programs



Acknowledgments

The Ontario Power Authority 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 Ontario Power Authority (OPA) 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 (e.g. Opower, Simple Energy);
- feedback technologies that allow consumers to observe their energy use at websites or from devices installed in their homes (e.g., Blueline in-home-displays, GE Nucleus, Rainbow, etc.);
- home automation technologies to consumers that help them consume less energy (e.g., Nest thermostat);
- time varying rates that help consumers lower their energy consumption to reduce demand on the electric system while saving money on their bills;
- public appeals for conservation (e.g., OPA's Summer Savings and Summer Sweepstakes Program and California's Flex Your Power Program);
- 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;
- technical support to large organizations to assist them in identifying energy efficiency investment opportunities, designing and evaluating solutions and implementing them; and

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 be 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 OPA 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. Ontario's load impact protocols for evaluating demand response resources are an example of this sort of protocols. 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. The OPA currently has a number of education initiatives under development. These include:

- **Builder training** – training and incentives designed to teach builders energy efficient building techniques
- **HVAC installation optimization** – training and installation incentives to contractors and other trades people aimed at increasing the quality of HVAC installation and maintenance
- **Building operator training** – training for building operators on operating buildings efficiently
- **Energy Manager training** – training and certification in energy management

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 – a (1) behavior change impact resulting in (2) 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 OPA EM&V Protocols and Requirements should be applied.

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

- **Process Evaluation Protocol** – OPA EM&V Protocols and Requirements, Process Evaluation Guidelines.
- **Market Effects Protocol** – OPA EM&V Protocols and Requirements, Market Effects Guidelines
- **Cost Effectiveness Protocol** – OPA 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 200C. 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 subpopulations 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 s 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

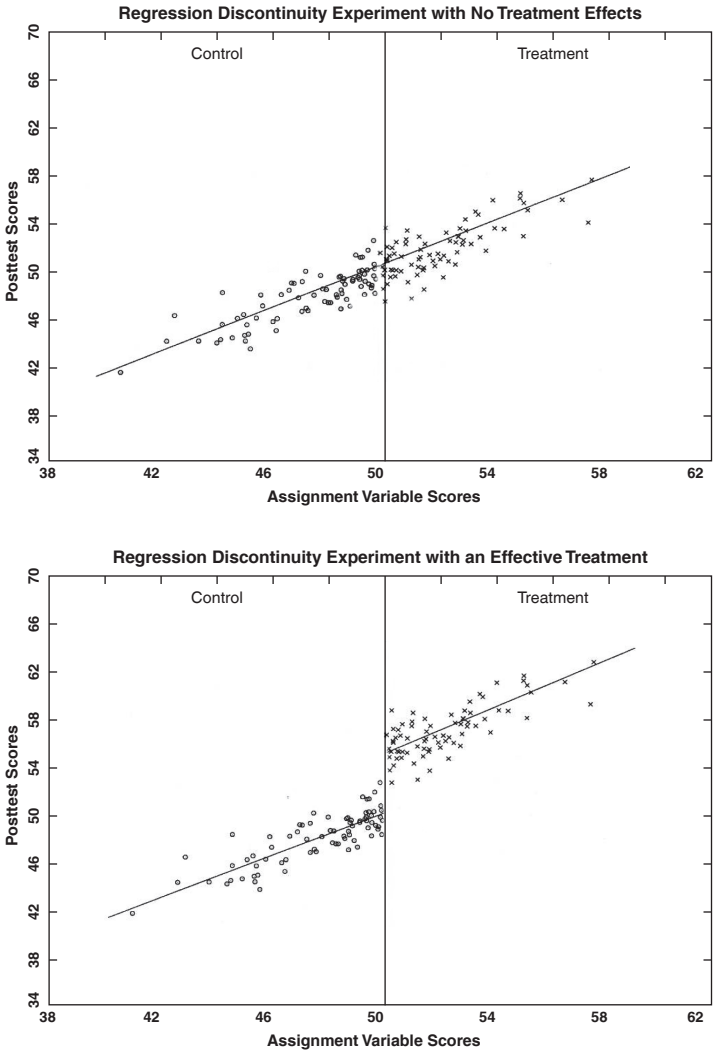
In the two true experimental designs discussed above (RCT and RED) subjects are randomly assigned to experimental groups – thereby establishing their statistical similarity. Under certain circumstances, assignment of subjects to treatments can be non-random provided subjects are assigned to treatment 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.

The OPA currently has a number of capacity building programs under development. These include:

- **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.

There are two basic types of capacity building programs in the mix of programs supported by OPA – training programs and segment support programs. 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 OPA)?
- 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 subpopulations (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-pout 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 OPA)?
- 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 OPA)?
- 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 at 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

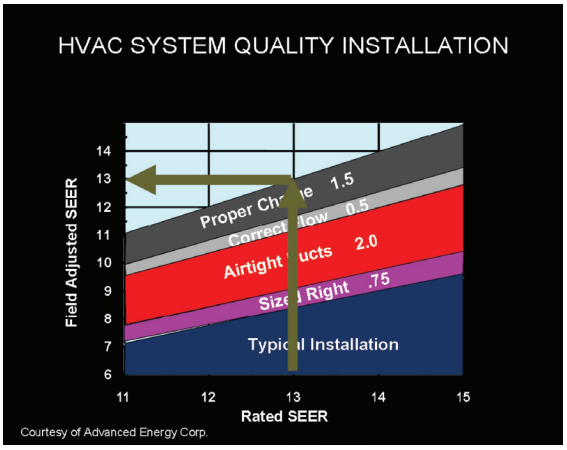
In this section, an example of the application of the evaluation protocols to a training program is presented. It is intended to show the level of depth that is required to meet the requirements of the protocols and to illustrate the types of information that are required to answer the questions in the protocols.

8.1.1 Introduction

The following example of a behavioral training program is sponsored by OPA and offered by Heating, Refrigeration and Air Conditioning Institute of Canada (HRAI) in 2013 and 2014. It is 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



OPA and the HRAI have developed and implemented a program for training personnel responsible for designing and installing AC systems; and OPA has made successful completion of the training course a condition of participation by contractors in its AC incentive program starting in 2014. The question is: how much impact is this training program having 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. It was offered by the HRAI in the winter and spring of 2013 and is being offered again in the winter and spring of 2014. It is offered on a first come, first served basis in a number of locations throughout the province of Ontario.

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 OPA's Heating and Cooling Incentive (HCI) program are required to complete the training course before they are eligible to participate in 2014. So, denying contractors access to the training would effectively deny them access to the HCI – an anticompetitive practice that OPA 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 (in 2013) with the knowledge, opinions and installation practices of those who have not (i.e., those who do not receive training until 2014). This effort requires:

- Surveying 2013 students concerning their knowledge, opinions and installation practices during the 2014 training period. This survey will be designed to observe the knowledge that 2013 student retained from the 2013 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 2013 training changed their practices.
- Surveying 2014 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 2013 students
- Comparison of installations of HVAC systems completed in the summer and fall of 2013 by parties who were trained in 2013 with HVAC systems installed during the same period by parties who are trained in 2014. Careful engineering reviews of the subject installations before and after training will 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 in 2013 and 2014 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 in 2013 and 2014, it will be necessary to intensively 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

Study Element	Sample Size
Survey of 2013 trainees	• 100 sales personnel • 100 installers
Survey of 2014 trainees	• To be completed on intake into the classroom for all 2014 trainees
Survey of installations	• 100 installations made by 2013 trainees in 2013 • 100 installations made by 2014 trainees in 2013 • 100 installations made by 2013 trainees in 2014 • 100 installations made by 2014 trainees in 2014

8.1.7 - Protocol 6: The Program Recruitment Strategy

Contractors are being recruited to the training on a first come, first served basis. All contractors who seek to participate in the HCI program must complete the training course prior to the 2014 cooling season.

All trainees in the 2014 training 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 completing their training in 2013 and 2014. 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 of $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 2013 trainees in the second year after their training. The period of the study will be 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 example of an awareness campaign is being sponsored by the California Public Utilities Commission in 2012, 2013 and 2014. During each of these years a randomly chosen subset of approximately 1/3rd of all small and medium sized commercial and industrial customers served by investor owned utilities in California is being defaulted on to time varying rates. 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, a public information campaign is being implemented. In this campaign customers who are about to be defaulted are 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 is being undertaken. The objective of the evaluation is to determine how effective the information campaign is 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 California Mandatory TOU Awareness Campaign is designed to inform selected non-residential customers that they are about to be defaulted on to time varying rates. The information campaign was carried out in the late summer and early fall of 2012 and 2013 and will be carried out again in 2014 prior to the default assignment of selected customers on to 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 is 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 explains why the rate change is necessary and that customers will no longer be able to subscribe to flat rates.

The Target Population

The target population includes non-residential customers that will be assigned to time varying rates in each defaulting period (i.e., November of each year). Within these overall populations there is 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 is 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 are expected to change the timing and magnitude of their electricity consumption after they are informed.

Whether the Exposure to Education Can Be Controlled

Education cannot be denied to parties who are about to be defaulted onto a time of use rate. Indeed the entire purpose of the information campaign is to ensure that all parties who are about to experience a significant rate change, are aware of it and understand 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**

While this is the case, the parties who are being defaulted onto time varying rates in each year are a randomly selected subset of all non-residential customers in California. A randomly selected subset of non-residential customers was defaulted onto time varying rates in November of 2012. In the fall of 2013, another randomly selected subset of non-residential customers was defaulted; and another randomly selected subset will be defaulted in 2014. 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, make 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 are the customers' understanding of how time varying rates work; their awareness of the fact that they are 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 are 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 2013.
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 are 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 makes 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 on to time of use rates in 2012. Another was defaulted in 2013 and the final group will be defaulted in 2014. Because of random assignment, the 2013 and 2014 groups are identical in all respects save the fact that the 2013 group received the awareness campaign in the fall of 2013.

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 2014).

The effects of the awareness campaign on customer knowledge and awareness of the impending rate change will be measured by surveying the following groups of customers:

- those who were exposed to the awareness campaign in fall of 2012, 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 2013 and were subsequently defaulted on to time varying rates in November of 2013 (i.e., those who experience the awareness campaign in the fall of 2013); 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 will be basically identical for all three surveys. However, customers who were defaulted in 2012 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 2013 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 will not experience the awareness campaign until 2014 will provide measurements of the levels of knowledge and awareness that are 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

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 2012 or 2013 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 incentive 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 rim 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 example of a pilot information feedback program is being implemented by one of Ontario's LDCs. 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; and

The program will be provided to 50,000 customers over the course of one year.

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 residing in the LDC's service territory.

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. A 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 is a randomized controlled trial (RCT) in which a random sample of 100,000 qualifying residential customers of the LDC will be randomly divided into two equal sized groups – treatment and control. The treatment group will be exposed to the feedback contained in the pilot. The experiment will take place over a two year interval with treatment group customers receiving 5 ERs per year. Treatment group customers will receive 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 will not receive ERs and will not have access to the website.

At the conclusion of the first year, treatment and control group customers will be 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 will be 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 LDC serves hundreds of thousands of customers, it will be necessary to select a sample of customers to participate in the pilot.

To select customers to participate in the pilot a random sample of 150,000 residential customer records will be randomly sampled from the LDC's customer information system and delivered to the energy report vendor. The vendor will use these records to identify customers who are eligible to receive the treatment. Typically, this involves removing customers for which it is impossible to define neighbor 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

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 the LDC. 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 OPA.

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