EPRI | ELECTRIC POWER RESEARCH INSTITUTE

Measurement and Verification for Behavioral Programs

Evaluating Programs That Have Gone Full-Scale

3002001269

Measurement & Verification for Behavioral Programs

Evaluating Programs That Have Gone Full-Scale

3002001269

Technical Update, December 2013

EPRI Project Manager

J. Robinson

ELECTRIC POWER RESEARCH INSTITUTE
3420 Hillview Avenue, Palo Alto, California 94304-1338 ▪ PO Box 10412, Palo Alto, California 94303-0813 ▪ USA
800.313.3774 ▪ 650.855.2121 ▪ askepri@epri.com ▪ www.epri.com

DISCLAIMER OF WARRANTIES AND LIMITATION OF LIABILITIES

THE FOLLOWING INDIVIDUAL, UNDER CONTRACT TO EPRI, PREPARED THIS REPORT:

Matthew Harding

This is an EPRI Technical Update report. A Technical Update report is intended as an informal report of continuing research, a meeting, or a topical study. It is not a final EPRI technical report.

NOTE

For further information about EPRI, call the EPRI Customer Assistance Center at 800.313.3774 or e-mail askepri@epri.com.

ACKNOWLEDGMENTS

The following individual, under contract to the Electric Power Research Institute (EPRI), prepared this report:

Matthew Harding

This report describes research sponsored by EPRI.

This publication is a corporate document that should be cited in the literature in the following manner:

Measurement & Verification for Behavioral Programs: Evaluating Programs That Have Gone Full-Scale. EPRI, Palo Alto, CA: 2013. 3002001269.

ABSTRACT

The evaluation of behavioral programs requires rigorous measurement and verification. While the randomized controlled trial (RCT) lies at the core of modern program evaluation, in most situations it is not feasible to implement a randomized approach. This is particularly true when programs have gone full-scale and the success of the program needs to be evaluated outside of the confines of an experimental framework involving the random allocation of households to treatment and control groups.

In the absence of this experimental framework, the main problem that needs to be overcome is the bias associated with self-selection of program participants into treatment groups, which leads to different baselines that cannot be directly compared to non-participants. As a result, naïve approaches comparing the average outcomes after program implementation of the different groups of customers lead to misleading results.

This report aims to describe methods for performing a scientifically sound program evaluation in situations where a properly randomized experiment is not possible. A number of different statistical approaches are available that when implemented correctly, can be used to recover unbiased estimates of the causal effects of a program, even in the absence of an RCT. These methods require the availability of additional observable variables on households participating in the behavioral program and on a sample of non-participating households. Variables such as demographics, property characteristics, neighborhood characteristics, and pre-program usage patterns can be used to profile customers and characterize the difference between participating and non-participating households. If the profiling is accurate, a range of statistical techniques involving matching are available to restore the balance between the treatment and the control group. These methods eliminate the selection bias inherent in comparing the outcomes of groups of households which are different along a number of different dimensions.

Matching and propensity score methods are widely available in contemporary statistical software packages such as SAS or STATA and their implementation is not too challenging computationally. It is important to recognize the fact that the methods rely on untestable assumptions about household behavior. If the analyst can make a persuasive argument that the underlying assumptions are correct then the results of these methods will be credible. Additional research is required to document the relative performance of the different methods in the context of utility programs.

Keywords

Behavioral program evaluation
Causal effects
Rubin Causal Model
Matching
Propensity score
Covariate balance
Weighted regression
Randomization

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1
INTRODUCTION

In recent years, analysts across industries have become increasingly aware of the need to provide rigorous measurement and verification for behavioral programs. Of particular interest are programs aimed to induce change in usage behaviors. Cost and credibility are the primary concerns for the development of scientifically sound program evaluation methodologies. We are increasingly witnessing a methodological move towards data driven decision-making, which is only going to gain momentum as new and large data resources become available (i.e., the Big Data phenomenon).

The randomized controlled trial (RCT) lies at the core of modern science and informs decision makers of the efficacy of a wide range of interventions, from the latest medical treatments to the allocation of foreign aid. While industry knowledge and experience are still very valuable, statistically rigorous evaluations promise to deliver scientifically accurate measures of the effectiveness of different programs and policies. Industries are moving away from an experience and opinion based decision process towards a formal data driven process which delivers the maximum performance at the minimum cost, free of subjectivity, human biases and preconceptions.

Randomized experiments, when correctly designed and implemented, represent the gold standard of evaluation and measurement strategies. An RCT can provide an unbiased estimate of the causal impact of a behavioral program. Furthermore, it is amenable to a fairly straightforward statistical analysis and has the advantage of producing results that are easy to convey to a broad audience. In recent years we have witnessed the increased adoption of RCTs in the implementation of various utility programs. Indeed, much of EPRI's recent work related to behavioral programs has focused on protocols for conducting RCTs.[1,2]

While the continuing use of RCTs should be wholeheartedly embraced, it important to realize that in many real-world cases it is not possible to adopt an experimental approach. This report aims to provide guidelines for performing a scientifically sound program evaluation in situations where a proper randomized experiment is not possible. The main problem that needs to be overcome in the absence of an experimental framework is the inherent self-selection of program participants into the treatment groups, which can lead to different baselines that cannot be directly compared to the non-participants. As a result, conventional analytical approaches will produce biased and misleading estimates of the effectiveness of a program. Decision-making based on biased evidence leads to costly mistakes, which could have been avoided.

The main message of this report is that when done correctly, it is possible to recover unbiased estimates of the causal effects of a program, even in the absence of an RCT. Extensive research

---

[1] *Quantifying the Impacts of Time-Based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines*. EPRI, Palo Alto, CA: 2013. 3002000282.

[2] *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols.* EPRI, Palo Alto, CA: 2010. 1020855.

exists on addressing these issues in the fields of statistics and econometrics. In fact these disciplines would be substantially less interesting if they were restricted to the analysis of randomized data. While the mathematical framework required for the analysis is easily available in most statistical software packages, the success or failure of evaluation strategies in the absence of a randomized setting depends crucially on the following two ingredients:

- In-depth understanding of customer preferences and program implementation details
- Availability of extensive customer level data at the population level

While the availability of extensive customer level data often presents the major challenge, the utility industry is somewhat uniquely situated to benefit from the increasing availability of data collected from smart meters, smart thermostats, etc., as well as the availability of property and household level information from third-party data aggregators.

This report is intended to provide a working knowledge of the current approaches to program evaluation for utilities and their M&V contractors that have moved away from pilots to full-scale programs, and need to evaluate the effectiveness of these programs. Additionally, this report will provide important insights to utilities which originally designed RCT pilots but where the randomization was compromised due to on-the-ground implementation realities. It is important to be aware that even when pilots are designed according to the most rigorous randomization principles, implementation constraints that may seem relatively minor lead to the pilots no longer being adequately controlled. Proceeding in the evaluation on the assumption that the program was randomized when in fact the implementation diverged from the randomization ideal leads to bias; in such circumstances, one of the analysis approaches described in this report would be much more appropriate.

This insight is by no means new to the utility industry. Practitioners have long been aware that small design features, introduced in an otherwise randomized experiment, may lead to different conclusions. In an analysis of one of the earliest time-of-day pricing experiments from Arizona in 1976, Aigner and Hausman (1980) show that when ignoring the best-bill guarantee that was offered to customers in the randomized pilot, one estimated a peak price elasticity of demand that was larger than the off-peak elasticity. Once the best bill guarantee was accounted for, the result was reversed and made consistent with other experiments conducted at that time. This shows that even something as simple as a best-bill guarantee may sufficiently distort an otherwise randomized pilot and lead to evaluation bias. Small design elements that may seem innocuous *ex ante* may sometimes lead to unintended consequences when conducting the program evaluation *ex post*. Nevertheless, such biases can often be removed using statistical methods such as the ones described below, and they do not invalidate the evaluation as long as we are aware of them and use an appropriate method.

This report is structured as follows. Section 2 introduces the conceptual framework needed to characterize causal effects in a non-randomized program environment. Section 3 describes the assumptions and data required for the evaluation. Section 4 introduces a number of statistical evaluation approaches for a variety of situations that are likely to be encountered. Section 5 discusses how we can test and further interpret our results.

2
NONEXPERIMENTAL METHODS

The modern statistical literature relies on a well-established framework to think about treatment effects, the *Rubin Causal Model*. The framework dates back to the early part of the 20th century when statisticians first started rigorously evaluating randomized experiments. This framework was extended in the 1970s to model non-experimental methods (Rubin, 1974; 1977).

To understand its implications, first, let us define some basic statistical concepts. We will denote by Y_i the variable of interest, for example electricity usage measured over a certain interval, such as an hour or month. The subscript i denotes the household or firm for which this outcome is recorded. For simplicity we will consider the units to be households. Throughout, we assume that we observe Y_i for a large sample of households $i = 1...N$, where N is the number of households under examination. Thus, a typical dataset may contain monthly billing data for a large number of households and the outcome of interest, Y_i, denotes the usage in kWh for each household in the data.

Let us further consider the case where a behavioral program is made available to these households. The range of programs that can be evaluated using this methodology is very broad and includes conservation programs (e.g., monthly reports that compare a customer's usage with that of their neighbors), price based programs (e.g., introducing a TOU rate), or technology based programs (e.g., installing a smart thermostat). For the purpose of setting up our model, the type of behavioral program is not essential. Each involves imposing a treatment intervention on the household that is intended to alter its electricity usage. We denote by D_i an indicator variable which captures whether or not a given household participates in the behavioral program that we are evaluating:

$$D_i = \begin{cases} 0 & \text{if household } i \text{ does not participate in the behavioral program} \\ 1 & \text{if household } i \text{ participates in the behavioral program.} \end{cases}$$

We shall refer to households who choose to participate in the behavioral program as "treated households", while households that don't are labeled as "control households". Note however that here we don't assume that the allocation to treatment or control necessarily followed from randomization, and in fact we are mostly interested in the case where treated households are not allocated randomly to the behavioral program, but rather are selected into the program or have chosen to be part of it when given the opportunity.

We also need to introduce one basic mathematical operator, $E[Y_i]$, the expectation of Y_i. The expectation designation means that the outcome is not certain, but depends on factors that are not controllable in the experiment or program. From a statistical perspective, Y_i is a random variable and can take any value in a specified relevant range. In any given sample, each household exhibits a value which we measure directly. We can think of the expectation of a random variable as the average value of this variable taken over a very large sample of households. This reminds us that most statistical methods are exactly correct only in very large samples. It is

easier conceptually however to define techniques in terms of very large samples even though in practice this may not always be a great starting point. For many situations of interest, however, samples involving hundreds or thousands of households are sufficient for the statistical concepts to be approximately correct.

An important extension of the expectation is the conditional expectation $E[Y_i \mid X_i = x]$. This denotes the expectation of the variable Y_i conditional (the vertical bar symbol "|") on the value of another variable X_i. This is a useful concept and we can think of it as measuring the average outcome of Y_i in large samples, if a certain condition has also been satisfied. The unconditional expectation may not always be of interest and in many situations we can learn more by considering conditional expectations. For example, the average electricity usage in a state may not tell us everything we are interested in and we may find it more useful to look at the average electricity use of households with children or households below the poverty line. Both are valid expressions that we can condition expectations on, but they convey different information and have different uses.

The Rubin Causal Model

The Rubin Causal Model asks us to imagine that for each unit (e.g., a household) we have two *potential* outcomes, which we shall denote by $Y_i(1)$ and $Y_i(0)$ corresponding to the outcome if the household is treated or not treated respectively. These outcomes are potential because in practice it is impossible to observe both outcomes at the same time. Any given household is either part of the program or it is not. Since we are interested in measuring the effect of a program it is conceptually important to think about *counterfactuals* – that is, what *would have been*. The potential outcomes framework allows us to ask, for a household that participates in the program, what would have been if a household had *not* been in the program. Similarly, for a household which is not in the program, what would have been if that household *had* been in the program. We are thus using counterfactual reasoning to make statements about the causal impact of a program. From a practical point of view, evaluating counterfactuals is not trivial because they are unobservable values. It is impossible for a household to be both in the program and not in the program at the same time.

By this logic, it follows that the actual outcome is one or the other potential outcome depending on the treatment status of a given household.

$$Y_i = Y_i(D_i) = \begin{cases} Y_i(0) & if \quad D_i = 0 \\ Y_i(1) & if \quad D_i = 1 \end{cases}$$

The left hand side of the equation asks the outcome of Y_i given its treatment state D_i. The right hand side says that outcome is designated by setting $Y_i = 0$ if the household was not treated, and $Y_i = 1$ if it was treated.

For every household only one of the potential outcomes is observed. The counterfactual outcome corresponds to the outcome which is not observed. Since it is impossible to observe both outcomes for a given households, we have:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = Y_i(0) + D_i(Y_i(1) - Y_i(0)) \tag{1}$$

The main problem of causal inference lies in the fact that in order to make causal statements as defined by Eq. (1), we need to rely on both observed outcomes and counterfactual outcomes (which are by definition unobserved). Statistically this means that we need to make inferences from missing (unobserved) data. As we shall see later, all methods attempt to provide a mechanism for imputing the missing data by finding reasonable values for the counterfactual outcomes.

This equation also says that in general the observed outcome is always going to be a function of the treatment choice, D_i, because in the calculation of Y_i, D_i is involved regardless of whether Y_i was treated or not. This says that when evaluating electricity usage it is always very important to understand the process through which households are allocated to either the treated or non-treated groups because the observed outcomes will always depend on the allocation process. The only exception is the RCT case, where the random nature of the allocation ensures that no additional information is contained in D_i that is needed for the evaluation of the outcomes.

In this section we have introduced the Rubin Causal Model as a modern framework for thinking about causality in the evaluation of behavioral programs. We have seen that in general we cannot make causal statements without also considering the process through which households are allocated to the treatment and non-treated groups. In the next section we will see what the implications of this insight are, and how it can be used to define and measure treatment effects,

3
TREATMENT EFFECTS

The ultimate goal of evaluating behavioral programs is to measure the treatment effect of the program on the outcome of interest. This enables us to determine whether a behavioral program succeeded in changing behavior, and quantify the magnitude of the expected change induced by the program. Ideally we would want to quantify the *individual treatment effect* for every unit in our sample: $Y_i(1) - Y_i(0)$. As we saw before, this is not directly implementable since we never observe the same individual both treated and not-treated at the same time. Returning to Eq. (1) above, we see that the outcome for a treated unit is the outcome if the unit were not treated plus the treatment effect of participating in the program. This also shows that in general we should expect the treatment effect to be different for every household in the sample. Our inability to observe both potential outcomes prevents us from evaluating the entire distribution of treatment effects. Instead we are forced to use the observed data which consists of measurements of units, which are treated, and units which are not, to make inferences about the unobserved treatment effect, while also relying on additional untestable assumptions, which we will introduce below.

In some sense we can think of the various techniques described below as a statistical attempt to impute the missing counterfactual observations using a series of assumptions about the data and household behavior. The discussion below does not assume randomization, which in general is a very special case of a behavioral program where the selection into the program happened at random.

We are interested in the case where the selection into the program does *not* occur at random. This is typical for full-scale utility programs where customers are free to choose to participate in the program, and no provision is made upfront to develop a randomly assigned control group for comparison purposes.[3] When selection into the program does not occur at random, this means that $D_i = 1$ depends on a whole range of factors involved in the decision to participate by customer i.

Statisticians typically focus on estimating two objects of interest, the average treatment effect (ATE) and the average treatment effect on the treated (ATT). The average treatment effect is measured by:

$$ATE = E[Y_i(1) - Y_i(0)] \tag{2}$$

and captures the average expected gain from the program for a household selected at random from the population. Intuitively, this says that what we are interested in is the average impact of treating a household relative to the counterfactual of not treating that same household. Of course

[3] Note there are situations where it is possible to have both customer self-selection and randomization. In "recruit and delay" or "recruit and deny" experiments, customer are invited to participate in a program, and all volunteers are subsequently randomized into either the treatment or the control group. This approach controls for selection bias, but results can only be generalized to other volunteers in the population. Regardless, recruit and deny/delay approaches are just variants on RCTs; the focus of this report is programs that rely on self-selection, and for which there are no pre-determined randomly assigned control groups.

in practice, for any household, we only observe one of the states of the household since at any given point in time the household is either treated or not treated.

The average treatment effect on the treated is measured by:

$$(3) \qquad ATT = E[Y_i(1) - Y_i(0) \mid D_i = 1]$$

Here we employ the conditional operator to look at treated households only. Eq. (3) captures the average expected gain from the program for a household that is actually treated. Notice that this definition is subtly different from the one for the ATE above. While the ATE measures the gain from the program that *would be* achieved if we were to treat any household selected at random, the ATT looks at the gain from the program for those households that are *actually* treated because they opt to select into the program. We know that in practice many programs are only attractive to a select group of customers, so this distinction is highly relevant when we think about considering the effect of a program that has gone full scale.

Let us now consider the *naïve* approach to estimating the causal effect of a behavioral program that ignores the process through which households decide whether or not to participate in a behavioral program. The most compelling approach would be to compare the outcomes for the households that participate in the program with the households that do not. Mathematically this leads to the following expression:

$$(4) \qquad E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0].$$

Notice however that this is not equal to the expression in Eq. (3) above. The difference comes from the fact that in Eq. (4) we are comparing observed outcomes only, when in reality computing the treatment requires us to compare actual outcomes to counterfactual ones, and ask the question what *would have* happened if a household would not have been treated, i.e., as specified in Eq. (3).

This is not what the naïve approach does, and in fact we can show that the naïve approach presents a biased measure of the actual causal effect of the behavioral program. To see that let's add and subtract $E[Y_i(0) \mid D_i = 1]$ to the right side of Eq. (4):

$$\begin{aligned} & E[Y_i \mid D_i = 1] - E[Y_i \mid D_i = 0] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1] + E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \\ (5) \qquad & = E[Y_i(1) - Y_i(0) \mid D_i = 1] + \{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\} \\ & = ATT + \underbrace{\{E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\}}_{} \\ & = ATT + \qquad \text{Selection Bias} \end{aligned}$$

The naïve approach estimates an effect which equals the ATT plus an additional term (labeled "selection bias" in the last line, and defined by the bracketed part of the line just above) which in general is not zero and is usually referred to as the selection bias. The reason for that is when households can choose whether or not to participate (i.e., when they are *not* randomly assigned) and when a "recruit and deny/delay" variant on an RCT is not used, the baseline outcome for households who choose to participate is not necessarily the same as the baseline outcome for households that don't. So in general, $E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0] \neq 0$, and the naïve approach produces a biased estimate of the ATT.

The main problem we need to overcome in evaluating behavioral programs is the presence of selection bias. When households decide whether or not to participate in a program, those that participate (the treatment group) are not directly comparable to the households that do not participate. The degree of bias is directly related to the degree of selection. There are many potential factors that can drive selectivity into a program. We can think of selection as being of two types:

- Selection on observables
- Selection on unobservable

Selection on observables happens when features of the household that can potentially be observed by a researcher are driving whether or not a household participates in a program. Examples of observables may include:

- Household demographics (e.g., income, family composition)
- Home features (e.g., age, HVAC, pool)
- Geography (e.g., city, school district)

As we shall discuss below, when selection is driven by (and only by) a specific set of observables we have a number of statistical techniques available to us that can disentangle the treatment effect from the selection bias.

Much more problematic is the case where selection is driven by unobservables. For example, participation in a program may be driven by a household's beliefs over the validity of climate change predictions. Another example is the case of households who plan to have children and, anticipating the change in future usage, decide to sign up for a behavioral program which they think would be beneficial to them due to future circumstances, which are not currently observable. It may be that some unobservables truly remain unknown, but it is also possible that in many situations of interest we can collect observable proxies which correlate strongly with the unobservable attributes.

Selection Bias and Treatment Effects

It may help to consider a simple graphical illustration of the selection bias problem and the issues related to the use of the naïve comparison to measure the causal effect of the behavioral program in a stylized example. This example is similar to a real world program which is offered to all customers. To keep things simple we will look at the situation where the program is attractive to all high income customers above a certain income threshold, and it is not attractive to customers below that threshold. As a result, selection into the program is purely conditioned by income, all high income customers participate, while all low income customers do not. While this is a rather stylized example of a real world example, it is simple enough that the analysis can be illustrated graphically using simple charts. More complicated examples will be discussed later, but the principles underlying the program evaluation are the same.

In Figure 3-1, we present a hypothetical example where household electricity consumption measured in kWh (the Y axis) is related to household income (the X axis). In particular, consumption increases monotonically with income (meaning it increases consistently, without ever decreasing).

The graph shows two functional relationships: $Y_i(0)$, household consumption if the household does not participate in the behavioral program, and $Y_i(1)$ household consumption if the household does participate in the program. Note that these functions are defined (by the solid line) as being mutually exclusive. Selection into the program is driven by income. Low income households choose not to participate in the program ($Y_i(0)$ is defined only for low income customers), while all households above the threshold level of income (defined by the vertical dotted line) participate in the program ($Y_i(1)$ is defined only for high income customers).

Participation is denoted by D_i, which takes the values 0 or 1 as before. Recall, however that we never observe electricity consumption in both counterfactual states of the world. We denote the consumption that we observe by the continuous lines and the consumption that we don't observe by the dashed lines. Notice that in this example we observe the consumption of low income households who do not participate in the behavioral program and the consumption of high income households who do participate.

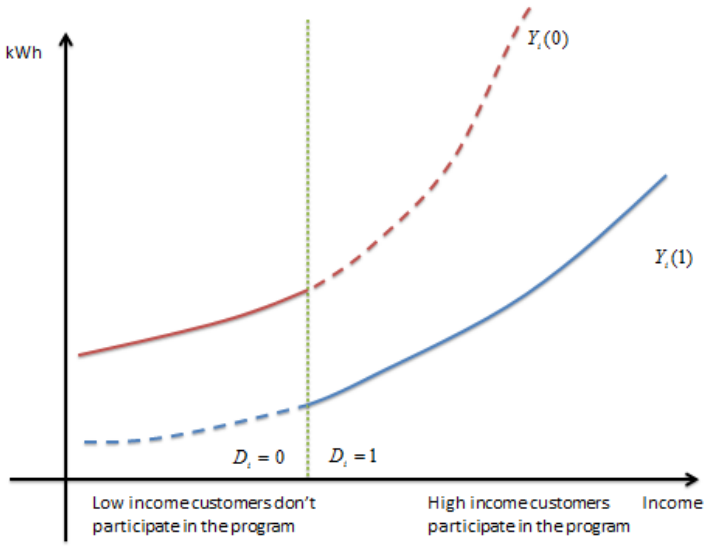

Figure 3-1
Energy Use and Program Participation

Let us now use this scenario to illustrate the problem of trying to estimate the causal effect of this program. The two quantities we are usually interested in are the average treatment effect and the average treatment effect on the treated as defined above. In Figure 3-2 we show what the ATE corresponds to. According to its definition the ATE captures the benefit in terms of kWh reductions for a random household irrespective of its income. Denote by A the average usage of households that do not participate in the program and by B the average usage of the same households participating in the program. Recall that these quantities are not actually observed in practice. But if they were, the ATE would be given by the kWh value of A minus B in Figure 3-2, which would suggest that the program was very successful in reducing energy use.

The other quantity of interest is the ATT. In Figure 3-3 we show a graphical representation of the treatment effect on treated households. Now we are only comparing the mean usage of

households that participate in the program with their mean usage if they did not participate in the program.

There is a subtle difference between the graphs in Figure 3-2 and Figure 3-3. In Figure 3-2 the points A and B correspond to the average of the electricity usage for all consumers irrespective of income. In Figure 3-3 the points A and B correspond to the average of the electricity usage for high income consumers. In both figures A corresponds to the average for the consumers if they were not treated, while B corresponds to the average for the consumers if they were treated. The difference is that Figure 3-3 is computed over the subsample of high-income consumers, while Figure 3-2 is computed over all consumers. Thus, both A and B are higher on the Y axis in Figure 3-3 compared to Figure 3-2.

This example captures the intuition that while ATE looks at the average program impact for a random consumer, ATT looks at the impact on a subsample of consumers that actually selected into the program, in this case a subsample of consumers with high income.

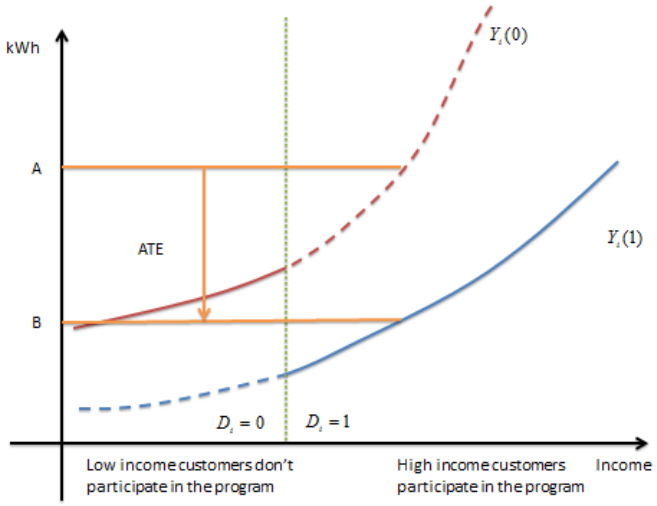

Figure 3-2
Average Treatment Effect from Program Participation.
A corresponds to the average value of all points on the Y(0) curve. B corresponds to the average value of all points on the Y(1) curve.

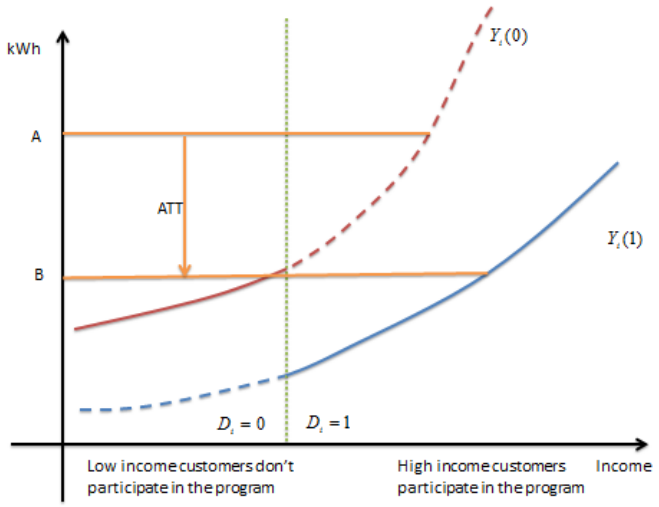

Figure 3-3
Average Treatment Effect on the Treated.
A corresponds to the average value of all points on the Y(0) curve for which D=1 (i.e., only high income customers). B corresponds to the average value of all points on the Y(1) curve for which D=1 (i.e., only high income customers).

Now consider the naïve evaluation approach introduced above which compares the usage of households which signed up for the program with the usage of the households that did not. In our example this means that we are comparing the usage of high income households which are part of the program (the line B) with the usage of low income households which did not sign up for the program (the line A). Figure 3-4 shows that this approach would be seriously misleading.

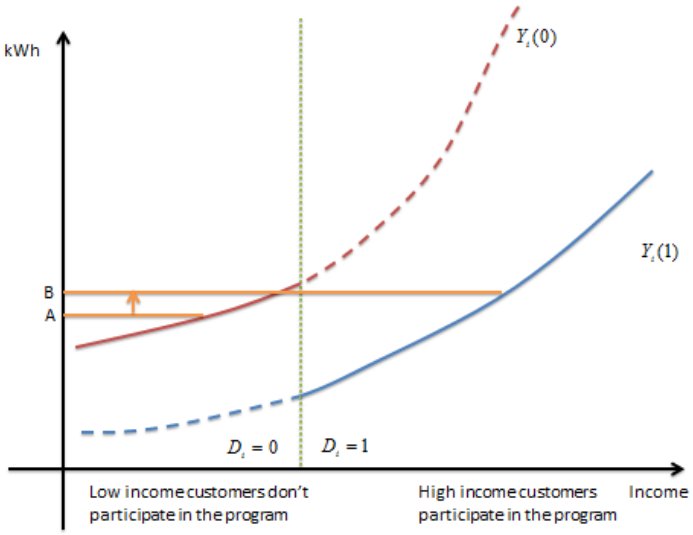

Figure 3-4
Naïve Approach to Measuring the Causal Effect.
A corresponds to the average value of all points on the Y(0) curve for which D=0 (i.e., only low income customers). B corresponds to the average value of all points on the Y(1) curve for which D=1 (i.e., only high income customers).

We can see from Figure 3-2 that the ATE is large and negative, meaning a decrease in electricity use, while Figure 3-4 shows that the naïve approach would estimate a small increase in electricity usage resulting from the implementation of the program. This is because we are now comparing the electricity usage of high income consumers to that of low income consumers. Even though the program was extremely effective, it is not measured correctly by the naïve approach.

Depending on what we are looking to measure, either ATE or ATT are the proper measures of the program's effect, but the naïve approach is never the correct measure.

The above discussion, while stylized, does capture the challenge posed by selection bias in estimating causal effects and the need for a more advanced statistical treatment than is provided by the naïve approach. Given the illustrative nature of this example, it is easy to see how the selection bias arises from comparing high and low income customers which are sorted into treatment and control group as a function of income. The challenge for us comes from the fact that in reality we would expect many different factors (covariates), other than income, to contribute to households' decision to sign up for a program.

These figures also show why randomization helps to solve the selection bias problem. If households are randomly allocated to treatment and control groups we would expect both groups to have households across the entire income spectrum. As a result it is enough to compare the mean usage of households in the treatment group with the mean usage of the households in the control group. If a large enough number of households are included we can get a good approximation to the two curves in Figure 3-3 and compute the ATE accurately. In a randomized trial ATE and ATT are identical. It is important to keep in mind that even in a randomized trial it may be difficult to enforce a perfect randomization. In general it is not possible to prevent households from opting out and if a large enough number of households with specific characteristics decides to opt out, the initial randomization is in effect annulled, and we are still dealing with a selection problem.

Addressing the Selection Problem: Assumptions

The assumptions we are willing to make will determine both the choice and credibility of the methods that we can employ to solve the selection problem. Consider the problem of correctly measuring the ATT discussed above. As we saw, the naïve estimator suffers from selection bias unless $E[Y_i(0) \mid D_i = 1] = E[Y_i(0) \mid D_i = 0]$. This means that the process, which determines whether a household participates in the treatment does not impact the potential outcome $Y_i(0)$. A stronger condition is to assume that the exposure of a household to the behavioral program is independent of both outcomes, $Y_i(1)$ and $Y_i(0)$. In statistics we use the symbol $\perp$ to denote the fact that a variable or set of variables is statistically independent of another variable or set of variables. In the context of our current discussion one assumption which would guarantee that selection bias is not a problem for the estimation of the causal effects of interest is:

$$Y_i(0), Y_i(1) \perp D_i$$

which says that the potential outcomes for each household are independent of the participation of the household in the program. By design, a randomized trial guarantees that this assumption is met and therefore we can estimate the treatment effects without contamination resulting from

selection bias. In a correctly implemented randomized trial the design mechanism prevents selection and therefore selection bias.

In order to develop methods which estimate the causal effects we need to replace the assumption above, which in general is only met in randomized trials, with a different set of assumptions which are also met in non-randomized settings and that can then be used to develop methods for quantifying the causal effects of interest.

Given the central role that assumptions play in our methodology, it is worth pausing to remind ourselves that assumptions are central to any attempt to make inferences. A randomized trial relies on the assumption of independence described above, and there is no such thing as an assumption-free method. It comes down to making sure that the assumptions we are making are appropriate for the program that we are evaluating. Not all assumptions are always credible or reasonable and thus we have to make sure that we are as careful as possible when choosing what assumptions we make. Above all it is crucial to be as clear as possible when imposing assumptions in order to give the reader the opportunity to decide for herself whether she believes our assumptions and the study results that follow from invoking these assumptions.

In a recent book discussing best practices for conducting policy analysis, Manski (2013) summarizes the role of assumptions in program evaluation as follows:

> "Policy analysis, like all empirical research, combines assumptions and data to draw conclusions about a population of interest. The logic of empirical inference is summarized by the relationship: assumptions + data = • conclusions. Data alone do not suffice to draw conclusions. Inference requires assumptions that relate the data to the population of interest. […] The fundamental difficulty of empirical research is to decide what assumptions to maintain."

The main assumption we will rely upon in order to estimate causal effects in behavioral programs *without* randomization is expressed as follows:

> Assumption 1 [Selection on observables[4]]: Conditional on observable attributes X_i, the assignment of the households into the treatment and control group is independent of the potential outcomes:
>
> $$Y_i(0), Y_i(1) \perp D_i \mid X_i$$

This means that once we account for observable characteristics of a household such as demographics, home features, or past electricity usage, the allocation of households into treatment and control groups is independent of the potential outcomes. The logic of this assumption is easy to understand once we apply it to a randomized trial. In a randomized trial we would expect the various observable attributes of households in the treatment and control groups to be very similar on average. When we randomize households we should see comparable levels of income or home features present in both groups. As we already discussed, selection occurs when households with certain characteristics are more likely to participate in the program than

[4] This assumption appears in the statistics literature under a variety of names, including *unconfoundedness*, *conditional independence*, *ignorability of treatment,* or *missing at random*.

others. In the examples above we saw how the selection of households into a program based on income levels makes it impossible for us to compare the mean outcome of the treated group with that of the control group in order to make causal inferences.

The selection independence assumption says that we may be able to recover the causal effect of a behavioral program if we can identify observable features of the households, which, when accounted for, would allow us to compare the treated and control groups. In the example above, since the selection is driven only by income and we can potentially observe household income from surveys or census data, the assumption is satisfied and a number of statistical methods will be available to us to quantify the causal effect of the behavioral program. This assumption highlights the central role that understanding customer behavior has in evaluating the effectiveness of behavioral programs. Saying that we need to account for the observable attributes of the household which determine participation is another way of saying that we need to understand the factors which drive some households to sign up for a behavioral program or the factors that drive households to opt out from such a program. If we can characterize their decision in terms of variables that we can observe and measure we will also be able to use statistical methods to accurately estimate the treatment effect of the behavioral program.

In a randomized trial we could in principle estimate the average treatment effect simply from data on electricity usage for the control and treatment groups. Without randomization, we need to collect additional data on household features that may explain why some households decide to participate in the program while others decide not to. Since the participation decision is most likely influenced by many different factors or may differ across customer groups, this assumption also implies that we'll need a substantial amount of additional information on the households. With the increased availability of data on customer demographics and behaviors from commercial data collection firms, numerous measures of household characteristics are now readily available. Statistical models and qualitative knowledge of customer behaviors need to be combined in order to determine the best choice of observable attributes X_i , which best capture the factors driving the household decision to participate in the program.

It is important to remind ourselves, however, that this assumption precludes the existence of factors which may explain the participation decision of the households, but are not observable. The decision to participate may be driven by factors that are not easily observed or quantified, such as a household's subjective evaluation of the state of the economy, its anticipated future energy use needs, or even whether a recruitment mailer just happened to catch a householder’s eye. To the extent that such unobservable factors are present and explain participation in the program, we should expect our measures of the causal effects to be biased. It is important to use all the observables that are available to us which may explain the participation decision in order to minimize the impact of unobservable factors that we cannot include in our analysis.

> Assumption 2 [Stable Unit Treatment Value]: the outcome for a household participating in the program does not depend on the mechanism or pattern of participation of any other household.

This assumption excludes many forms of programs where the success of the program is conditional on the participation of other households. It assumes that the impact of the program on a given household which participates in the behavioral program is not dependent on the number of households that also participate or the composition of the group of households that form the treatment group. It explicitly excludes any form of program that depends on a "snowball" effect

where success is a function of the number or type of other participating households. The causal effect of a program is not allowed to increase in intensity with the number of households which sign up. Furthermore, the impact on electricity usage of a behavioral program for a household that participates should not depend on whether the neighbors also participate. It should also not depend on how many or which neighbors participate. We are explicitly excluding programs where the effectiveness of the program depends on the overall popularity of the program. An example would be a program aimed at energy conservation, for which the conservation effort of a household depends on the perceived conservation effort and the number of other households that are also participating in the program.

This assumption may strike us as surprisingly strong since it excludes the usual network or social interaction effects that have made social media platforms so popular in recent years. The causal effect of a behavioral program should be driven by the features of the program and not by what other households may or may not do. It also reminds us of the dangers of extrapolating too far into the future since the assumption also excludes what economists call "general equilibrium" or systemic effects. For example, a behavioral energy efficiency program may induce households to purchase photovoltaic panels or some other energy related technology. As the panels increase in popularity and sales go up we would expect their price to go down, which over time would induce more households to also purchase photovoltaic panels and eventually overall electricity usage is lowered.

These types of long-term adjustments may induce further macroeconomic effects by changing the allocation of labor across industrial sectors, until the economy reaches a new equilibrium. For most behavioral programs that we are interested in this is likely to be of limited concern, but in general it is something we ought to be aware of. While the adoption of electric vehicles is still very limited we may reasonably conclude that the adoption of such a technology on a large scale may induce broad changes to the electricity industry and we ought to be careful when measuring the causal effects of any program that may potentially lead to broad (and sometimes unanticipated) consequences.

Before introducing our last assumption, we need to define a very important statistical concept, the propensity score, P_i. The propensity score measures the conditional probability of participating in the behavioral program, i.e., of being treated, conditional on the set of observables X_i for each household:

$$P_i = \Pr(D_i = 1 \mid \mathrm{X}_i) .$$

The propensity score measures the probability that a given household, characterized by observable attributes X_i, participates in the program.

Let us review the stylized example described above. In the example households with high incomes above a threshold level participate in the program, while households with income below that threshold do not. In this example X_i is measured household income. Denote the threshold that determines whether or not households participate by c. According to this definition if a given household has income $X_i \geq c$ then:

$$P_i = \Pr(D_i = 1 \mid \mathrm{X}_i \geq \mathrm{c}) = 1,$$

while a household with income below *c* has:

$$P_i = \Pr(D_i = 1 \mid \mathrm{X}_i < \mathrm{c}) = 0$$

What makes our example stylized is the fact that it only considers the case where the decision to participate or not is a very deterministic function of income which depends purely on whether or not a household is above the income threshold.

Our last assumption says that if household participation is fully determined by the income threshold *c* it will be impossible to quantify the treatment effect. In the real world there is no precise income threshold but it is still possible that we find ourselves in a situation where a program is more popular with high income than with low income customers. Thus, the probability of participating is closer to zero for low income customers and closer to 1 for high income customers, even if no customer has a probability of participation which is exactly 0 or 1. But this does help us since as we shall see below an important part of our strategy will be to compare households which participate with households that do not, but which have similar income levels.

Measures such as matching, which will be described below, seek to match customers in order to develop treatment and control households for comparison. In the stylized example, we have households participating in the behavioral program that are not comparable to the households which do not participate. What we require is the ability to find households which do not participate in the behavioral program at levels of income which are comparable to those of the households that participate. This assumption is labeled the *overlap condition* or the *common support condition*:

> Assumption 3 [Overlap]: for each value of the observable characteristics X_i of a household, we have a non-zero probability of participating or not in the behavioral program:
>
> $$0 < P_i = \Pr(D_i = 1 \mid X_i) < 1$$

This means that for every level of the observable household attributes, e.g., income, we should see a fraction of households greater than zero that participate in the program and a fraction greater than zero that does not. Overlap refers to the observables between the treatment and control group "overlapping" and not creating distinct partitions of the sample size. As we shall see below this ensures that when we observe a household that is treated we can find a household which is similar in terms of observable characteristics that we can compare it to the treated household.

In light of this assumption, a graphic example which satisfies the last assumption looks as follows:

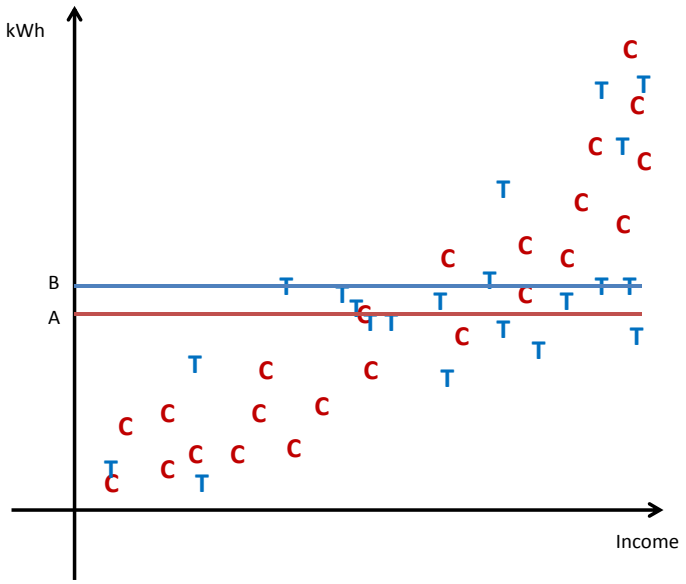

Figure 3-5
Electricity Consumption for Participating (Treatment) Households and for Nonparticipating (Control) Households

Households that participate in the behavioral program, denoted by T, are the treated observations. Households that do not participate in the program are denoted by C, are the control households. We plot household electricity usage of each household against household income. In this example, we compare the mean usage of the households which participate in the program with the mean usage of the households which do not participate in the program. We find that the electricity usage of the households participating in the program is actually somewhat higher than that of the households that are not part of the program, as the diagram illustrates. The program appears to have been more popular with high income than with low income customers (more Ts on the right-hand side of the graph) and as a result we may expect that customers with higher baseline electricity usage are more likely to have selected into the program.

However, this simple comparison of the mean usage between the two groups is misleading by not accounting for the fact that customers with different observables such as income selected into the program. Since program participation was not randomized it is impossible for us to quantify the causal effect of the program by comparing the average electricity usage for the two groups without employing a more advanced statistical technique, such as matching, to correct for the bias, which will be discussed below.

Using Additional Data

Before introducing the different methods to quantify the causal effects of programs where participation was not randomized, it is worth discussing the crucial role that additional data plays in our analysis. The assumptions introduced above and the resulting statistical methods require us to be able to compare the observable characteristics of households that participate in the program with the characteristics of the households that do not.

An overarching consideration is that the validity of the outcomes of the methods we employ depends on our ability to account for the factors that explain the households' decision to

participate in a program. While we will never observe the exact combination of factors which motivate households to participate in a program, in many cases we can remove the bias to the extent that we can control for a rich enough set of proxies to the underlying motivating factors.

The energy sector has a number of advantages in terms of data availability over other industries, which may potentially lead to more accurate evaluations of behavioral programs. While in many situations it is difficult to obtain data on program participants, utilities have detailed billing information for their existing customers. This includes at the minimum the service address and a history of past usage.

Based on the service address it is possible to populate a database with numerous variables describing the property, the demographics of the household and the neighborhood, countless measures of behavior and interests based on the household's purchases of different goods and services, political affiliations, subscription to different mailing lists and catalogs, etc. While it is easy to feel overwhelmed by the massive availability of data obtainable from third party providers, it is important to realize that a thoughtful selection of such variables can provide real value at relatively low cost. While, third party data aggregators typically take good care in collecting these additional variables, it important to keep in mind that these variables are not always perfect. In some cases they are very accurately recorded in administrative databases, while in others they are imputed and may be only approximate or outright false for a given household.

Let us review some of the useful pieces of information which can relatively easily be added to complement an existing customer database. Based on the service address it is possible to populate a database with property information based on public records. These data may include square footage, number of bedrooms, presence of a pool, etc. Additional real estate information is easily purchased from sources such as DataQuick or CoreLogic. As a quick search on Zillow.com shows, for any given property it is possible to obtain information on several variables characterizing it.

Using the customer address, a number of third-party data aggregators such as Acxiom offer a large menu of household level variables which characterize the demographic composition of a household. Additionally it is possible to learn about the previous purchasing behavior of a household. Participation in a behavioral energy program may relate to other purchases the household has previously made. Purchases also allow us to profile a household's preferences from revealed choices. For example, if we believe that attitudes towards the environment may explain the popularity of a program within a demographic segment it may be possible to proxy for these preferences using other purchasing data that are available. For example households which are more likely to sign up for a given behavioral energy program may also be more likely to make "green" purchases of environmentally friendly products in supermarkets. These households reveal a preference for the environment through frequent purchases of outdoor or camping products.

Another factor that may explain the households' decision to participate in a behavioral program may be captured by neighborhood characteristics. Neighborhood characteristics may include the distribution of income, professions, unemployment, race, or age. In some cases the similarity or dissimilarity of a given household to its neighborhood may also be informative. Political preferences and ideology may also be important and can be captured from previous local and national election results, party enrollment, voter registration and participation, or political activism. Neighborhood characteristics are easily available from the Census and other surveys

such as the American Community Survey. In most states voter registration, political donations, and election results are also easily available. Providers such as Aristotle can easily match a variety of political data for a given address.

The amount of data available on a given household can be difficult to handle. In practice it is important to rely on customer insights based on industry experience that can be backed up by statistical evidence. Statistical methods exist that can help identify a manageable subset of variables that explain customer signup decisions from a very large and comprehensive set of variables. The use of large datasets may seem cumbersome but we should remember that the bias in the evaluation of a behavioral program is only removed when we account for all relevant observables.

Let us not forget that the accuracy of the information is also important. Many unreliable variables will not deliver the solution we seek. In addition to the raw household data, a number of data providers also offer customer segmentations which are useful for providing a broad perspective on the differences between households. These segmentations are constructed by averaging over a large number of demographic indicators for the households and the neighborhoods. While these may be accurate in many areas, their use in the evaluation of behavioral programs has not been researched enough and their aggregate nature make them *a priori* suspect and they should be used with care. Until further evidence on their effectiveness emerges from the scientific community, they should not replace more precise household level measures if these are available.

As mentioned already, the electricity industry benefits from the availability of pre-program usage data. For most households historical monthly usage and billing data is available, usually for the past 1-5 years. Increasingly, high frequency usage data are also available from installed smart meters, with recordings at 15 minute or hourly intervals. Past usage (before the behavioral program was initiated) is a function of household characteristics and preferences. As such we can think of it as a convenient summary statistic that provides a baseline against which to compare households. There are numerous measures of past usage, from the mean and variance over a specified period, to the precise monthly usage. All of these may be used to compare households, and as we shall see below, match households in the treatment group to households in the control group. At the moment insufficient research exists that would provide guidance on how best to summarize past usage data in behavioral program evaluations. It is important to stress we are only referring to historical usage data before the program was initiated and not usage data collected after the program start. Usage data before the program start is already fixed and is not subject to the program itself. Once the program is initiated however, electricity consumption becomes an outcome subject to the program and it would be misleading to use it the same way we would use demographic variables, which do not change as a result of the program.

From a statistical perspective it is a best practice to identify the required set of information on the characteristics of the treatment and control households before the start of the program. During evaluation, the mean and variance of the observable characteristics for the two groups can be calculated. In a randomized trial environment we would expect these differences to be minor. Conversely, we expect to find more significant differences in a non-randomized setting where households are free to choose whether to participate in the program or not.

One way to quantify the statistical significance of the difference between the characteristics of the groups of households which participate in the program and those which do not, is to perform

pairwise tests, such as testing for the difference in means (or variances). The test for the difference in means is given by:

$$t = \frac{\overline{x}_C - \overline{x}_T}{\sqrt{\frac{\hat{\sigma}_C^2}{N_C} + \frac{\hat{\sigma}_T^2}{N_T}}},$$

where $\overline{x}_C, \overline{x}_T, \hat{\sigma}_C^2, \hat{\sigma}_T^2, N_C, N_T$ correspond to the means, variances and household counts of the control and treatment groups respectively. The statistical significance of this test can then be reported in the form of p-values. The p-values correspond to probabilities of obtaining a test statistic which is as extreme as the one computed "by accident". In general we consider the difference to be statistically significant if the p-value is less than 5%, meaning that there is a very low chance of the two means being the same.

As a simple illustration, consider the following figure from Harding and Rapson (2013) which compares participants and non-participants from a sample of households offered the opportunity to participate in a carbon offsetting program in California. The Climate Smart program was launched in 2007 and offered customers the opportunity to voluntarily offset the carbon emissions resulting from their consumption of electricity. Customers paid a small monthly surcharge on their electricity bill, which was used to fund local carbon offsetting activities such as preservation efforts or investments in methane capturing facilities. Information on the program was distributed to customers using a variety of channels and many customers chose to sign up for the program. Since this was not a randomization, the first question of interest is understanding what the differences are between customers who sign up for the program and those who do not.

Figure 3-6 reveals that participants in the program likely are statistically different than non-participants at a high degree of confidence because the p values are almost all close to zero (many of these are so small that they were rounded to zero). By looking at the differences in characteristics we can profile adopters. Adopters use less electricity, are older, wealthier and have smaller family sizes. They are more likely to be involved in environmental activities, concerned with wildlife and enjoy camping. They also tend to live in older, smaller but more expensive homes without a pool. Note that the household groups are not significantly different along a couple of dimensions. This is to be expected since we would not expect them to be different along every single dimension, but as long as the vast majority of the attributes are different we can safely conclude that adopters have a different statistical profile than non-adopters.

This example is consistent with the premise that when households are free to decide whether or not to participate in a program, households with different demographic characteristics are likely to select themselves into the treatment and control group based on those characteristics, and therefore not in a way that replicates random assignment. This means that the two groups are "unbalanced" from an observational point of view and requires methods such as the ones described below to restore balance and allow for the comparison of the effect of the program on the different groups. Recall that in a randomized setting we would have expected all means for the variables reported in the table to show no statistically significant difference.

	E1 Rate Customers				
	Adopters (N=9,445)				
	Non-adopters (N=22,895)				
	Mean		SD		p-value
	Non-adopters	*Adopters*	*Non-adopters*	*Adopters*	
(Average) kWh*	635.507	530.212	360.500	309.619	0.00
(Average) Bill*	109.731	82.263	97.345	73.102	0.00
Age	56.545	59.093	14.306	15.418	0.00
College	0.294	0.268	0.456	0.443	0.00
HHIncome $80k+	0.480	0.528	0.500	0.499	0.00
Children	0.471	0.382	0.499	0.486	0.00
Working Woman	0.467	0.412	0.499	0.492	0.00
HH Size	2.877	2.582	1.471	1.374	0.00
Home Owner	0.966	0.958	0.180	0.201	0.00
Environmental	0.097	0.182	0.296	0.386	0.00
Green Living	0.615	0.622	0.487	0.485	0.33
Charity	0.370	0.483	0.483	0.500	0.00
Charitable	0.547	0.500	0.498	0.500	0.00
Outdoors	0.551	0.634	0.497	0.482	0.00
Wildlife	0.073	0.130	0.261	0.336	0.00
Camping	0.266	0.322	0.442	0.467	0.00
Home Age	38.547	48.472	23.403	25.809	0.00
Heating	0.273	0.264	0.446	0.441	0.11
Cooling	0.111	0.061	0.314	0.239	0.00
Sqft 2500+	0.161	0.128	0.367	0.334	0.00
Home Value $500k+	0.161	0.186	0.368	0.389	0.00
Pool	0.139	0.087	0.346	0.283	0.00
ClimateZone	6.670	5.180	4.459	3.915	0.00

* The average kWh and bill amounts for adopters are computed for the pre-adoption periods

Figure 3-6
Comparing Demographic Characteristics for Participants and Non-participants. P-values less than 0.05 indicate that the households are significantly different along a particular demographic dimension. (Note: Many of the variables are binary at the household level, e.g., homeowner. As a result the group means correspond to the fraction of households with that characteristic.)

In this section we have introduced some of the fundamental statistical concepts that define what we mean by a "causal effect" in a rigorous scientific study. The main insight is that in the absence of perfect randomization, the treatment effects estimated from the data will be biased; we call this "selection bias". Households typically signing up for a program have different characteristics than household choosing not to. As a result, a direct comparison of the electricity consumption of households participating in a program with that of households not participating is going to be very misleading. In order to address this problem, we need additional data. In particular, we need to capture the characteristics that distinguish participating households from those who choose not to participate in a behavioral program. In the next section we will look at some of the main statistical methods which use the additional data to estimate the causal effects of interest in behavioral programs without randomization.

4
OBSERVATIONAL APPROACHES TO PROGRAM EVALUATION

Our discussion above shows that when we do not have randomization, comparing the average outcome for the households participating in the behavioral program with the average outcome for the households that are not participating will not produce reliable results. This is a result of the fact that households are able to decide whether or not to participate in the program, which leads to a selection problem. The best way for estimating the causal effect is to compare the observed outcome with the counterfactual outcome. Unfortunately, that is not possible since only one outcome is observed. Below we shall review the three main approaches developed in the field of statistics to solve this problem: matching on covariates, propensity score matching, and propensity score weighting.

Before we proceed, it is worth reemphasizing that whenever you deviate from an RCT, as we do with all the situations discussed in this report, you must be prepared to identify all important observable covariates, meaning those variables that predict and describe how the participants differ from the non-participants. It is also worth emphasizing that all of the following evaluation approaches described require this additional data for both the participants, *as well as* for the non-participants. The ideal situation is to have these data on all the non-participants, although this is not always possible or practical. Regardless, the data requirements need to be considered at the outset. They may not be insubstantial, and in some cases, they may make an RCT a better way to proceed.

Matching on Covariates

The idea behind matching is that the missing data problem can be potentially addressed by an imputation algorithm, which develops a counterfactual, under the assumptions discussed above. To illustrate the mechanics of the matching algorithm let us consider a very simple example where we attempt to compute the ATT for the data presented in Table 4-1.

Note the difference between computing the ATT and the ATE is that for the ATT, we only need to find matches for the treated households, while for the ATE we need to find matches for both the treated and the control households. Depending on the question we are interested in, we will compute either the ATT or the ATE. Recall that the ATE computes the causal effect for all households, while ATT only computes it for the treated group. Which causal effect we choose depends on whether we think that participation in program once the program is scaled up will be broad or narrow. If we think that the program is only appealing to a small number of all customers, then it is indicated to compute the ATT, since the ATE includes households that would never be subject to the treatment.

Below we show how to compute the ATT but the principle is identical for the ATE. The number of steps for the ATT is smaller since we only need to find matches for the treated households and it is thus easier from an expositional point of view. In practice computer software will easily perform the required matching operations for either causal effect.

In practice this opens the question as to what the control group is in a real-world program. In principle we are interested in all non-treated households. If the number of households not participating in the program is very large this may not be very practical, however. We may not wish to collect observables or conduct matches on every single household not participating in the program. Thus in practice, we choose a random sample of the households not participating in the program as our control group. Since we may wish to find more than one match for every treated household, it is common practice to choose a sample of non-participating households which is several times (usually 3-5 times) larger than that of participating households. It is important to choose this sample at random, e.g., using suitable database software, to make sure it is representative of the population of households not participating in the program.

In Table 4-1, we are imagining a behavioral energy efficiency program which is offered to a group of 10 households. Four of these households sign up and participate in the program, and six do not (indicated by a 1 and 0 respectively in the second column labeled "Participate"). Again, one of the features of this approach is the requirement that data is available for households that *do not* participate in the program, as well as those that do.

We are interested in measuring the effect of the program in the month after the program was offered. For each household we also measure household income. First we compare the average household income for households that participate in the program with that of households that do not. We find that on average, participating households have higher income than non-participating households.[5] If we assume that income is the only variable which explains why some households participate in the program and others do not, we can proceed by matching households on income.

Table 4-1
Computing ATT by Matching on Income

Household	Participate	Income $'000	kWh		Match	Difference
			Y(1)	Y(0)	Y(0)	
1	1	62	300	?	350	-50
2	1	120	400	?	375	25
3	1	80	500	?	600	-100
4	1	90	600	?	650	-50
5	0	60	?	350	-	-
6	0	105	?	250	-	-
7	0	125	?	500	-	-
8	0	80	?	600	-	-
9	0	87	?	650	-	-
10	0	30	?	150	-	-

Our outcome of interest is monthly usage measured in kWh. For each household we observe either Y(1) or Y(0) depending on the participation status. The counterfactual usage that we would like to know but is not available, is denoted by "?". In order to compute the ATT, we need

[5] In this example we look at whether there are differences in means, but ignore whether these differences are actually statistically significant. In general we should employ the difference in means test described in the previous section to determine whether a difference is actually statistically significant or not.

to impute the missing values of Y(0) for the participating households. Let's consider the household 1. It has an income level of $62,000. In the sample we also observe household 5, which is not participating in the program with an income level of $60,000. Since this is the household in the control group with income closest to our treated household we will impute the value of Y(0) for household 1 with the corresponding value for household 5. Thus, household 5 is matched with household 1. Now consider household 2. In the control group we find that both household 6 and household 7 may be reasonable matches for this household. In this case we should consider the average of the two households and use the average as the matched value of Y(1) for this household.

We can proceed by finding matches for each treated household. Later in this section we will show how the matches can be found using a more rigorous algorithm. Note that in this example we have found more than one match for one household. At the same time, household 10 which did not participate in the program was not matched to any participating household because of its low income.[6]

If we now compute the pairwise difference for the participating households between their usage and the counterfactual usage if they had not participated (obtained by matching) we would find that on average the program reduced usage by an average of almost 44kWh per household.

$$Average: Y(1) = \frac{300+400+500+600}{4} = 450$$
$$Average: Y(0) = \frac{350+375+600+650}{4} = 493.75$$
$$ATT = 450 - 493.75 = -43.75(kWh)$$

On the other hand, if we compute the naïve estimator for the ATT using the approach described in Section 3, we find that the program *increased* usage by 33.3 kWh:

$$Average: Y(1) = \frac{300+400+500+600}{4} = 450$$
$$Average: Y(0) = \frac{350+250+500+600+650+150}{6} = 416.67$$
$$ATT = 450 - 416.67 = 33.33(kWh)$$

In Figure 4-1, we return to the stylized example in Figure 3-5 and show how after matching, the computed ATT changes sign compared to the naïve approach of comparing participating and non-participating households. Households in the control group which are not matched are found in the lower tail of the income and usage distribution (these are lightly colored and denoted by ~~C~~). They are ignored in the analysis.

---

[6] Note that if we wanted to compute the ATE, we would perform the same matching procedure for the control households and match them to the treated household, thereby imputing their corresponding values of Y(1). The ATE is then evaluated as the average of the differences between the observed Y(1) or Y(0) and the corresponding imputed values of the counterfactual for the whole sample. The principle is the same as the one presented above for the ATT, but needs to be executed for every single household.

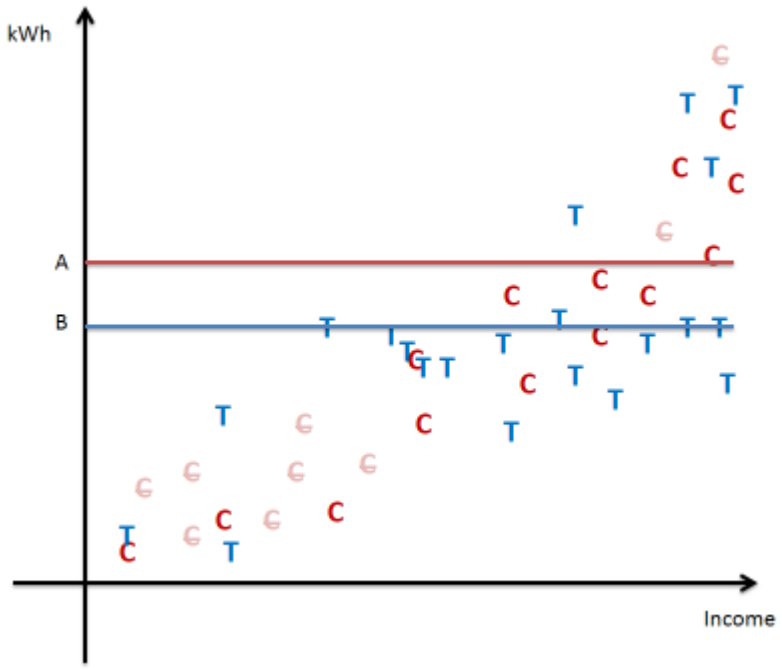

Figure 4-1
ATT with Matching

This discussion, while intuitive, does pose an interesting technical challenge. How do we decide what the best match or matches are to a given household? The traditional approach is to use a distance measure to define the nearest neighbor or neighbors to a given household in the space of observable covariates. In Figure 4-2 we consider the case of one treated household and 3 control households. For each household we observe two different variables which characterize each household (e.g. income and house size). In the space defined by the two variables we can compute the Euclidean[7] distance between the treated household and the control households. We find that $d_1 < d_2 < d_3$. In practice we have to decide how many nearest neighbors to include in the matched set. When the pool of control households is large and a number of matches can be found, it is advisable to use more than one match in order to improve the accuracy of the estimate.

[7]The Euclidean distance is the "usual" distance we use in everyday life, which satisfies the Pythagorean Theorem. Here it measures the distance between observations with the coordinates defined by the values of the observable household characteristics.

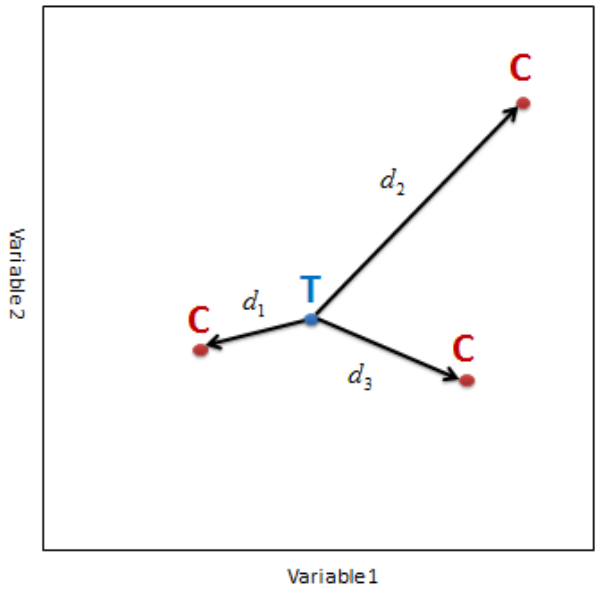

Figure 4-2
Matching on Two Variables

Statisticians have developed a number of different metrics that can be employed to measure distances between households when multiple variables (defined in different measurement units) are available which capture different attributes of the households. The most popular are the Euclidean and the Mahalanobis distances. Expressed in vector notation the Euclidean distance is given by:

$$d_E(x_T, x_C) = \sqrt{(\mathrm{x}_T - x_C)'(\mathrm{x}_T - x_C)}.$$

Here we have employed linear algebra notation to denote the dot product between the vectors of observables for each household. In the two variable case, $x_T = (\mathrm{v}_T, \mathrm{z}_T)$ and $x_C = (v_C, z_C)$ and $d_E(x_T, x_C) = \sqrt{(v_T - v_C)^2 + (\mathrm{z}_T - z_C)^2}$.

Let us consider a simple example. Household T has income (in $'000) of 50 and lives in a house valued at 400. Household C has income 75 and lives in a house valued at 650. The Euclidean distance between these two households, as measured by the two attributes is given by

$$d_E = \sqrt{(50-75)^2 + (400-650)^2} = 251.25$$

Notice that this measure depends on the units in which the variables are measured, and does not take into account that the variables can be correlated with each other. Statistical software typically uses a modified version of this measure called the Mahalanobis distance, which is given by:

$$d_E(x_T, x_C) = \sqrt{(\mathrm{x}_T - x_C)'\mathrm{S}^{-1}(\mathrm{x}_T - x_C)}.$$

What this expression says is that the Mahalanobis distance takes into account the variance-covariance matrix S between the variables and adjusts the Euclidean distance by the extent to which observables are correlated with each other. This computation is difficult to perform by hand by is easily handled by a computer.

As previously mentioned, one of the advantages of evaluating behavioral energy programs is the availability of usage data for households before the start of a program, which can be used to match households. Care should be taken, however, since households may exhibit anticipation effects. In some cases households expect the introduction of a given program and may start changing their behavior before the program officially starts, anticipating to participate in the program later on. In other cases it may be that a change in usage behavior is what induces the household to sign up for the program in the first place. Harding and Rapson (2013) show that households attempt to engage in energy conservation before signing up for a carbon offset program and may potentially view the carbon offset program as an easier way of achieving their desired environmental goal without having to reduce consumption. This suggests that when using pre-adoption usage, it is important to use a long period of time and match on the series of consumption as opposed to the consumption in one single period.

One of the challenges of using matching is that it is natural to want to include as many variables as possible in order to best account for the factors which determine participation in our behavioral program. In fact, we have established that all relevant covariates (observables) have to be included to avoid bias. Once we include information on pre-program usage, the set of observables can become very large since we can treat the observations in every single period as an observable attribute of the household, rather than the average over the period. This could make it very difficult to find good matches for the treated observations. Households may be close on some dimensions but quite different on other dimensions. This is sometimes referred to as the "curse of dimensionality" that arises in matching. Propensity score methods were developed to deal with this problem.

Matching on the Propensity Score

One important contribution to the statistics of matching was the introduction of methods based on the propensity score (Rosenbaum and Rubin, 1983). The propensity score is the probability that a household is treated, i.e. participates in the program. It takes on values between zero and 1. Instead of attempting to match households in the multidimensional space corresponding to the potentially large number of observables for each household, we can compute the propensity score and then match based on the propensity score. This reduces our problem to a single dimension which facilitates making judgments about alikeness.

In Figure 4-3, we show a stylized representation of the process of constructing the propensity score. In this example all observations, corresponding to both treated and control households have two variables measured, the household income and the house size (in sqft). We can now use a statistical model which takes as its input the two variables characterizing these attributes and map them on to the propensity score, which is bounded between 0 and 1 and measures the probability that a given household participates in the program.

$$P_i = \Pr(D_i = 1) = G(x_i)$$

The propensity score is a function G(.) of the variables x_i which are observed for each household. We can use a standard binary choice model such as the logistic model to compute the propensity score.

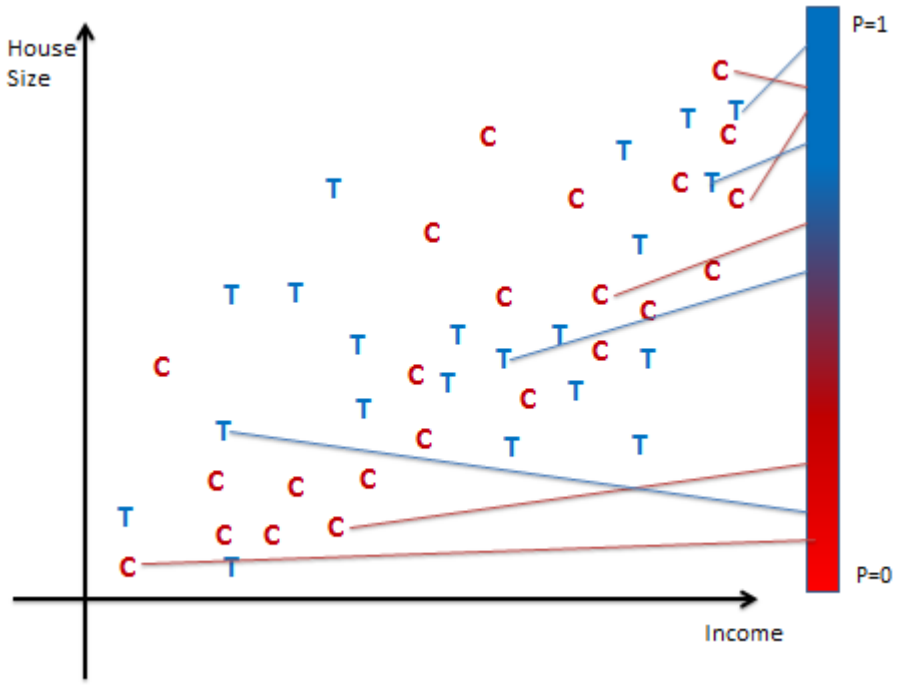

Figure 4-3
Constructing the Propenstity Score

As before, when constructing the propensity score it is crucial to choose factors that are likely to be important in explaining the households' selection into the program. For each behavioral program, we have to try and understand the motivations behind the decision to participate and what observables are available to provide proxies to these explanatory factors.

Once the propensity score for each household has been estimated from the data, we can use it to compute matches and then proceed to evaluate the causal effect of interest.

The most common approach to using the propensity score to generate matches is to compute nearest neighbor matches based on it. In Figure 4-4 we show how one would determine the nearest neighbor match for a two treated households. The advantage of using this method is that instead of having to find matches between observations characterized by a large number of characteristics, we are now dealing with a problem where we only need to find the nearest neighbor to an observation along the propensity score dimension. We can think of lining up all households in ascending order of their propensity score from 0 to 1 and for each treated household we choose the control household with the closest propensity score value as its match.

Figure 4-4
Matching on the Propensity Score

In practice, we have a number of options available to improve the performance of the matching algorithm. We have the option of obtaining more than one match for each household. One common procedure is to define a fixed radius or "caliper" and use all the households that fall within that radius as matches for a given observation. It is generally advisable to match "with replacement", meaning that the same household may be used as a match more than once (as shown in Figure 4-4). Otherwise, the outcome of our matching may produce results which depend on the order in which the households are being matched. If we choose a large radius it may also be advisable to construct a weighted average of the outcomes for the matched households where the weighting function depends on the distance from the propensity score of the household for which the match is constructed.

Many of these matching varieties are available in standard statistical packages. Naturally, they also require us to make choices over their implementation, e.g. choosing the radius in propensity score matching and when implementing such a method is important to show that the results are not driven purely by these choices. We would like the estimate effects to be similar for different matching algorithms and parameter choices. Unfortunately, there is no clear guideline to which approach to choose and the academic literature has shown that all methods imply trade-offs in terms of bias and precision (Guo and Fraser, 2009). In particular we know that using a larger number of matches may improve precision but at the cost of introducing bias by comparing households which become less and less comparable as the number of matches increases. It is advisable to report a number of different matching estimates based on a different number of matches. If the results are robust there should only be minor differences between the obtained numbers.

Propensity Score Weighting

An alternative to matching is using the propensity score to construct weights. Since the issue that we are addressing is the lack of balance between households in the treatment and control group, this idea seems natural. Weighting can be implemented in a two-step procedure. First, we can compute the propensity score using a standard binary choice model such as the logit:

$$P_i(x_i) = \Pr(D_i = 1) = \frac{\exp(\sum_{k=1}^{K} x_{ik}\,\beta_k)}{1 + \exp(\sum_{k=1}^{K} x_{ik}\,\beta_k)}$$

for a vector of observable characteristics x_i . In this expression, exp denotes the exponential function. For notational simplicity we assume that for each household we observe k different characteristics. The coefficients β_k are then estimated from the data and can be thought of as

weights on the different characteristics which best predict the probability that a household participated in the behavioral program. Once these weights have been computed from the data, the predicted propensity score $\widehat{P}_i(x_i)$ is computed for each household. The predicted propensity score is a function of the data x_i.

In the second step we use the estimated propensity score to compute observation weights which are inversely proportional to the propensity scores. The idea is to restore the balance between the groups by weighting the households by the inverse of the probability of being in their respective group. Denote the household specific weights by w_i , then

$$w_i = w_i(D_i) = \begin{cases} \dfrac{1}{1-\widehat{P}_i(x_i)} & if \quad D_i = 0 \\ \dfrac{1}{\widehat{P}_i(x_i)} & if \quad D_i = 1 \end{cases}$$

This equation says that for treated households, the weight is given by 1 divided by the propensity score. For untreated households it is given by 1 divided by 1 minus the propensity score.

The estimator for the ATE is given by:

$$\widehat{ATE} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{D_i Y_i}{\widehat{P}_i(x_i)} - \frac{(1-D_i)Y_i}{1-\widehat{P}_i(x_i)}\right).$$

If we wish to compute the ATT instead we the required are:

$$w_i = w_i(D_i) = \begin{cases} \dfrac{\widehat{P}_i(x_i)}{1-\widehat{P}_i(x_i)} & if \quad D_i = 0 \\ 1 & if \quad D_i = 1 \end{cases},$$

and the estimator for the ATT is given by:

$$\widehat{ATT} = \frac{1}{N}\sum_{i=1}^{N}\left(D_i Y - (1-D_i)Y_i \frac{\widehat{P}_i(x_i)}{1-\widehat{P}_i(x_i)}\right).$$

The intuition here is relatively simple. The naïve estimator which compared the average of the outcome for the treated households with that of the untreated households is biased because it fails to take into account the imbalance between the different groups of households. Once we determine the magnitude of that imbalance as measured by the propensity score we can use it to reweight the observations. Thus while the raw difference in averages was biased, the weighted difference corrects for this bias.

Weighting on the propensity score also provides a convenient approach to incorporating matching in a traditional regression approach. We saw earlier that

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) = Y_i(0) + D_i (Y_i(1) - Y_i(0))$$

This suggests that in a randomized experiment the following regression will provide an unbiased estimate the average treatment effect (δ_1):

$$Y_i = \delta_0 + \delta_1 D_i + \varepsilon_i \ .$$

In a non-randomized behavioral program, this will produce an inconsistent estimate. Using the propensity score methodology we can estimate the model using the following weights in the regression analysis:

$$w_i = \sqrt{\frac{D_i}{\hat{P}_i(x_i)} + \frac{1 - D_i}{1 - \hat{P}_i(x_i)}} \ .$$

This approach can be particularly useful when we need include additional controls in our regression, such as variables designed to capture seasonal variations, or weather effects. If we label these additional variables Z_i , then the resulting estimating equation is:

$$Y_i = \delta_0 + \delta_1 D_i + \delta_2 Z + \varepsilon_i \ ,$$

where we can use the same weights w_i from above.

Can We Match When No Information is Available Regarding the Non-Participants?

In some cases, no information about the non-participants is available. This often occurs when a program was implemented some time ago without randomization and the original design did not involve a control group. In such cases it is generally impossible to sample non-participants and attempt a matching strategy such as the one described above. Under certain conditions it may, however, still be possible to recover an estimate of the ATT.

If households enter the program at different points in time, it may be possible to use the variation in the timing of adoption to measure the impact of program participation. The intuition behind this result is that at a specific point in time we can compare a household that is part of the program with another household which is not part of the program but will be part of the program later on. If the timing of adoption can be thought of as random then this is a valid identification strategy. Such an approach was implemented by Harding and Hsiaw (2013) to analyze the effectiveness of a behavioral energy conservation program in Northern Illinois for which no control group was available. This program was advertised through a variety of media channels including TV and Internet and participation was open to all residents in the relevant geographic area. Households were able to sign up for the program over the next 12 months; sign-ups over that time are shown in Figure 4-5.

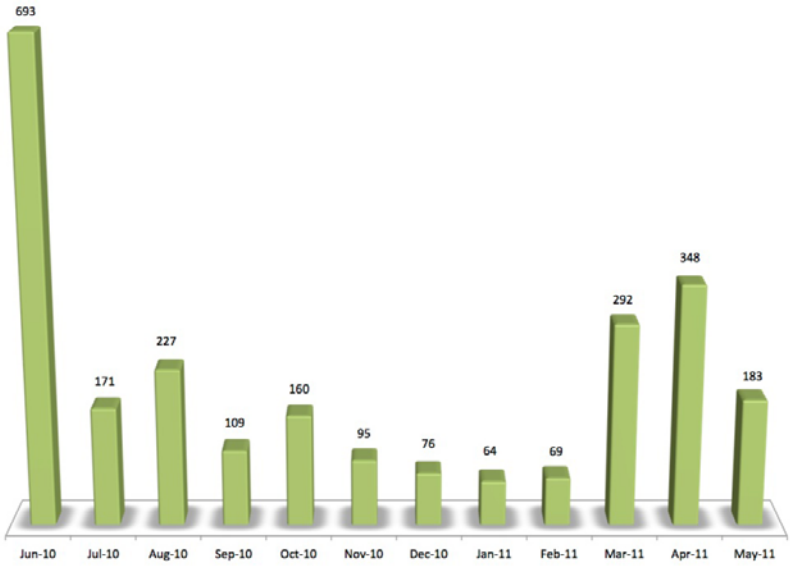

Figure 4-5
Signup for a Behavioral Energy Program Over a 12 Month Period

The number of households signing up for the program varies from month to month with the recruitment effort and amount of exposure to the program that the households had access to. Since the program launch received extensive media exposure the number of households signing up for the program is large. If we are willing to assume that the characteristics of the households signing up at different points in time are the same then we can compare the electricity usage in July of households signing up in June with that of households signing up in August.

In order to measure the effect of the program in the month immediately after signup, we can proceed as follows. We can define the variable D_{it} to be equal to 1 in the month immediately after the household signs up for the program and 0 in all previous months. Thus, in the example above, households signing up in June have a value of 1 in July and 0 otherwise. Households signing up in August have a value of 1 in September and 0 otherwise. The short-run impact of the program can be estimated by the following regression equation:

$$Y_{it} - Y_{it-1} = \delta D_{it} + \varepsilon_{it} \ .$$

To give an example, consider the month of June. We look at the average outcome of households which have signed up in May and compare their electricity consumption with that of households which will sign up later on in the year but have not signed up by June. In effect these households are the control households. Then we repeat the process for July, August, etc.

Note that this approach requires a bigger leap of faith and acceptance of assumptions that can't be verified in practice. Using this approach requires that the analyst can convince herself that the sequencing of subscription is truly random. Whether the reader believes that will depend on the degree to which the analyst makes a sound case.

It may be possible that individuals signing up at different points in time may have different characteristics. If these differences are relatively minor, in the sense that the observed characteristics of households signing up at different points in time are fairly comparable, then this approach can be combined with matching to reduce the estimation bias. For example, it may

be that younger households are more likely to sign up in the early stages of a program because they are more involved with the different media channels. Harding and Hsiaw (2013) show how to employ matching in this situation following the approach outlined by Sianesi (2004). It consists of comparing households signing up for the program in one month with similar households signing up later on where similarity is determined by matching on observable characteristics. The paper also shows how the analysis can be extended to measure the medium-run effectiveness of the program over an 18 months period.

5
IMPLEMENTATION AND TESTING

Now that we have introduced the main statistical approaches designed to estimate causal effects when no randomization is available. it is worth revisiting the assumptions on which these are based.

First, we see the crucial role that the first assumption plays, which says that the selection process can be characterized by observable attributes of the households. Both matching and propensity score analysis rely heavily on this assumption and use the observable attributes to "balance" the treatment and control groups before comparing the outcomes of interest. If this assumption fails, so will our attempt to restore balance.

The second assumption states that the outcomes are independent of each other and the pattern of enrollment. The methods discussed herein treate each household as independent of any other household. In particular the decision to participate in the behavioral program for a household was not determined by the participation decision of any other households. If this assumption is violated our procedures will produce biased estimates since they don't take into account the complex relationships that may emerge when households jointly decide on participation.

First, it is important to note that these two assumptions are not directly testable. The second assumption relates primarily to program design and implementation and a careful consideration of the program may indicate how likely it is to hold in practice. It also cautions us to think about the potential long run implications of the program when implemented at full-scale.

While we cannot directly test the first assumption, we can follow a number of best practices to ensure that it is likely to hold in practice. A number of tests have also been developed to ensure that we guard against likely failures of this assumption.

Understanding the process that leads to some households selecting into the program is at the very core of the analysis. It is important to explore this from all possible angles and build a comprehensive understanding of customer behavior. Some of the factors that should be considered are:

- What marketing approaches are used to recruit customers into the program?
    - How were the program benefits and costs framed?
    - What were the channels through which customers were reached?
- What are the geographic or institutional factors that make the program more appealing to certain customer segments?
    - Climate
    - Rates
    - Other utility programs that are be thought of as complements or substitutes to a particular program
- If technology is part of the behavioral program, what role does it play in program adoption?
    - Are there any technological constraints to eligibility, e.g. type or age of home?

- Are there hidden costs in terms of installation time or learning to use the technology that create differential propensities for adoption across demographics?

A clear understanding of the factors that drive adoption helps select the most important variables that will be used in the matching process and also highlights the potential for bias if certain important factors are not easily measured in the available data. In such cases it is important to think of proxies that account for the factors thought to be important.

In addition to a careful analysis of the program implementation details, a few simple rules should be followed when choosing the additional data (the observables) that is to be used in the analysis in order to maximize the chance that it satisfies the assumptions we have made. It is important to make sure that the variables selected for inclusion in the matching process are measured using similar instruments (survey or third-party) for both the treatment and control groups. Furthermore, the variables should only denote outcomes that at predetermined at the start of the program and not in any way outcomes of the program itself. It is best if the measurement of these variables occurs before the program start to avoid contamination with information that relates to program outcomes. Both continuous and discrete variables can be employed.

When a large number of variables is available, statistical tests can be used to select the subset of variables which best explains the selection into the program. The inclusion of irrelevant or only weakly related variables may potentially decrease the precision of the matching process.

Once the choice of variables was made we can perform balancing tests such as the comparison of means described above for the new matched or weighted groups. If the variables are chosen correctly we should not see any statistically significant differences in variables denoting household characteristics that are not program outcomes. In our stylized example, once we have chosen appropriate match variables we should not find a statistically significant difference between the mean income and house size of the matched or weighted treatment and control groups.

If statistically significant differences remain, we need to reconsider the selection process and investigate other variables which may better characterize the selection process and have higher predictive power to explain the propensity to participate in the behavioral program. If this is not possible, or if statistical differences still remain, this might lead us to conclude that selection is so unbalanced that any measurement of effects is too biased to be credible. This may seem like admitting defeat (or gross negligence in the study design), when in fact it is better to acknowledge that the results have little probative value than convey them as though they merit use in making important decisions about program implementation.

The third assumption is more subtle. It relates to the propensity score, and says that conditional on any set of observable characteristics, we observe both participating and non-participating households with positive probability. In Figure 4-5 we present a case where this assumption is violated. We plot the density of the propensity score for both participating and non-participating households and we notice that their distributions do not overlap. This means that in the middle of the distribution of the propensity scores we have numerous households which share common characteristics. For these households it is easy to find matches between the two groups. The area of overlap is called the common support.

The figure shows, however, that there are tails of the distribution of the propensity scores where no overlap exists. On the left side we have households which do not participate in the program and which have characteristics that make them very different from the participating households.

On the right side we have households participating in the program for which there are no comparable households. We can think of these households as being those who find the program so appealing that it is difficult to find households with similar characteristics that would not also participate in the program. While we can proceed to evaluate the causal effects for the area of common support, our results will no longer be externally valid. This may or may not be a big problem for the evaluation of energy programs depending on the type of households which are included in the common support.

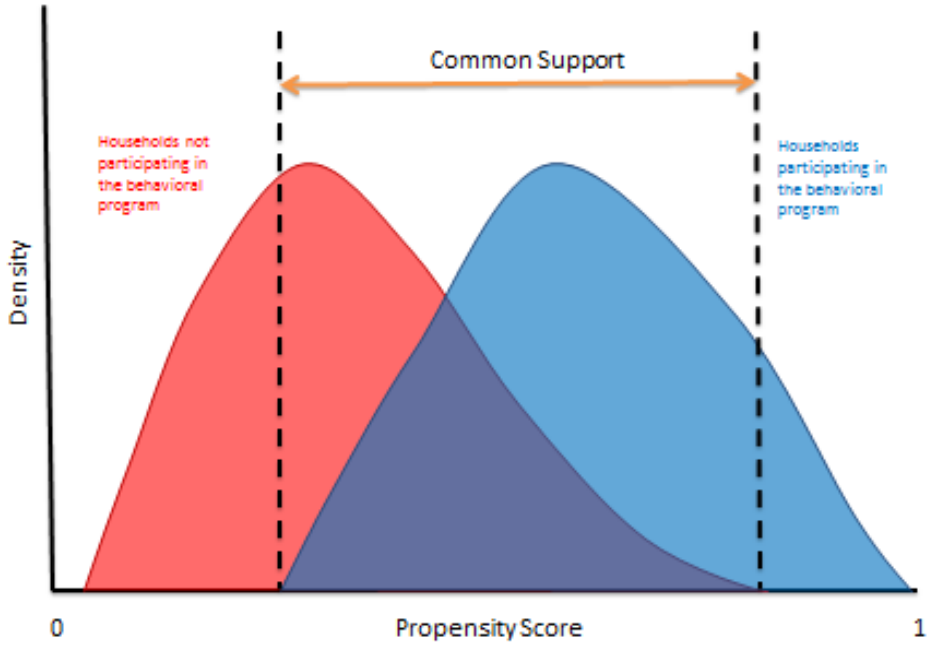

Figure 5-1
Common Support in Propensity Score Analysis

In practice, it is a good idea to plot the distribution of the propensity scores for the two groups and visually inspect the degree of overlap. In fact, we can plot the distributions of the propensity scores for the groups before and after matching. Not only should we observe the distributions overlapping but we should also be possible to see the effect of matching directly. Before matching the two distributions will exhibit noticeable differences in shape. After matching the distributions will be much more similar. It is possible to test for the statistical significance of the difference between the distributions using standard tests such as the Kolmogorov-Smirnov test available in most statistical packages.

If we determine that the common support does not cover the entire range of observations, it is common practice to simply discard observations above and/or below a certain threshold where overlap does not hold. This approach will produce a valid estimate of the causal effect subject to the implications of the data censoring, but may lack external validity since it only holds for the subpopulation defined by the characteristics, which are found to be common to the treatment and control groups.

Lastly, as the above discussion shows there may not be a single answer to the problem of estimating the causal effect of interest. Disagreement over the precise selection mechanism or the variables that should be included in the analysis may persist. It is thus good practice to perform a

sensitivity analysis, whereby the results are re-derived under different sets of variables. Ideally our results will not be too sensitive to the inclusion or exclusion of small subsets of the variables used to construct the propensity scores and quantitatively similar results for the estimated causal effects will be obtained.

We summarize the different steps involved in the analysis in the following flowchart (Figure 5-2). Analyzing a behavioral program without randomization starts by determining the observables driving the sign-up process and explaining the differences between the treated and control groups. If the number of variables is small they can be used to find matches for the treated and control households. If the number is large the propensity score can be computed and used either directly in matching or to compute observation weights. Either approach will produce estimates of the causal effects ATE and ATT. While these methods rely on assumptions which are not completely testable in practice, additional sensitivity analysis should be performed to confirm that the results are robust to different factors that went into the analysis such as the choice of observables used to construct the propensity score or the number of matches used in the analysis.

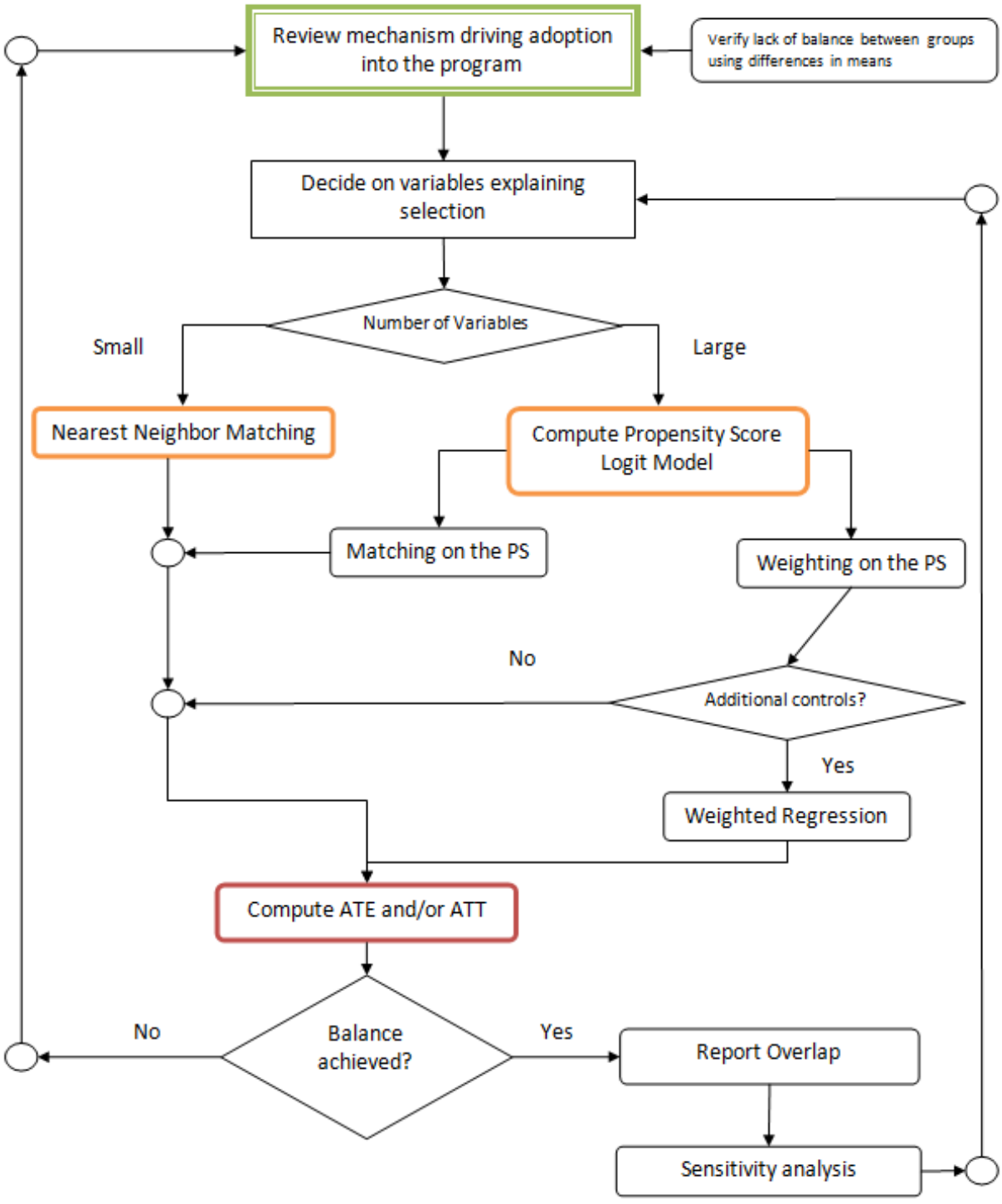

Figure 5-2
Implementation Flowchart

6
CONCLUDING COMMENTS

The aim of this report is to introduce utility practitioners to a broad set of methods developed by statisticians and econometricians, which enable us to evaluate the success of behavioral programs. We are particularly interested in techniques that can be used when these programs have gone full-scale. While randomized trials remain the gold standard of program evaluation, a number of different techniques are available to estimate the effectiveness of these programs when randomization is not possible because it is impractical, or even when randomized experiments suffer from design flaws which lead to imperfect randomizations.

The methods discussed in this report aim to measure the causal effect of the program. The notion of causality is introduced within the framework of the Rubin Causal Model. Consider the case of a household participating in a program. The Rubin Causal Model says that for every outcome there is a counterfactual outcome corresponding to what the household would have done if it had not signed up for the program. Similarly for a household not participating in a program we have to consider what the outcome would have been if that household had signed up for the program. The fundamental problem of evaluation is that for each household we only observe outcomes in one of the states since the household either participates or not in the behavioral program.

If we can select observables such as customer demographics, which capture the differences between customers who choose to participate and those who do not, then a range of statistical techniques involving matching are available to us. These methods were designed to eliminate the selection bias inherent in comparing the outcomes of groups of households which are different along a number of different dimensions. Matching and propensity score methods are widely available in contemporary statistical software packages such as SAS or STATA and their implementation is not too challenging computationally. It is important to always be cognizant of the fact that the methods rely on untestable assumptions about household behavior. If the analyst can make a persuasive argument that the underlying assumptions are correct then the results of these methods will be credible.

Academic research has not yet shown that any one method strictly outperforms all other approaches. Different methods tend to perform better or worse depending on the analysis conducted and the available data, including how large the sample is and the quality of the additional data on demographics and customer behavior included in the analysis.

Additional research is needed in the electricity sector to evaluate the performance of these methods and provide direct evidence on the success of different approaches in the types of data usually encountered in the evaluation of utility behavioral program.

The ideal environment for future research would consist of a program which was initially implemented as a randomized experiment and the evaluation of which can be used as a reliable baseline. Building from this we can conduct a series of statistical analyses where we attempt to evaluate the program without using the “true” control group. Instead additional variables can be collected on a sample of households not included in the RCT and an analysis can be constructed using the methods outlined in this report. Since utility data has many features not present in other economic data, this analysis would provide valuable evidence on the performance of these

methods in the utility context. By comparing the estimated causal effect of the program using techniques such as matching to the causal effect measured from the RCT, the degree to which approaches designed for non-randomized programs can be formally evaluated. Ideally this research would show that with sufficient data, methods such as matching can replicate the results obtained from a RCT even when no randomized control group is available. We do not expect the methods to work perfectly in practice and this research would help quantify how well they work and what potential issues may arise when applying them to utility programs.

A
REFERENCES

Aigner, D. and J. Hausman 1980. Correcting for truncation bias in the analysis of experiments of time-of-day pricing of electricity. *Bell Journal of Economics* 11(1), 131-142.

Caliendo, M. and S. Kopeinig 2008. Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys* 22(1), 31-72.

D'Agostino, R. 1998. Tutorial in Biostatistics: Propensity score methods for bias reduction in the comparison of a treatment and a non-randomized control group. *Statistics in Medicine* 17, 2265-2281.

Dehejia, R. and S. Wahba 2002. Propensity score-matching methods for non-experimental causal studies. *Review of Economics and Statistics* 84(1), 151-161.

Gerber, A. and D. Green 2012. Field experiments: Design, analysis and interpretation. *Norton.*

Gertler, P. et. al. 2011. Impact evaluation in practice. *World Bank*.

Guo, S. and M. Fraser 2009 Propensity score analysis: Statistical methods and applications, *Sage*.

Harding, M. and A. Hsiaw 2013. Goal setting and energy conservation. *Journal of Economic Behavior and Organization*, under revision.

Harding, M. and D. Rapson 2013. Do voluntary carbon offsets induce energy rebound? A conservationist's dilemma. *Journal of Environmental Economics and Management*, under revision.

Imbens, G. 2007. Estimation of average treatment effects under unconfoundedness, *NBER*.

Manski, C. 2013. Public policy in an uncertain word: Analysis and decisions, *Harvard.*

Morgan, S. and C. Winship 2007. Counterfactuals and causal inference: Methods and principles for social research. *Cambridge*.

Rosenbaum, P. and D. Rubin 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70, 41-55.

Rubin, D. 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology* 66, 688–701.

Rubin, D. 1977. Assignment to a treatment group on the basis of a covariate. *Journal of Educational Statistics* 2, 1–26.

Sianesi, B. 2004. An Evaluation of the Swedish system of active labor market programs in the 1990s, *The Review of Economics and Statistics* 86, 133-155.

Export Control Restrictions

Access to and use of EPRI Intellectual Property is granted with the specific understanding and requirement that responsibility for ensuring full compliance with all applicable U.S. and foreign export laws and regulations is being undertaken by you and your company. This includes an obligation to ensure that any individual receiving access hereunder who is not a U.S. citizen or permanent U.S. resident is permitted access under applicable U.S. and foreign export laws and regulations. In the event you are uncertain whether you or your company may lawfully obtain access to this EPRI Intellectual Property, you acknowledge that it is your obligation to consult with your company's legal counsel to determine whether this access is lawful. Although EPRI may make available on a case-by-case basis an informal assessment of the applicable U.S. export classification for specific EPRI Intellectual Property, you and your company acknowledge that this assessment is solely for informational purposes and not for reliance purposes. You and your company acknowledge that it is still the obligation of you and your company to make your own assessment of the applicable U.S. export classification and ensure compliance accordingly. You and your company understand and acknowledge your obligations to make a prompt report to EPRI and the appropriate authorities regarding any access to or use of EPRI Intellectual Property hereunder that may be in violation of applicable U.S. or foreign export laws or regulations.

The Electric Power Research Institute, Inc. (EPRI, www.epri.com) conducts research and development relating to the generation, delivery and use of electricity for the benefit of the public. An independent, nonprofit organization, EPRI brings together its scientists and engineers as well as experts from academia and industry to help address challenges in electricity, including reliability, efficiency, affordability, health, safety and the environment. EPRI also provides technology, policy and economic analyses to drive long-range research and development planning, and supports research in emerging technologies. EPRI's members represent approximately 90 percent of the electricity generated and delivered in the United States, and international participation extends to more than 30 countries. EPRI's principal offices and laboratories are located in Palo Alto, Calif.; Charlotte, N.C.; Knoxville, Tenn.; and Lenox, Mass.

Together…Shaping the Future of Electricity

3002001269

Electric Power Research Institute
3420 Hillview Avenue, Palo Alto, California 94304-1338 • PO Box 10412, Palo Alto, California 94303-0813 • USA
800.313.3774 • 650.855.2121 • askepri@epri.com • www.epri.com