

---

---

---

**A WHITE PAPER:**

# **Residential Portfolio Impacts from Whole-Premise Metering**

**Thought Experiments in Random Assignment  
and Top-Down Modeling**

**Prepared for the California Investor Owned Utilities**

**Date:** 1-14-2016

CALMAC ID: SDG0295.01





**Prepared by:**

Miriam Goldberg  
Ken Agnew

**Acknowledgements**

DNV GL wishes to acknowledge many reviewers who were essential to the development of this white paper. External reviewers included Chris Russell, Russell Meyer, and Lynn Hoefgen from NMR and Caroline Chen from Statwizards. California IOU reviewers included Rob Rubin, Brian Smith, Andy Fessel, Loan Nguyen, Corinne Sierzant, Jesse Emge, Miriam Fischlein, and Kevin McKinley. Michelle Marean and Noel Stevens from DNV GL provided additional input.

## Table of contents

1	SYNOPSIS.....	1
1.1	Background and objectives	1
1.2	Conclusions	1
1.2.1	Applying RCT methods at the portfolio level	1
1.2.2	Top-down analysis	2
1.2.3	Using AMI data in evaluation	3
2	INTRODUCTION.....	4
2.1	How the work was conducted	5
2.2	Organization of the paper	5
3	COUNTERFACTUAL AND RCT CONTEXT.....	6
3.1	A utility-level counterfactual	6
3.1.1	The counterfactual scenario	7
3.1.2	The counterfactual	20
3.2	Counterfactual vs. randomized assignment experimental designs	23
3.2.1	Randomized controlled trial experiments	24
3.3	Using RCT for program evaluation	26
3.4	Literature review	29
3.4.1	Evaluation protocols for random designs	30
3.4.2	HER program evaluations	31
4	QUASI-EXPERIMENTAL METHODS: THE SPECIAL CASE FOR TOP-DOWN MODELING.....	32
4.1	Overview of quasi-experimental methods for program evaluation	32
4.1.1	Methods summary	32
4.1.2	Regression-based quasi-experimental methods	39
4.1.3	Top-down modeling approaches	41
4.1.4	Results of recent top-down studies	43
4.1.5	Guidance from recent top-down studies	46
4.2	Potential applications to net savings load shape development	51
4.2.1	Use of top-down methods to estimate net annual energy savings	51
4.2.2	Use of top-down methods to estimate net hourly energy savings	52
5	LEVERAGING AMI DATA FOR PROGRAM EVALUATION.....	53
5.1	Background	53
5.1.1	Automated measurement and verification	54
5.1.2	Current research on the performance of automated M&V methods	56
5.2	Potential applications to net savings load shape development	56
5.2.1	Assessing automated M&V for program evaluation	56
5.2.2	Program evaluation with net load shapes using AMI data	57
6	SUMMARY AND CONCLUSIONS.....	58
6.1	Applying RCT methods at the portfolio level	58
6.2	Top-down analysis and other quasi-experimental methods	58
6.3	Using AMI data in evaluation	59
6.4	Where to next	60
7	REFERENCES.....	61

---

---

---

## List of exhibits

Figure 1. Utility population .....	7
Figure 2. Utility population duplicated to create counterfactual .....	8
Figure 3. Counterfactual would-be program participants behave similarly to actual participants .....	9
Figure 4. Households in counterfactual and actual general populations show same load shape.....	10
Figure 5. Counterfactual and actual customers who purchase widgets show same load shape.....	11
Figure 6. Difference in energy consumption between counterfactual would-be participants and actual program participants represents total program savings.....	12
Figure 7. Natural EE adopters would install widget regardless of program .....	13
Figure 8. Natural EE adopters show same load shape in counterfactual and actual populations .....	14
Figure 9. Some share of natural EE adopters will participate in programs .....	15
Figure 10. Natural EE adopters who participate in programs are free-riders.....	16
Figure 11. Amount of overlap in natural EE adopters and the program participants determines the degree of free-ridership .....	17
Figure 12. The portion of the program population producing net savings.....	18
Figure 13. Example of overlapping programs within a portfolio .....	19
Figure 14. Comparison of counterfactual vs. actual participants in overlapping programs .....	20
Table 1. Summary of common quasi-experimental methods for program evaluation.....	36
Table 2. Effects typically designed to be captured by quasi-experimental methods .....	38
Table 3. Basic regression equation and difference-in-difference estimator .....	40
Table 4. Comparison of RCT estimates with quasi-experimental estimates, Schellenberg et al. ....	41
Table 5. Features of two California top-down studies .....	44
Table 6. Electricity savings estimates from California top-down studies.....	45

---

---

# 1 SYNOPSIS

## 1.1 Background and objectives

This whitepaper explores a methodology for developing an energy “savings impact load shape” for an entire portfolio of residential energy efficiency (EE) and integrated demand-side management (IDSM) programs. An initial objective was to build on the successful evaluation model of the Home Energy Reports (HER) programs using randomized control-treatment (RCT) design and explore if it could be extended to a comprehensive evaluation of the entire portfolio using advanced metering infrastructure (AMI) data to provide results on an hourly basis. Because of the randomized assignment, the savings load shape would be net of free-ridership and would fully account for interactions between program offerings within households. Utilizing newly available AMI data, the analysis would provide an hourly estimate of net savings for all programs across the whole residential population.

In discussion with California investor-owned utilities (IOUs) and the California Public Utilities Commission (CPUC) Energy Division (ED), this broad initial goal led to consideration of several interrelated issues:

- The potential for using an RCT method similar to the evaluation design for HER to obtain full portfolio “all-in” net savings.
- The potential to conduct such an analysis using hourly consumption data available from AMI, so that the evaluated savings would be hourly, that is a full portfolio, net load-impact shape.
- Lacking the ability to operate a full portfolio of programs as a randomized assignment, the potential to use a quasi-experimental analysis of whole-premise consumption data to develop portfolio savings. This approach is essentially an application of “top-down” analysis to derive net savings, an approach that has been the subject of recent work in California.
- Additional considerations beyond the prior work that are raised when addressing hourly impacts via top-down methods.
- The potential for some form of standard analytics applied to AMI data to provide an inexpensive, quick turn-around, and comprehensive net savings estimate. Use of automated ongoing consumption data analytics is often referred to as measurement and verification (M&V) 2.0.


Based on the discussions with the IOUs, the emphasis of this work was directed toward the top-down analysis approaches, with general framing related to the issues noted above. This whitepaper addresses these issues from three primary perspectives:

- Exploration of the potential for applying RCT methods at the portfolio level, with hourly data
- Exploration of quasi-experimental approximations to RCT at the portfolio level, in particular top-down analysis, including hourly top-down analysis
- Exploration of other uses of AMI data for evaluation

## 1.2 Conclusions

### 1.2.1 Applying RCT methods at the portfolio level

Randomized assignment designs derive their power from the approximation of the ideal “no-program” counterfactual against which actual participant consumption is compared. The counterfactual approximation



is provided by the control group. The control group is equivalent to the treated or participant group in composition and external forces at work, except for differences due to random assignment to one group or the other, and the treatment or program effect itself. The impact estimate based on comparing the treatment and control groups is statistically unbiased, with quantifiable uncertainty. As a result, the impact estimate based on the random assignment design is rigorous, unambiguous, and reliable, potentially even for relatively small impacts.

Central aspects of a randomized assignment design make literal application of this approach impractical at the level of a full residential portfolio net-savings load shape. In particular, it is not possible to operate a full portfolio of programs using a range of delivery and outreach methods, where customers are randomly assigned either to have the full portfolio available or to be denied all of it.

### 1.2.2 Top-down analysis

Top-down analysis fits a model of aggregate consumption across geographic areas and time, as a function of portfolio activity along with demographic and economic factors. This modeling is designed to isolate the effects of the EE portfolio while controlling for other factors. That is, the model with program activity variables set to zero estimates the no-program counterfactual. This quasi-experimental approach has particular appeal for portfolio-level analysis because in principle it captures comprehensive portfolio effects including spillover, market effects, cross-program interactions, free ridership, and take-back. The approach has thus far been applied only at the level of annual energy savings, but could in principle be extended to an hourly framework.

California has established a database that can be used for developing top-down models for portfolio-level impact analysis. This database addresses one of the common impediments to development of such models, with monthly data.

There remain several limitations to the top-down methods. One is that any such analysis will retain the potential for method bias, or model specification uncertainty. Even with a well-established database, the only variables that can be used in the model are those that are available from the compiled sources, and at the level of geographic and temporal detail provided by those sources. As a result, some level of model misspecification bias is inevitable, as with virtually any modeling exercise. Even if all possible variables were available, different reasonable models yield different results.

Self-selection bias is also an issue with top-down analysis, as it is for other impact estimation based on individual customer consumption analysis. In effect, top-down models estimate program effects by comparing consumption across areas and times, "all else being equal." This means the models control analytically for other factors that vary over areas and time. To the extent that consumption is lower for areas and times with higher program activity, all else equal, this could be due to the program, or could reflect customers with greater naturally occurring EE interest, all else equal, being more inclined to engage in programs.

Along with the potential for bias or specification uncertainty from various sources, a second set of limitations of top-down models has to do with the level of detail the model can provide. A top-down estimate cannot provide separate estimates of the various net-to-gross components, such as free ridership and spillover. The estimate also provides an overall net savings per unit, and does not identify how a particular program year or area was more or less effective.

---

---

---

To extend the top-down methods to address hourly data would require a number of technical issues to be addressed. Most of the predictor variables for such models are available only at the annual or at best monthly level. The model structure must take into account variation by time of day and calendar, while appropriately reflecting lag effects. Work would be required to develop and test some approaches. Availability of clean hourly data for this work also remains a challenge. Hourly AMI data are collected but not comprehensively cleaned and compiled.

### 1.2.3 Using AMI data in evaluation

The availability of AMI data for all customers introduces new opportunities into evaluation, most of which are just beginning to be explored. These data make possible consumption analysis at the daily or hourly level, for essentially all customers, rather than relying on monthly data for all or higher frequency data for a sample. However, data quality and comprehensiveness remains a concern for any methods relying on AMI data. Key applications of AMI data for evaluation are:

- **Estimation of energy impacts using extensions of established analysis methods from consumption data analysis.** These methods include using essentially the same models with finer timescale data, such as daily and expanded models incorporating additional terms.
- **Estimation of hourly impacts using methods familiar in the context of load research and demand response analysis.** Some of the basic modeling approaches used for load data analysis can be applied to estimate net savings at the hourly level, in contexts where a comparison group is well defined and meaningful.
- **Automated M&V methods.** Commonly referred to as automated M&V, these methods process large volumes of individual customer data on an ongoing basis. A training period of up to 12 months establishes a baseline model prior to an intervention. Comparison of actual consumption with this baseline provides a measure of savings for any subsequent period. Automated M&V tools are primarily designed as program support tools, not for evaluation, but after validation of their predictive accuracy can be used in the evaluation context.

While such tools provide the capability of rapid results, the quality of these results is similar to what would be obtained by existing consumption data analysis methods. They are effective for providing savings estimates when the baseline is the prior equipment, and an appropriate comparison group is available. All the challenges of defining an appropriate comparison group remain.

At this stage automated M&V tools are mostly using monthly, not hourly data. Hourly models would be natural extensions, with many details yet to be worked out.

- **Data quality.** This remains an issue for any applications of AMI data to evaluation, whether for annual energy or hourly impacts. Data attrition is an issue for any consumption data analysis, but potentially is worse with AMI data. Work is needed to establish whether there are any patterns to the prevalence of missing and anomalous data for a particular service territory, and what biases these might lead to in an evaluation.

## 2 INTRODUCTION

The goal of this whitepaper is to explore a methodology for developing a “savings impact load shape” for the entire EE and IDSM residential program portfolio offered. A starting objective was to build on the successful evaluation model of the HER programs using RCT design, and explore how that approach could be extended to a comprehensive evaluation of the entire portfolio, using AMI data to provide results on an hourly basis. The savings load shape would be net of free-ridership and would fully account for interactions between program offerings within households. Utilizing newly available AMI data, it would be an hourly estimate of net savings for all programs across the whole residential population.

In discussion with the CA IOUs and the ED, this broad initial goal led to consideration of several interrelated issues:

- The potential for using an RCT method similar to the evaluation design for HER to obtain full portfolio net savings.
- The potential to conduct such an analysis using hourly consumption data available from AMI, so that the evaluated savings would be hourly, that is a full portfolio, net load-impact shape.
- Lacking the ability to operate a full portfolio of programs as a randomized assignment, the potential to use a quasi-experimental analysis of whole-premise consumption data to develop portfolio savings. This approach is essentially an application of top-down analysis to derive net savings, an approach that has been the subject of recent work in California.
- Additional considerations beyond the prior work when addressing hourly impacts via top-down methods.
- The potential for some form of standard analytics, often referred to as M&V 2.0, applied to AMI data to provide an inexpensive and comprehensive net savings estimate.

Based on the discussions, the emphasis of this work was directed toward the top-down analysis approaches, with general framing related to the other issues.


This whitepaper will discuss how the ultimate goal of a full portfolio-level residential net savings load shape can only be attained in a theoretical world of parallel universes. The paper will expand on this thought experiment to illustrate and clarify what aspects of this unattainable thought experiment are essential to get to the desired net savings load shape. Practical, real-world energy program evaluation attempts to attain this ideal with a variety of methods at both the single program level and when estimating aggregate system-level savings estimates through top-down modeling. The purpose of this paper is to place the range of methods used in program evaluation into the framework provided by this theoretical world of parallel universes. In particular, we focus on top-down modelling because it aspires to produce portfolio-level residential net savings via regression and available data.

We expand the discussion to include AMI data because the question of how the newfound wealth of data will inform the evaluation effort is directly relevant to the discussion. AMI data moves the idea of hourly interval data analysis on a utility population from the imaginary to the feasible. This paper will address what other ways AMI data impacts our efforts to attain the elusive counterfactual with practical methods and data.

To meet this objective, the paper focuses on a number of research objectives:

- Provide a simplified illustration of the counterfactual to act as a framework within which to understand the range of evaluation methodologies discussed.



- 
- Discuss how the successful evaluation model of the HER programs using RCT design is an example a methodology that reasonably approximates the counterfactual in practical terms.
  - Explore how the general set of methods applied to randomized and quasi-experimental design approximate the counterfactual with different techniques and making different assumptions.
  - Explore how top-down modeling extends this process to a comprehensive evaluation of the entire portfolio.
  - Finally, address how the availability of AMI data may enhance the ability of some methodologies to approximate the counterfactual but will not address fundamental aspects of the evaluation challenge.

Based on the discussions, the emphasis of this work was directed toward the top-down analysis approaches, with general framing related to the other issues.

## **2.1 How the work was conducted**

Two webinars were conducted with IOU evaluation, measurement, and verification (EM&V) project managers and ED staff to discuss and refine the study objectives, identify key literature to be reviewed, and explore initial directions and considerations in the work. Through these conversations, the emphasis of this work (described in Section 1.1) was clarified.

After reviewing the key literature on the topics, DNV GL prepared a draft version of this whitepaper and discussed the findings with the working group. We then distributed the draft for public review, and conducted a public webinar on it. The work has been revised based on comments from the IOUs and the general public.

## **2.2 Organization of the paper**

The work starts with a review of the RCT framework, then discusses top-down modelling approaches that apply quasi-experimental methods to savings analysis using aggregate consumption data, and finally, addresses incorporation of AMI data.

---

---

### 3 COUNTERFACTUAL AND RCT CONTEXT

In this section, we describe an idealized framework within which to develop a portfolio-level net saving load shape. This is the ideal to which feasible methods should aspire. We explore the concept in the following steps:

1. We develop a thought experiment that illustrates how such a portfolio-level net saving load shape could be derived. We discuss the nature of the parallel universe counterfactual.
2. Next, we offer a practical picture of what that counterfactual world would look like and discuss how a comparison of consumption load shapes between the counterfactual and reality would offer exactly the net savings load shape that would provide the reliable estimate of true program effects.
3. Then we contrast this portfolio level scenario to existing, actual RCTs for behavioral programs.
4. Finally, we discuss the clear limitations to feasibility to the development of an actual RCT design at the portfolio level.

#### 3.1 A utility-level counterfactual

Large-scale RCT experimental designs conducted across a range of behavior programs have piqued the imagination of energy program evaluators. HER programs offer the promise of precise and statistically valid savings estimates at an annual, monthly, or hourly level. These designs offer a template that makes the idea of a system-wide experimental design appear tantalizingly close to feasible.

The following figures illustrate how a system wide, portfolio-level net savings load shape would be developed in a theoretical world and the idea of the counterfactual. The counterfactual does not exist in reality, but it can be defined to good degree of complexity in a theoretical world. Developing the scenario clarifies why this extension of the RCT is so compelling; it clearly illustrates the counterfactual basis for unbiased estimates of program impacts.

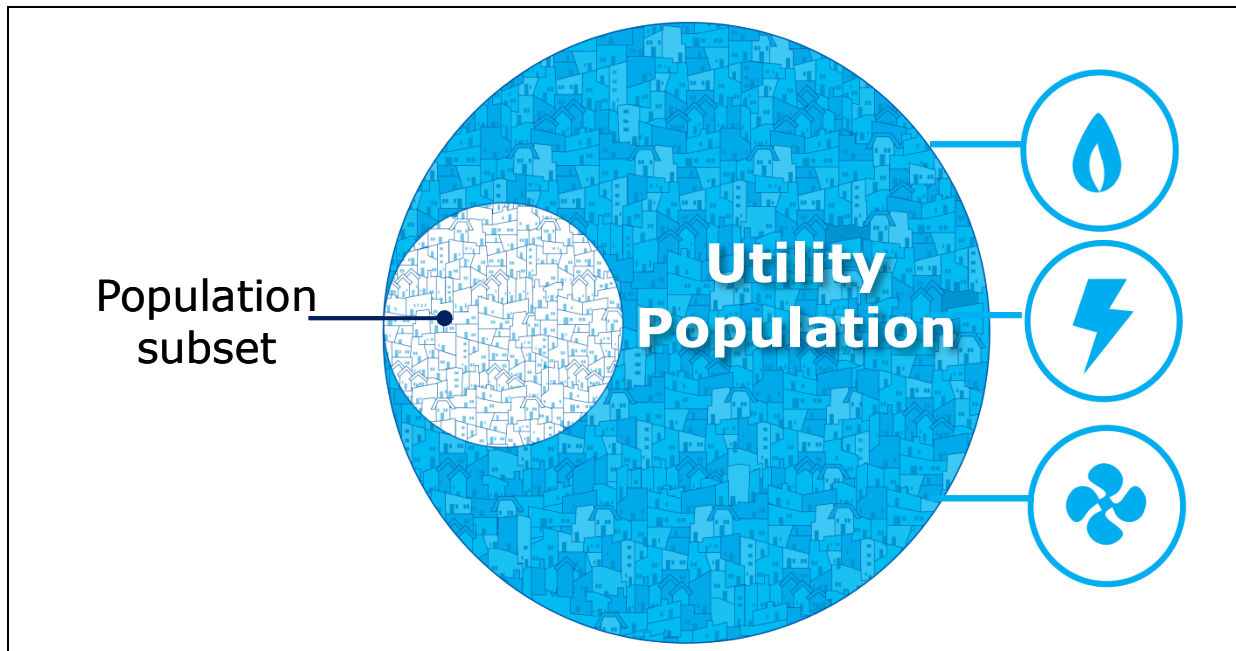
The scenario will also demonstrate how the ideal of a system-wide, portfolio-level net savings load shape is not feasible without access to parallel universe. Instead, the scenario provides a useful heuristic of the evaluation process that feasible methods must approximate. The counterfactual scenario as an ideal provides the starting point for discussion of the RCT experimental designs that are feasible under certain conditions. In later sections, the discussion will expand to include other methods, ultimately focusing on top-down modeling, which comes back to the idea of system-wide, portfolio-level results making use of feasible methods.

In the section that follows, we first define the counterfactual and outline the challenges associated with identifying a true counterfactual. Then, we discuss how a randomized assignment experimental design approximates the counterfactual. Finally, we discuss the literature regarding randomized assignment experimental design.

### 3.1.1 The counterfactual scenario

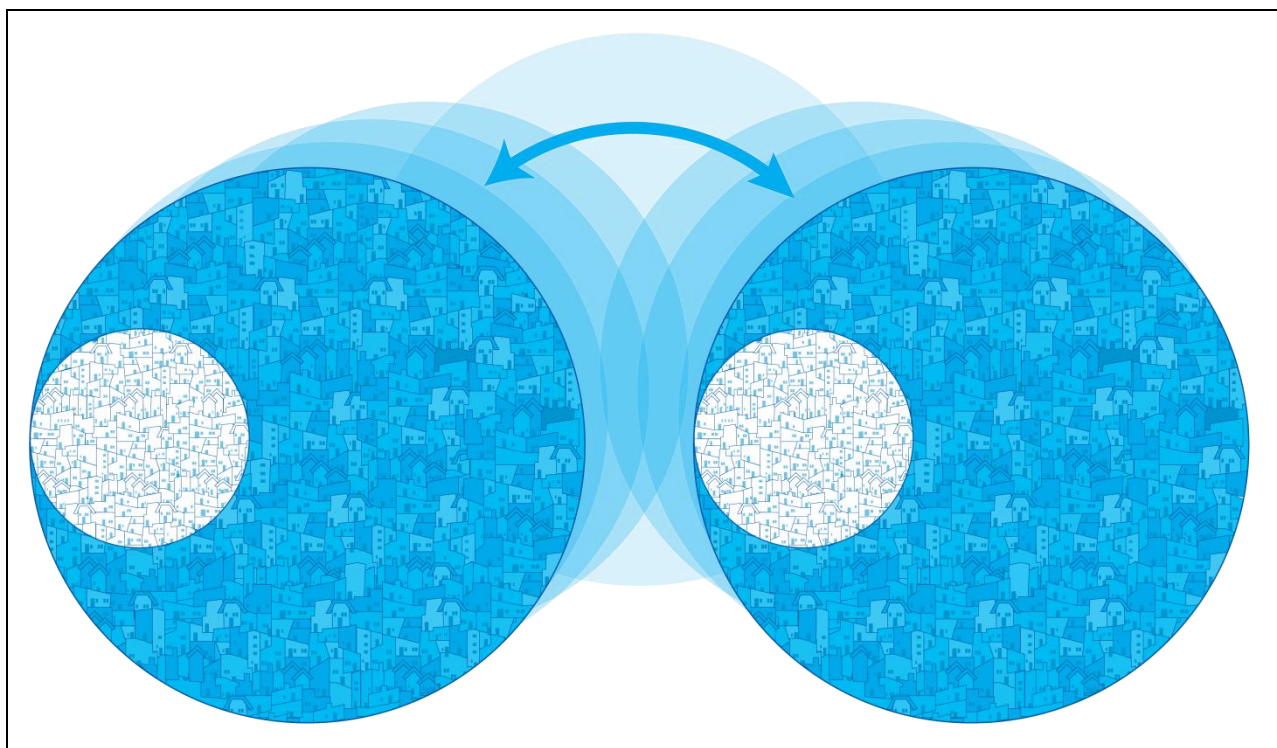
Figure 1 offers a stylized picture of a utility population. Everything within the larger blue circle represents the actual utility residential population. Each household, which is represented by a pixel, has its associated energy consumption. Among that population, there is a subset of households that will install a widget this year. To simplify this scenario, we will imagine that widgets are a measure common to all homes and are replaced after some variable period. In this respect, widgets are similar to refrigerators and furnaces. Every year, a subset of the population will purchase or replace these items. While the pictures illustrate the process for regularly replaced widget, the general concepts are transferable to other EE measures such as insulation, which are installed once.

**Figure 1. Utility population**



Imagine duplicating that utility population so that, in a parallel universe, there is an identical and parallel utility population moving through time (Figure 2). This parallel universe amounts to the ultimate experimental design where we have two identical populations and can test the effect of a treatment by treating only one, called the actual, and not the other, called the counterfactual.

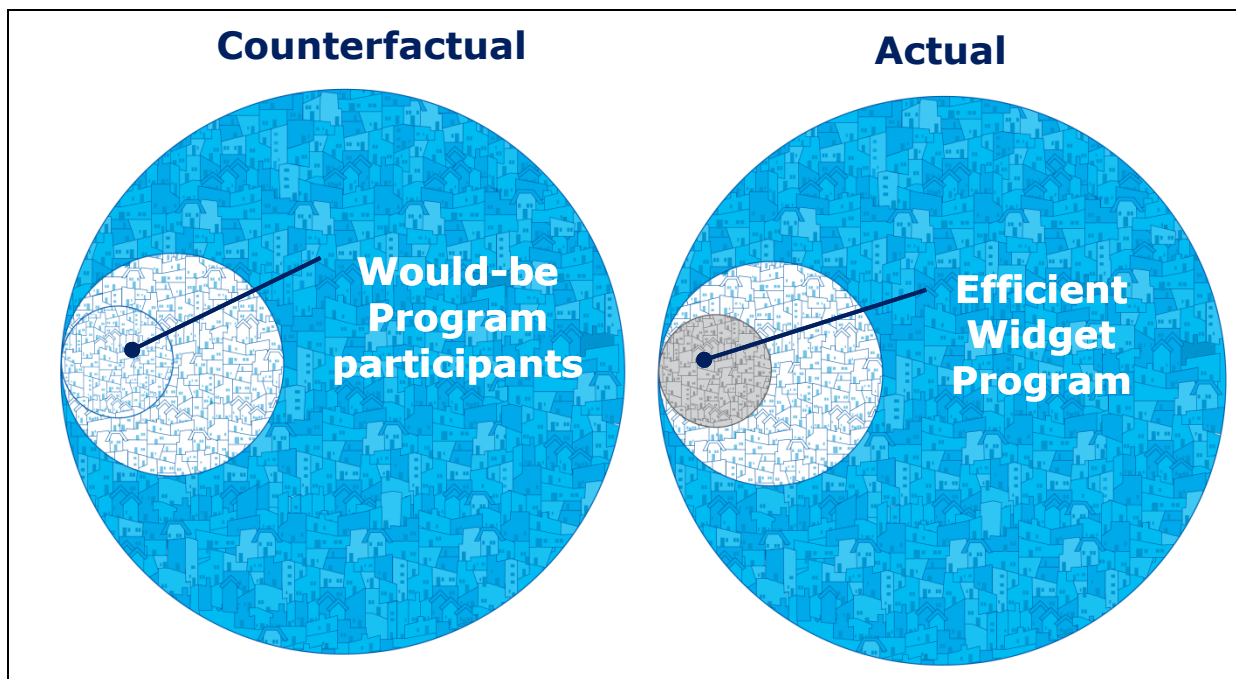
**Figure 2. Utility population duplicated to create counterfactual**



For the actual population in Figure 3, an EE program provides incentives for households to upgrade the standard widget they intend to install to a high efficiency widget. Only a subset of the widget-installing population takes advantage of the incentive. All of those households that take part in the program will have lower than standard efficiency energy consumption because of the high efficiency widget.

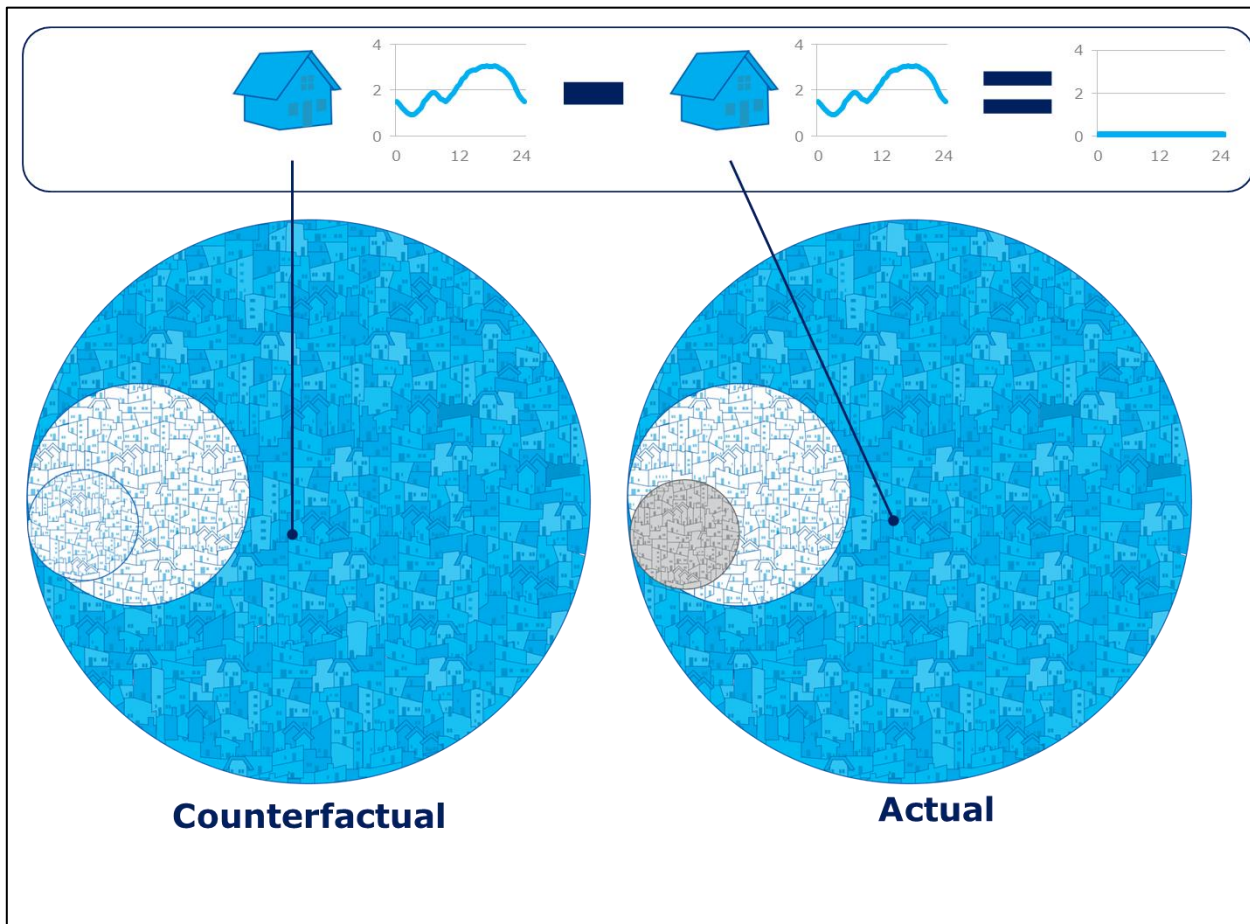
Opting into an efficient widget program is an act of self-selection. Despite the program incentives, these households will usually make a greater up-front investment. Program participants will have characteristics, observable and unobservable, that may be different from the rest of the widget purchasing population or the utility population as a whole. A key part of the counterfactual scenario is recognizing that this “would-be” participant subgroup will likely be different from the rest of the population even in the absence of the program. In the counterfactual scenario, would-be participants are identical to the actual world participants in every respect aside from the program effects. They are also may be different from all other population subgroups as is consistent with the set of characteristics that lead to their self-selection into the program.

**Figure 3. Counterfactual would-be program participants behave similarly to actual participants**



In Figure 4, the actual and counterfactual worlds can be compared on a household-by-household basis across the full populations. In the scenario that has been described, households in the blue portion of the circle are identical in both the actual and counterfactual worlds. These households are wholly unaffected by the presence of the program in this simplified example with no spillover. The aggregated hourly load shapes for the blue households in the actual and counterfactual populations will be the same as represented by the equation that shows counterfactual minus actual equals a zero load shape.

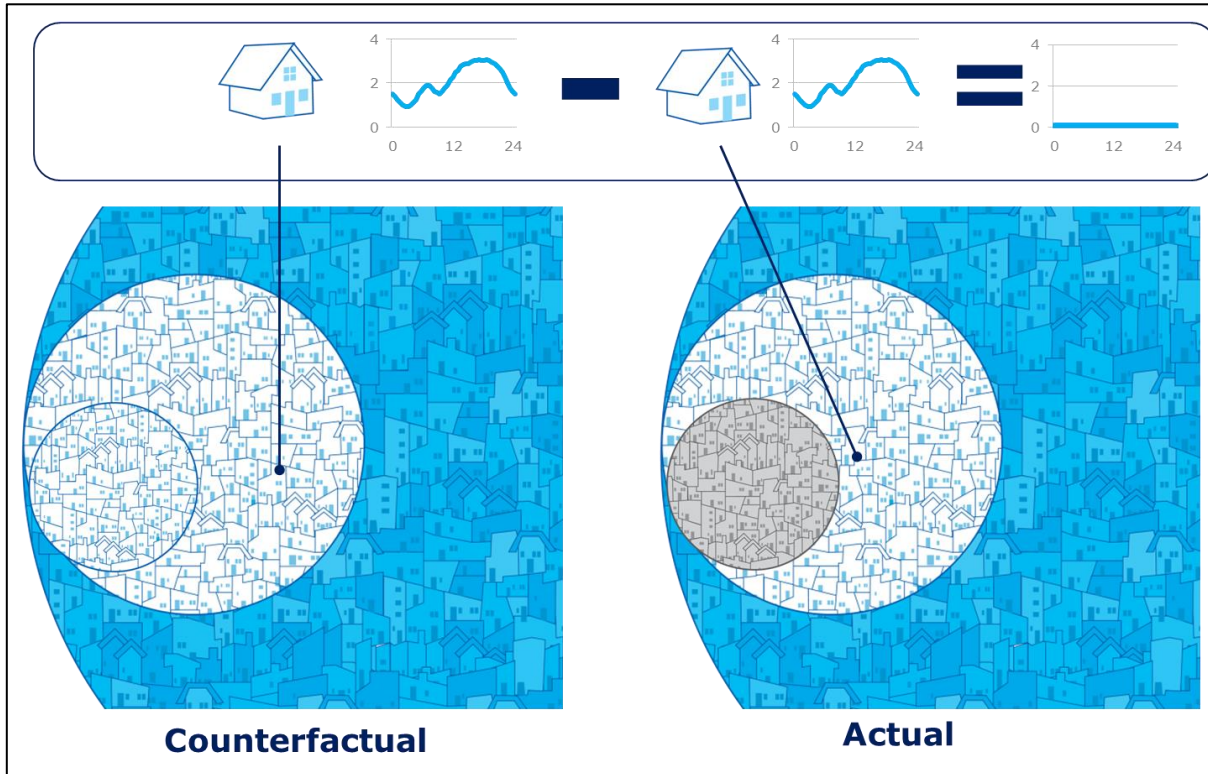
**Figure 4. Households in counterfactual and actual general populations show same load shape**





The households inside the white circles, but outside of the program or would-be program circles, will purchase widgets (Figure 5). They are also identical in the actual and counterfactual worlds. These households are also wholly unaffected by the presence of the program. These households may differ from the blue households because they purchased widgets, but the aggregate hourly load shapes for the white circle households not in the program will be the same in the actual and counterfactual populations.

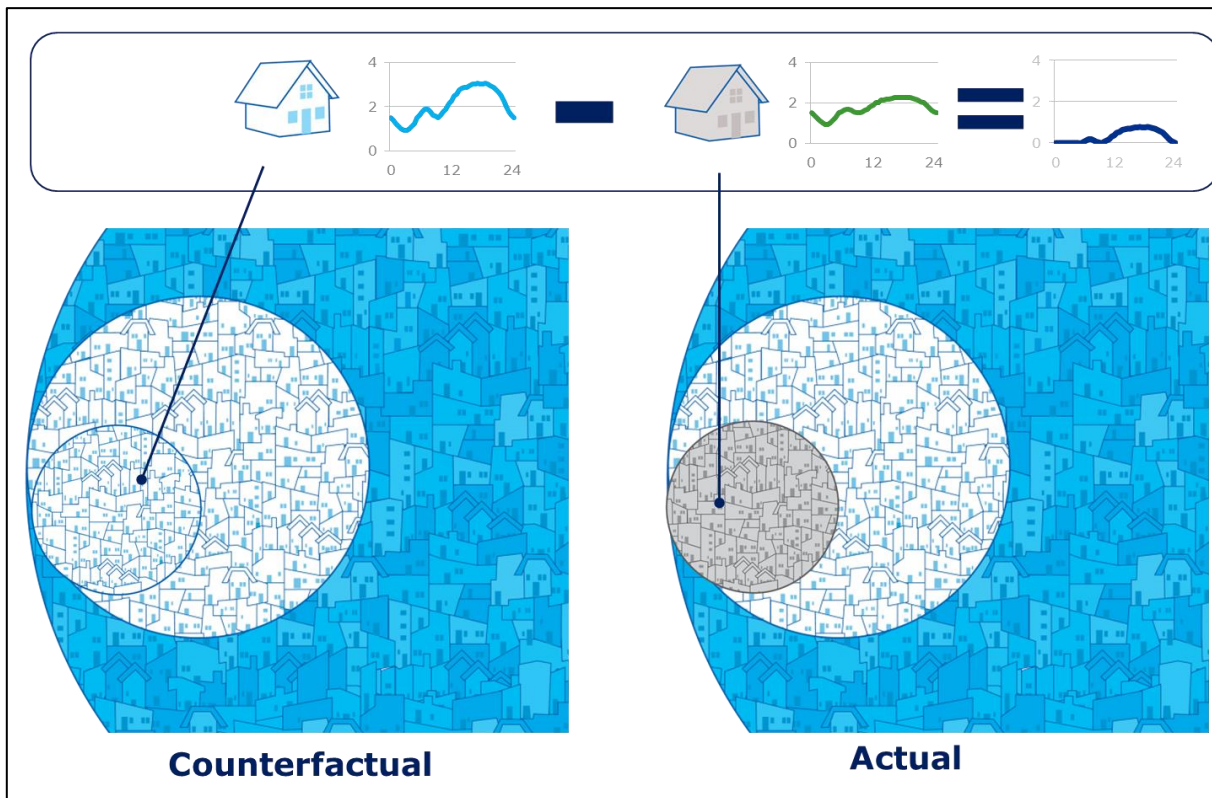
**Figure 5. Counterfactual and actual customers who purchase widgets show same load shape**



In Figure 6, the smallest circle represents the households that participate in the program. In the actual world, the circle is gray to indicate that the households participated in the program. In the counterfactual world, the circle is present, but not gray, because there are no EE programs in a counterfactual world. We refer to these households as would-be program participants. The same households are in this circle in both the actual and counterfactual worlds. The only difference between the two circles is that the program is involved in the process in the actual world.

The aggregated hourly load shapes of consumption for the program circle households will be different in the actual and counterfactual populations. The difference will represent the program-related savings due to all program households having installed high efficiency widgets. The difference in consumption between treatment and counterfactual will represent the total program savings. If the total consumptions of the full actual and counterfactual populations are differenced, the result will also be the total program savings.

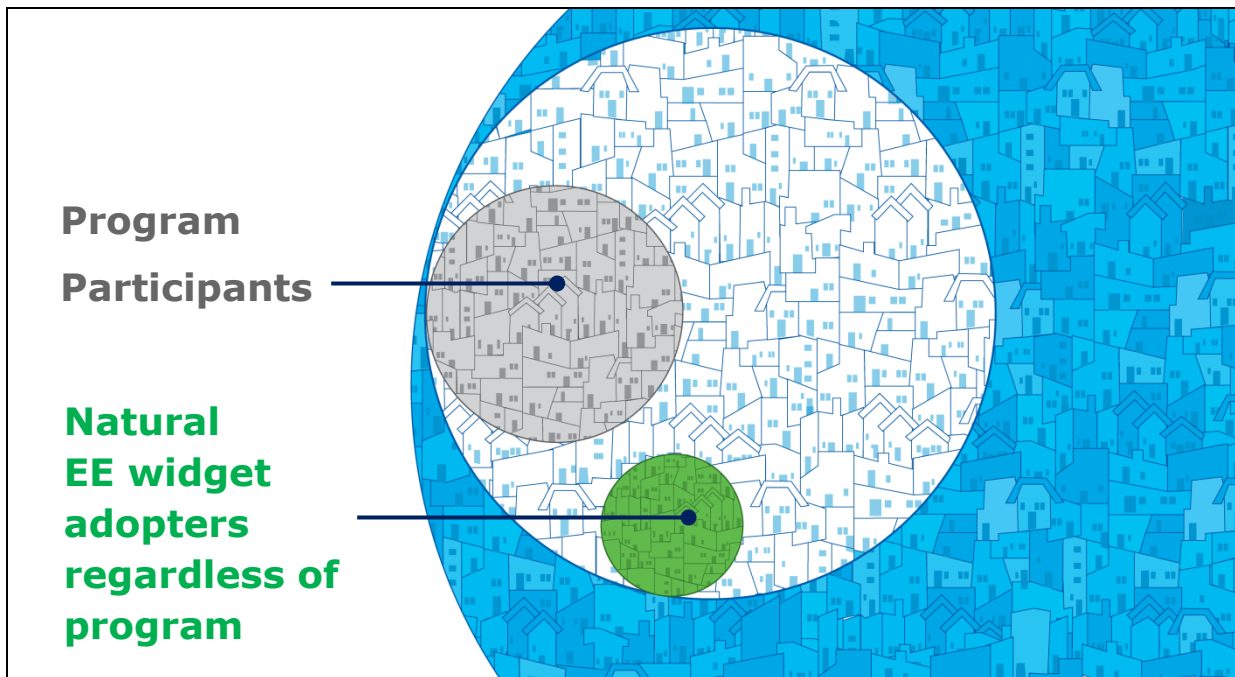
**Figure 6. Difference in energy consumption between counterfactual would-be participants and actual program participants represents total program savings**





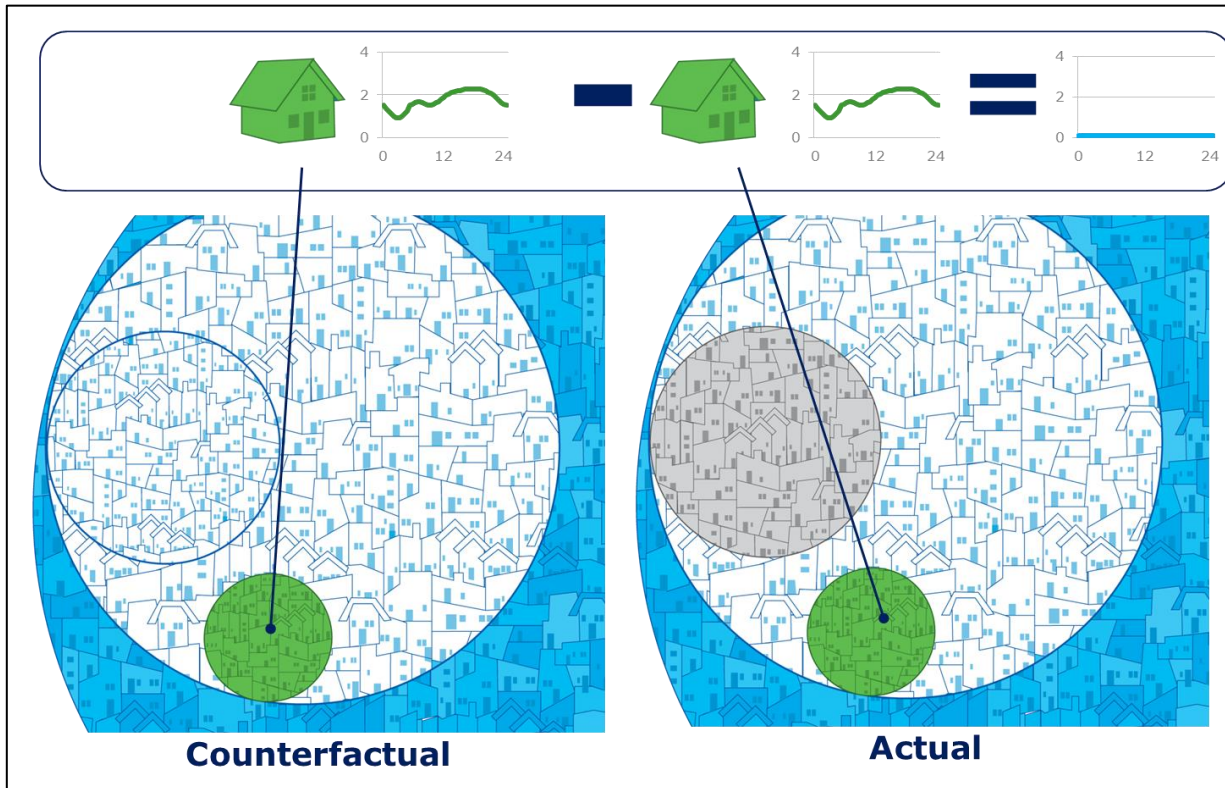
To make the scenario more complex, in Figure 7, another group of households is identified. We refer to this group as natural EE adopters. In this scenario, the term refers to households that would install an energy efficient widget even in the absence of the utility program. In this simple scenario, the natural EE households are installing the same widget as those households that participate in the utility program.

**Figure 7. Natural EE adopters would install widget regardless of program**



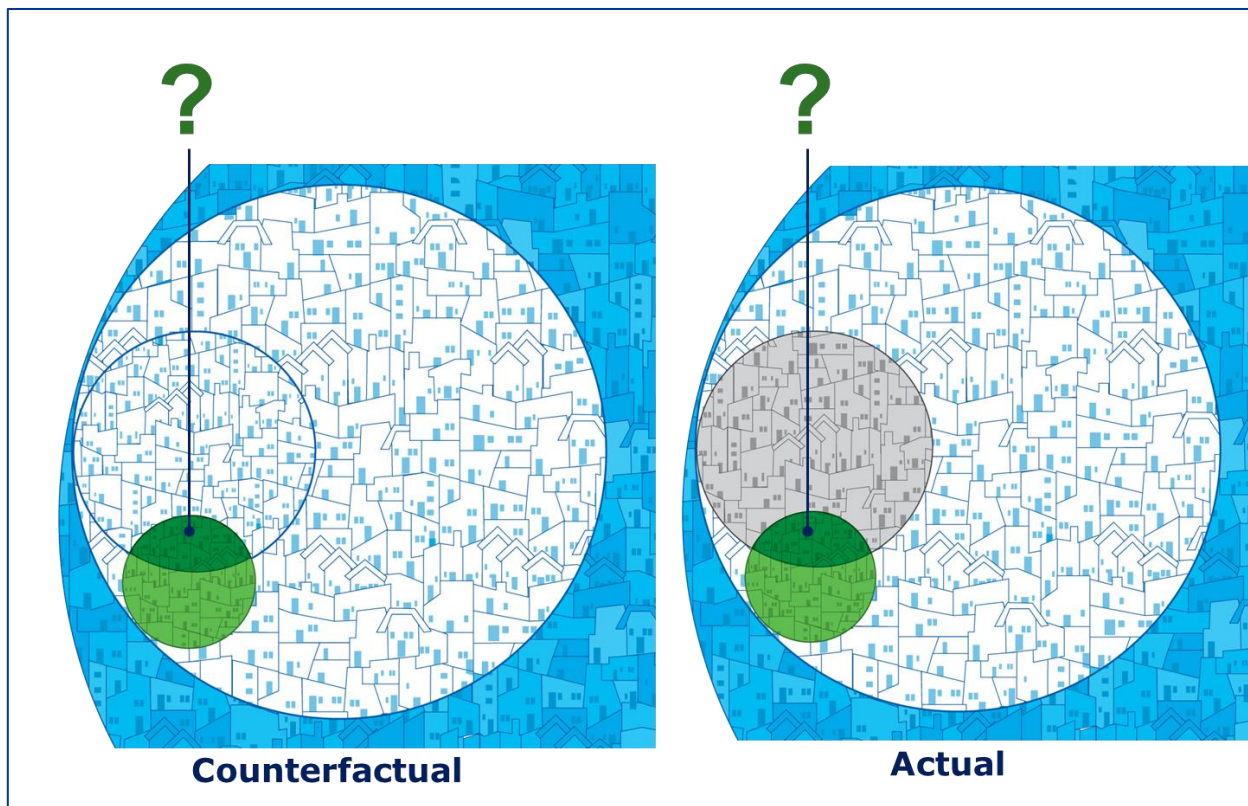
Natural EE adopters, pictured in the green circles in Figure 8, are wholly unaffected by the program. As with all of the other sub-populations that are unaffected by the program, the aggregated hourly load shapes of consumption for the natural EE adopter households will be identical in the actual and counterfactual populations. The level of consumption for natural EE adopters will be lower than the consumption in households that installed a standard efficiency widget (the remaining white circle outside of the program households), but it will be identical between natural EE adopter households across the two worlds.

**Figure 8. Natural EE adopters show same load shape in counterfactual and actual populations**



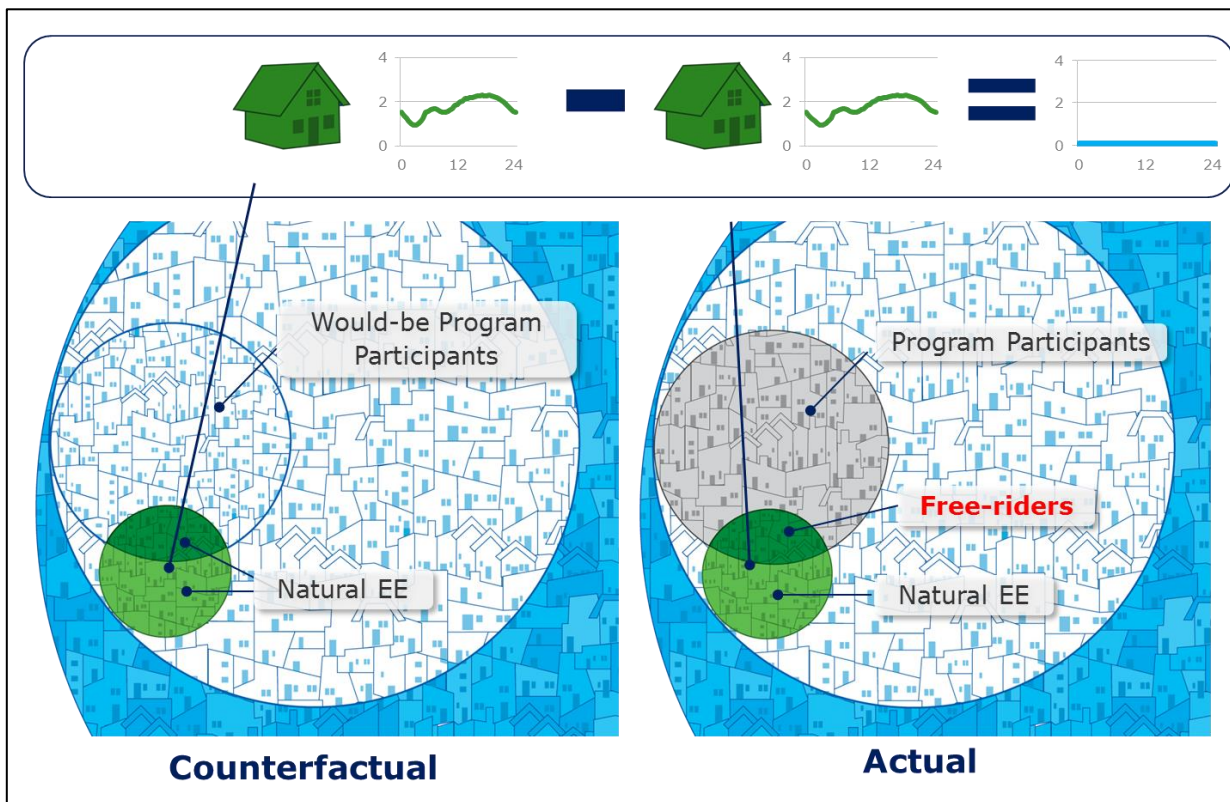
Natural EE adopters frequently participate in programs (Figure 9). There is a clear motivation for natural EE adopters to do this. They get the efficient widget that they intend to purchase, but the program incentive offsets some of the cost. Participating in a program has its own cost with respect to time and effort, so not all natural EE adopters will necessarily participate. In addition, programs frequently try to target their message, so it is possible that natural EE adopters would not know that they could get a rebate for the EE widget they intend to install. To the extent possible, programs attempt to avoid providing high efficiency widgets to natural EE adopters but it can be difficult to deny customers access to programs.

**Figure 9. Some share of natural EE adopters will participate in programs**



In Figure 10, the households of natural EE adopters that overlap with households of program participants are referred to as free-riders (right). There is no difference in consumption between actual and counterfactual for households in the overlap region between program and natural EE adopters. For the portion of the program circle where there is no overlap, the households in the actual world have lower consumption due to the program's EE widget. Where the natural EE adopters overlap with the program circle, that difference does not exist. In both the actual and counterfactual worlds, the households in the overlap region have the consumption of an EE widget. Only the households in the non-overlapping part of the program circle will produce program savings even though the program is serving the full program circle.

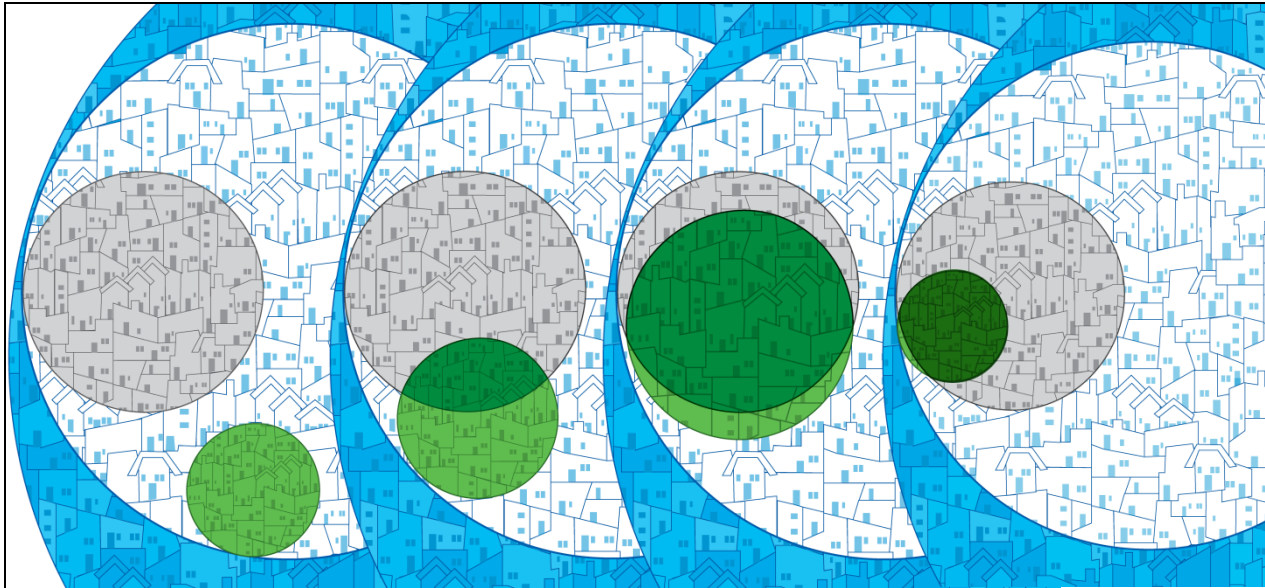
**Figure 10. Natural EE adopters who participate in programs are free-riders**





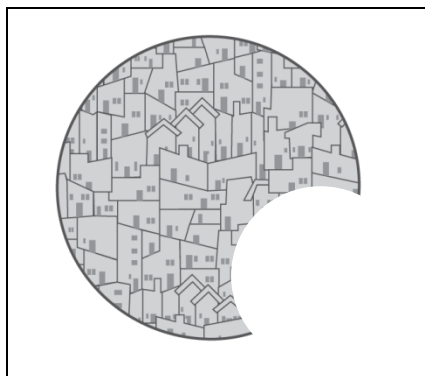
The amount of overlap between the natural EE adopters (green circles) and the program participants (gray circles), determines the degree of free-ridership (Figure 11). The leftmost crescent illustrates a utility program with no natural EE adopters. Although there is a clear economic motivation for natural EE adopting households to take part in programs, the amount of overlap between the natural EE adopters and program participants, or free riders, varies among programs and participants. Natural EE adopting households might not take part in the EE program because of the required level of effort relative to the widget rebate, other barriers intentionally built into utility programs to deter EE adopters from participating, or from a simple lack of knowledge that the EE widgets program exists. The amount of overlap will be determined by these factors along with the relative sizes of the program and the natural EE adopter population.

**Figure 11. Amount of overlap in natural EE adopters and the program participants determines the degree of free-ridership**



If we compare aggregated hourly load shapes of consumption across the complete actual and counterfactual populations, only the portion of the program circle not overlapped with natural EE adopters will be different (Figure 12). Typically, in an evaluation setting, the evaluator does not know who the natural EE adopters are, so they do not know if the overlap region exists or how big it is. Instead, for instance, the full participant circle is compared to a comparison group constructed from the remaining actual world households that is intended to approximate the counterfactual would-be program participants. To successfully estimate net savings, that actual world comparison group would need to have the same household characteristics as the self-selected program group and have the same proportion of natural EE adopters. By definition, the characteristic of self-selection into the program is not present in the rest of the population and, given that those characteristics are likely unobservable, would be difficult to locate. Locating the natural EE adopters is similarly challenging and, if many or most have opted into the program, would be impossible.

**Figure 12. The portion of the program population producing net savings**



The simple widget scenario illustrated here can be expanded to cover most realistic situations. For example, Figure 13 shows a program portfolio with multiple programs that, in this case, overlap and are shown as gray circles. This picture illustrates the range of possible overlaps between two programs targeting different widget-like measures. The two populations of widget installers may overlap (e.g., yellow households that install a new refrigerator and a new furnace in the same year). Both individual widget-installing populations and those doing both measures may or may not take part in the relevant programs. Finally, natural EE adopters exist in both individual widget-installing populations and those doing both measures.

**Figure 13. Example of overlapping programs within a portfolio**

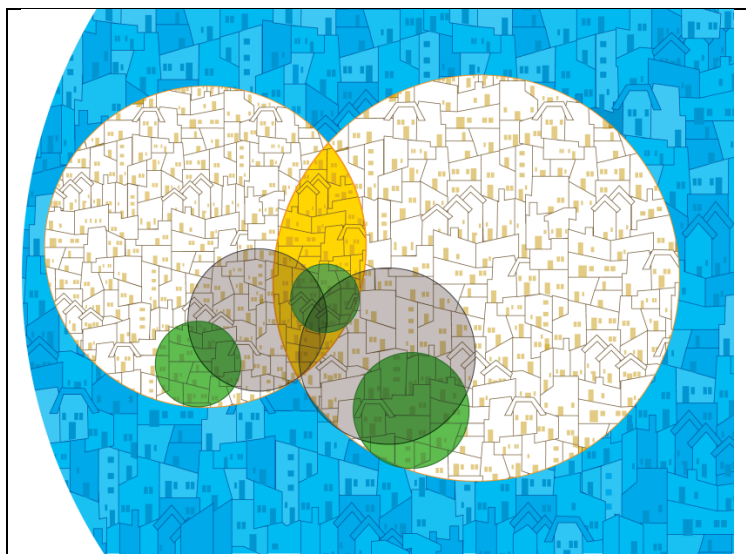
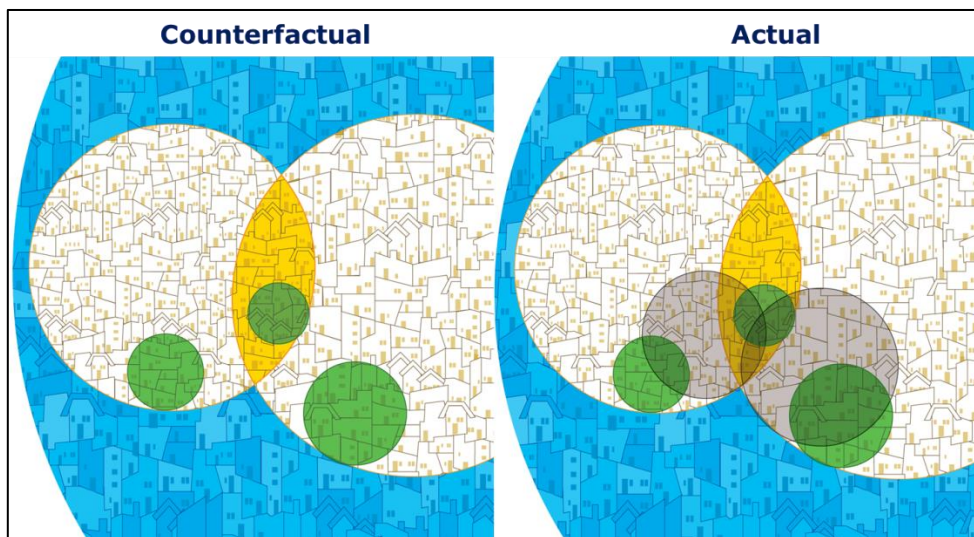


Figure 14 shows the comparison from actual to counterfactual. Only households in the gray program circles that are not overlapping with the green natural EE adopters will achieve the savings for the program EE measures. Overlapping program circles will capture efficiency savings from both widgets less any interactive effects.

**Figure 14. Comparison of counterfactual vs. actual participants in overlapping programs**



### 3.1.2 The counterfactual

At its most basic level, the counterfactual heuristic identifies three groups within the overall population:<sup>1</sup>

- The group that will purchase a widget of any kind
- The natural EE subgroup that will purchase the EE version of the widget, regardless
- Participants in the program that promotes the EE version of the widget.

The program participants and the natural EE subgroup are both subsets within the group that will purchase a widget. The wildcard among these three groups is the natural EE adopter subgroup that will install the EE version of the widget, regardless. The size of this subgroup represents the natural market share for EE widgets. It is the goal of the program to expand that market for high efficiency widgets while limiting the funding of EE widget purchases by households that would purchase the EE widget, regardless. The fundamental question of all energy program evaluation is quantifying the overlap between these two populations. That is, how big is the natural EE group and how many in the natural EE group take advantage of the program to fund their EE widget? The overlap of natural EE and the program represents free-ridership.

<sup>1</sup>This scenario is necessarily simplified. In this case, the widget is most easily thought of as a common measure that is regularly replaced like a furnace, AC, or refrigerator. There are different levels of efficiency for the widget and the program is designed to overcome market barriers to wider adoption of the higher efficiency version of the widget. There are other kinds of programs, like programs promoting insulation and weatherization, where the program might not be designed to increase the efficiency but, instead, to increase adoption of the measure altogether. The goal of the figures is to illustrate the most basic scenario. With modifications, all realistic scenarios can be illustrated.



---

---

---

As the schematic figures above illustrate, the extent of free-ridership is a function of the relative sizes of the program and natural EE populations and the degree of strategic behavior among natural EE populations to save money on their purchase. A number of key concepts are regarding the scenarios are discussed next.

### **3.1.2.1 Perfect and complete match**

The counterfactual is a perfect and complete match for the population that receives the program. At the pixel or household level, the match is one to one. Absent the effects of the program, an aggregate hourly difference between the system population and the counterfactual population would be zero for all hours.

This goes for all subpopulations across the distribution of all characteristics. The distribution of annual consumption is identical, but also the distribution of specific subsets of the population, such as employees at that major employer that may have major layoffs. The number of households with new babies is identical as are the number with kids leaving for college for the first time. All of the possible drivers of household consumption are present in an identical way across the population and the counterfactual.

Of particular importance, the households that are natural EE adopters (and free riders if they use the program) have the same distribution across the counterfactual as they do across the actual population. There are similar proportions of natural EE adopters and they are located in households with similar characteristics.


In the most simplistic sense, these households' EE actions and subsequent energy consumption are unaffected by the presence of the program. In the fullest sense of "free-rider," they would have taken EE actions without the program. Thus, even prior to the insertion of a program into the scenario, these households are present, equally distributed, and identical in energy consumption outcomes.

### **3.1.2.2 Clear identification of program effect**

In the simple example above, as we compare each actual subgroup with its corresponding counterfactual, the program effect is a simple aggregate difference in energy consumption for the compared subgroups. In the example, the consumption of all subgroups outside of the program circle for the actual and counterfactual are identical because the program has had no effect. For program participants (the gray circle), the counterfactual provides a perfect baseline for both non-free-riders and free riders. For the non-natural EE adopter participants (the participant crescent excluding the natural EE group), the counterfactual reflects the standard efficiency widget that the participant would have done in the absence of the program. The program ability to lower consumption from the counterfactual is the true measure of programs savings for these households. Where the natural EE contingent overlaps with the program participants, the program has no additional effect, consumption levels are the same. The counterfactual natural EE group would install the EE widget even without the program intervention.

### **3.1.2.3 Clear tracking of change over time**

In the counterfactual example, the perfect match and the resulting clear identification of the program effect extends through the dimension of time. While the counterfactual scenario in the figures has the appearance of a snapshot, we can imagine that a "film version" shows the actual and counterfactual moving though time in a parallel fashion. In this scenario, exogenous factors such as weather, economic change, and population characteristics affect both groups equally and at the same time. For instance, the increase and decrease of household occupancy would remain the same between the two worlds. The counterfactual will track the natural trends in energy consumption related to economic signals, adoption of new technologies, and changing social or political cues. For example, households that get motivated to improve household EE will



make consumption changes without assistance from the various measure-based and opt-in behavioral programs. Their counterfactual consumption will provide the appropriate baseline against which to measure additional program assisted changes in consumption.

### 3.1.2.4 Additional considerations

As with all simplified scenarios, they only tell part of the story. The framework for a full-portfolio, net load impact shape should extend its range to accommodate additional and more complex interactions. Spillover and early replacement are two particularly important ideas that can be discussed in the counterfactual scenario figures but would needlessly complicate the visual approach. Rather than push the counterfactual scenario pictures too far, we describe the issues in terms of the figures and how they might be expressed visually.


**Increasing complexity:** In the interest of clarifying the basis for a net savings load shape, the figures treat both program widgets and natural EE adoption widgets as being identical and producing the same decrease in consumption. This simplifies the comparison, actual to counterfactual, of the overlap region between natural EE adopters and the program. In the actual world, natural EE adopters that join the program purchase the identical high efficiency widget they would have without the program. In the counterfactual world, the natural EE adopters purchase the same high efficiency widget. In comparing households in this overlap region across these two worlds, there will be no difference in consumption despite the presence of the program in the actual world. This facilitates the fundamental understanding that, in the counterfactual scenario, free-ridership occurs when natural EE adopter households interact with programs.

In actuality, there will be a range of high efficiency widgets available and natural EE adopters may purchase a mix of high efficiency widgets that are on average more or less efficient than the program high efficiency widgets. Program high efficiency widget consumption effects will also be variable depending on such factors as program delivery and customer interaction with the program, such as if a monetary incentive increases the likelihood of take-back. With this in mind, the simple negation, in the overlap region, of the actual world program effect by the corresponding natural EE adopter households is an oversimplification. Furthermore, for EE adopters, a program incentive could still have a savings effect by providing incentive money to re-direct into other EE purchases.

**Spillover:** Spillover describes all savings caused by the program that are not related to the rebated high-efficiency widgets. None of the counterfactual scenario figures presented earlier included any spillover.

All kinds of spillover, including participant and non-participant and “like” or “unlike” spillover, can be placed in the counterfactual scenario. Like-spillover would be the program-induced, indirect adoption of EE widgets by households within the white circle of widget purchasers but outside of the gray program sphere (i.e., with no incentive), and outside the green natural EE adopter circle. Visually, like-spillover would effectively increase the size of the gray program circle without additional program costs. Compared with the counterfactual, like-spillover households would have the lower consumption of the EE widget.

Unlike spillover would show up as any non-widget, program-induced reductions in consumption across the whole actual utility population. Among program participants, examples of unlike spillover would be the installation of additional EE measures because of a new appreciation of EE due to the EE widget or because of the relationship with EE widget contractor. Unlike spillover could occur among non-participants through interactions with participants or contractors energized by the program, for instance. These savings could



show up anywhere in the actual population where a link back to the program process motivates a change in consumption.

**Early replacement:** Early replacement refers to program participant households that are induced to install a high-efficiency widget when they do not intend to replace their existing widget at all. In this case, the program population circle does not stay within the circle of widget purchasers. If the program convinces households to replace their existing widget sooner than they would have, then it qualifies as an early replacement. The extension of the program circle outside of the widget-replacers subgroup is important because it illustrates the appropriate baseline for early replacement. In the counterfactual population, those households would have the consumption of their older, existing widget. Unlike the typical program savings that represent the comparison of new standard to new high efficient widget consumption, these households will have savings that reflect the difference between existing widget consumption and the new high efficiency widget consumption.

**Expansion to many programs:** The final steps of the figures illustrate the complexity underlying a true residential net-savings load shape. Many different programs with different approaches, different target measures, and different expected savings would interact in the population. There would be overlap across programs; there would be overlap across natural EE groups. The final comparison, for any given household would be between the counterfactual household's consumption with no program influence and the actual population household's consumption given their interactions with one or more programs or spillover via other pathways. The result would be net of free-ridership and inclusive of spillover and would fully and accurately capture the savings effect of the full residential portfolio of programs.

## 3.2 Counterfactual vs. randomized assignment experimental designs

The counterfactual example provides an ideal model that we would like to replicate in a real world evaluation. Of course, in reality, the perfect one-to-one match of the parallel universe counterfactual is not possible. Various randomized assignment experimental designs are the closest we can come to approximating the counterfactual.

In place of replicating a world in parallel universes, the RCT approach approximates the counterfactual by creating two similar groups prior to applying the program (or treatment). The groups are created by randomly assigning a subset of the population at question to a treatment or control group. The randomly assigned groups do not mirror each other identically on a household-to-household basis. In aggregate, though, the two groups will approximately mirror each other. The exact degree to which they mirror each other will be determined by the way the randomization is performed, the size of the two groups and the variation in the characteristics of interest. RCT replaces the thought problem of the counterfactual with the statistics of random assignment.

RCT experimental designs and randomized encouragement designs (RED) are similar in their use of a randomly assigned control group to approximate the counterfactual. The distinctions between the two approaches relate to program design and implementation of the design as well as savings estimate methodologies. For both RCT and RED, the intent is to capture the effect of a treatment as applied to the treatment group. In an RCT, all treatment group members are assumed to receive the treatment. In an RED, all treatment group members are encouraged to take the treatment and the program knows who does accept the treatment. For an RCT, the control group does not receive any treatment. For a RED, the control

---

---

---

does not receive any encouragement, but may or may not be able to access treatment depending on the set-up of the experiment. This option for the control group to receive treatment gives the RED approach additional flexibility for applications in the energy program space.

To estimate program effects for both RCT and RED, the evaluator compares consumption or other characteristics across the whole treatment and control groups. This was illustrated in the counterfactual figures shown previously. It may only be a subset of households that have changed consumption, but it is the average household effect across the whole treatment group versus the average household effect across the whole control group that provides the most basic estimate of savings. All RCT and RED methodologies build off of this basic, full group comparison. The challenge, for both RCT and RED, is that there is enough effect among treatment group households to be estimated with desired precision given the natural variability in the data and the randomness of the allocation. This is an issue for HER programs where many who receive the treatment likely do nothing.

This is an even more basic challenge for RED approaches where only the proportion of households successfully encouraged to take the treatment will register the effect. If a RED design can encourage 20% of the treatment group to adopt a treatment that will have a 10% effect on their consumption then the average effect across the whole treatment group would be just 2%. While getting this basic estimate of savings may be challenging, because the RED approach generally knows who receives the treatment, the overall average savings can be transformed to a “local” average savings for those who were treated.<sup>2</sup>

### 3.2.1 Randomized controlled trial experiments

While it is not the place of this whitepaper to explain the details of an RCT and RED, we provide a brief review of important resources that provide substantial information regarding RCT and other randomized assignment design options in the context of energy program evaluation. To set the stage for the review of these studies, we first review how a randomized assignment design performs like the counterfactual scenario as well as where it differs.

1. Like the counterfactual scenario, a randomized assignment design approximates the perfect and complete match between the treatment and control groups. Because of the randomization of the eligible population, the households in each group will be approximately equivalent prior to the application of a treatment. The distribution of subpopulations and relevant characteristics within the eligible population should be similar to the distributions within the randomly assigned treatment and control groups.

Unlike the parallel universe scenario, randomized assignment design requires that an eligible population be divided into treatment and control groups. Because of the random assignment process, it is not possible to create a fully valid control group for a whole, pre-determined group of households like a utility residential population. If the full utility residential population were the eligible population, then part of it would have to be removed to create the control group. The utility-level distribution of household characteristics would be represented in each subset but the whole, pre-defined group, like a utility residential population, cannot be made into a valid treatment group.

---

<sup>2</sup> The exact calculation of savings and the interpretation depends on whether control group households were allowed access to the treatment. The protocols discussed in the next section do a good job of describing these details.

- 
- 
- 
2. Given the randomly assigned treatment and control groups and a treatment offered solely to the treatment group, the randomized assignment design will identify an aggregate program effect in the same way as the parallel universe counterfactual. This will be the case if the treatment is the installation of an efficient widget or a behavioral treatment. HER programs are evidence of the RCT's ability to support the estimation of savings caused by the diverse activities that occur as a result of the mailed reports. Despite the fact that the savings are relatively small in aggregate, diverse in source and may be spread across a large proportion of the treatment group households,<sup>3</sup> large-scale, RCT experimental designs are the accepted method for organizing these programs because they support unbiased estimates of the program effect.

Unlike the counterfactual scenario, in an RCT or RED, the control group must be denied any direct or indirect effects of the treatment or the encouragement that is being tested. This requirement creates a number of practical challenges including:

- Denying (or delaying) program benefits may be problematic for utilities on equity terms. Program rebates and services frequently offer customers substantial monetary and non-monetary value. Denying those benefits to a subset of customers on an arbitrary basis is difficult for utilities to justify even in the name of supporting good program evaluation. RCTs are used successfully to evaluate home energy reports because the reports are a low-cost (limited repercussion) service the utilities can deny to households without creating inequities between the treatment and control group.
- Targeting treatments to a single group creates difficulties for program delivery. RCT Programs cannot use general media marketing and/or trade-assisted delivery approaches because they may contaminate the control group.

Utilities are also not willing to force treatments on customers. It will always be easier to distinguish a savings effect from the noise of natural consumption variation if the savings are present in all treatment group households. However, it is not feasible for utilities to get all randomly assigned treatment group households to replace their furnace.<sup>4</sup> A portfolio-wide RCT would require denying the control group households access to all direct program advantages as well all non-participant spillover. Even if equity were not an issue, the practical application of a randomized assignment design for many programs across a large dynamic population would simply not be feasible.


3. Finally, within the confines of a feasible design, an RCT control group provides an approximate tracking of treatment group change over time. With regards to the important time-varying characteristics, both exogenous factors (e.g., weather and economics) and changing household characteristics will be tracked by a randomly assigned control group. Given that these exogenous effects may be of a similar magnitude as the aggregate program effect (especially for behavior programs, in general) this aspect is essential.<sup>5</sup>

---

<sup>3</sup> Survey efforts to better understand the drivers of HER program savings have had limited success likely because of the large number of diverse and small sources of savings.

<sup>4</sup>The RED approach addresses this issue by randomizing the encouragement to do some savings or demand control activity. It only partially solves the problem; however, as the success of the approach is still determined by the combination of the magnitude of the effect, proportion of households encouraged to take the treatment and the overall numbers in the experimental population. In practice, it can be challenging to locate a sufficiently large starting population and encourage a sufficiently large subset of the treatment group to support a viable estimate.

<sup>5</sup> See section 4.1.2.2 below where non control groups were assessed on their ability to track consumption over time.



The counterfactual scenario tracks change over time because the whole system is duplicated. Households changing occupancy, people coming and going—those dynamics would simply occur in both worlds. In an actual RCT, these kinds of dynamics are challenges to maintaining a valid control group. Natural attrition decreases the size of both treatment and control groups over time. New households are difficult to include in randomized assignment designs that start at a particular point and frequently include a requirement of consumption history.

In summary, randomized assignment designs do maintain some of the key characteristics of the parallel universe counterfactual. While each household in the treatment group will not have its perfect mirror in the control group, the control group will, on average, contain households that mirror the key attributes of households in the treatment group. On the other hand, there are a range of practical challenges to implementing even individual programs in an RCT or RED context and a portfolio-wide RCT is not feasible.


The natural EE adopters are the group that is the most important to track in this discussion. A random allocation should produce treatment and control groups that have similar natural EE adopter groups in terms of proportion of the population and the range of unobservable characteristics that define that group. In this respect, they will mirror the green circles in the actual and counterfactual worlds, respectively, in the figures. In the aggregate, comparison of consumption across the full treatment and control groups, the control group natural EE households will provide an appropriate baseline for the treatment group's natural EE households. If a program were offered to the treatment group then, then only additional program-motivated, net savings would be captured in the aggregate consumption difference between treatment and control groups. That is, an RCT or RED design does have the capability to provide valid net savings estimates of savings within a limited, feasible experimental design.

### 3.3 Using RCT for program evaluation

It is easy to understand the desire to take advantage of randomized assignment at any level of EE program evaluation. There are aspects of randomized assignment designs that support the ultimate goals of a portfolio savings. In particular, addressing free-ridership and complex interactions among programs are two particular strengths of RCT and RED designs. However, as discussed above, practical challenges of randomized assignments make them unlikely to proceed for large scale, portfolio scenarios.

There is a growing interest in the economics community in using RCT as the foundation for program evaluation. This interest stems in part from the potential method bias inherent in any observational study, something the EE evaluation community has long struggled with. DNV GL's whitepaper *Evaluating Opt-In Behavior Programs* lays out the challenges. A widely cited example of this newfound interest in randomized assignment designs is the recent paper by Fowlie, Greenstone, and Wolfram, *Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program*. For this analysis, the authors used a RED approach to evaluate aspects of the federal weatherization program.

In this section we provide an overview of approaches have been proposed that insert randomized assignment into a practical evaluation framework. Across the examples, there is a theme that is consistent with the counterfactual scenario illustrated above. The effect that is measured by randomized assignment designs is always incremental to the condition of the population prior splitting into parallel worlds or the randomized assignment. It is not particularly difficult to randomly assign groups. The challenge is recognizing the full implication of the counterfactual—who do the treatments affect and what is the baseline in the counterfactual.



This challenge can be understood also in terms of internal and external validity. A properly implemented randomized assignment design provides a rigorous, unbiased estimate of the treatment effect on the particular population studied (internal validity). Applying that result to other populations, treatments, or times requires justification of the similarity of those conditions (external validity) and is always subject to some question. In the context of program evaluation, even for a single program, it is difficult to implement a randomized assignment design that allows the full functioning of the program both to be available to the general population and to be observed as a randomly assigned treatment effect. The HER model does have full internal validity for a full program, but at the cost of restricting the program offering to the randomly assigned customers.

The following six example scenarios further illustrate the challenges and limitations of applying random assignment as the basis for program evaluation at the individual program or portfolio level:

1. A full program is delivered (with opt-out option) only to a randomly assigned treatment group. The randomly assigned remainder is the control. This is the model used first by Opower, and now others, for HER Programs. This approach has proved acceptable to utilities and regulators because the reports provide no tangible benefit to the treatment group and, thus, nothing is denied the control group. As discussed above, it is more difficult to get buy-in if tangible benefits would be denied to the control group.
2. A regular program operates as is, with an incremental encouragement/recruitment provided on a random assignment (RED) basis. This is the Fowlie et al. model. The limitations to this approach include:
  - a. It measures the effect of the program on the incrementally recruited customers, not on those recruited through standard channels. The program effects on these incrementally recruited customers may not be representative of savings for customers recruited through standard channels. For example, they could have lower savings potential which would explain why they were insufficiently attracted by the base program to join it without the additional encouragement.
  - b. It may be fairly expensive. RED may require recruiting very large recruitment sample to obtain relatively small incremental participation (this was experience for Fowlie et al.).
  - c. It may produce small average effects when averaged across the whole treatment group due to low opt-in rates. This can lead to low precision on results from simple, standard approaches. More complicated models may extract statistically significant parameter estimates but this in turn introduces again the potential of model specification bias.
3. A regular program operates as is, with incremental offerings made available on a random assignment basis. This approach can be useful for testing potential program modifications. The limitations to this approach include:
  - a. It measures the effect of the incremental program in the presence of the base program, not the effect of the base program. This is useful for assessing the value of the increment, but not for determining the full program effect.
  - b. It may get low opt in for the incremental offerings, which could result in similar concerns to those listed in 2b and 2c as above.
4. A regular program operates as is, with a randomly assigned subset of customers deferred to the next year. This approach can be implemented with or without customers knowing about the



possibility of deferral in advance. This design has been used for time-varying rates pilots, where a self-selected group interested in trying the new rate is randomly assigned to current year or next year. A “natural experiment” form of the design is used when wait-listed or next year’s participants are used as a control group. The limitations to this approach for most EE programs include:


- a. Customers go to programs when they want to buy something they need. Deferring a year when someone is ready to remodel their house or needs a new water heater or furnace will annoy people, whether it’s identified as a program condition up front or only when they ask for a rebate. For many programs, customers become aware of the program opportunity only when they already want to do something. For instance, when someone goes to buy a new refrigerator they may find there is a rebate for an efficient unit as well as a recycling program.
- b. Customers who are denied access have a wide range of possible responses and in aggregate there is no reason to believe the full set of responses will reflect the ideal counterfactual of no program available. The first two options below are the two recognized in the counterfactual scenario above. The remainder are options introduced by the particular design:
  - i. Customer will buy the needed widget, but not get the efficient one, as they would not have without the program (would-be participant and non-free-rider, appropriately counted by the Control minus Treatment difference).
  - ii. Customer will go ahead and buy the efficient widget without the rebate, as they would have done without the program (would-be participant and free-rider, appropriately counted in the Control minus Treatment difference).
  - iii. Customer who would otherwise have bought a widget this year will instead wait until the next year and buy EE (provides baseline of older existing widget, when the ideal baseline is a standard efficiency widget).
  - iv. Customer will go ahead and buy the efficient widget now without the rebate, but they wouldn’t have done it without the program (would-be participant who becomes spillover and counts against the program in Control minus Treatment difference).
  - v. Customer will buy the needed widget, but not go EE, even though with no program they would have gone EE, because they’re annoyed at being denied (negative program impact due to RCT design).

The resulting control versus treatment difference could have any mix of these possible outcomes.

1. If the goal is for the control group to represent purchasing a widget with no rebate ever available, then the control group should include the first two options (i and ii) in the correct proportions. There is no way of avoiding the presence of the other options or assuring the proportion reflects the true counterfactual.
2. If the goal is for the control group to represent buying EE vs not buying anything, then the control group should include ii and iii but not i. That is, the baseline should include the correct mix of doing nothing or purchasing EE, but no standard widgets are purchased. This is not feasible to implement or control.

The recruit/delay approach is useful if customers are unlikely to take program-encouraged action on their own, as in low-income programs or time-varying rate. The quasi-experimental variation-in-adoption approach, commonly used for low income programs, takes advantage of this dynamic as well.



- 
5. A regular program operates as is, a (small) random subset of the population is denied a rebate if they ask for one but given a buy-off payment. Similar to the recruit and delay/deny approaches, the denial will affect the behavior of the control groups in ways that may not reflect the simple absence of the program.
    - a. Customers will buy the efficient widget they would have bought through the program and use the buy-off payment in lieu of a rebate. This group includes some natural adopters and some program-induced adopters. Program-induced adopters within the control group have the effect of spillover counting against the program in the Control minus Treatment difference).
    - b. Customers will buy the standard efficiency widget and use the payment to defray that cost. This would include mostly non-natural adopters.
    - c. Customers will buy the standard efficiency widget and use the payment to buy other energy-using equipment, or to increase use of existing equipment. This would include mostly non-natural adopters.

The overall effect would include both program-induced increases in consumption and program-induced EE purchases in the control group. Ultimately this model measures the difference between offering the program as is and offering program marketing and a payment without a tie to EE actions. This is not a measure of the effect of the program.

6. A full portfolio of programs operates as is; a small random subset of the population is denied access to any programs if they ask, but given a buy-off payment instead.
  - a. This case has all the issues of case #5.
  - b. The program-induced EE in the control group would be the effect of multiple programs, not just one.
  - c. The inability to limit control group customers access to upstream programs means that upstream programs are part of the base and not included in the effects measured by the RCT design. Any additional upstream activity due to the buy-off payments would count as spillover and would count against the non-upstream program effect measured in the RCT design.

These different scenarios illustrate the importance of understanding against what counterfactual a program effect should be measured to obtain the desired result. While the ideal counterfactual is frequently easily envisioned, in actuality creating the right conditions for a control group that reflects that counterfactual is not straightforward. The above examples provide illustrations of the process of identifying the range of possible responses to a program design and then a consideration of whether there is any mechanism in the design that would lead the evaluator to expect those responses to mirror the idealized counterfactual.

### 3.4 Literature review

The purpose of the previous section on RCT Context is to explore an ideal experimental design and define the challenge of building experimental methods that approximate the. In this section, we give an overview of the important reports, papers and evaluations that discuss RCTs in general and their role in energy program evaluation.

The inspiration for this idealized experimental design clearly grows out of the success of RCTs in the energy program area over the last several years. The evaluations coming from those designs offer high levels of validity and precision on particularly difficult-to-measure programs. It is little wonder that we would dream

---

---

---

of an application at the system level. In contrast, the documents highlighted here are all practically oriented works. They include protocols discussing how to set up and use an RCT correctly and examples from evaluations and other analyses.

### 3.4.1 Evaluation protocols for random designs

The following three protocol documents address the evaluation of behavior-based interventions designed to affect energy consumption. The documents provide a primer on the practical mechanics of RCTs and REDs. They discuss the strengths of such designs including that RCTs and REDs make it possible to produce estimates that are on average, unbiased estimates of effects causally linked to a specific treatment. They also discuss the practical challenges of such designs. In addition to the limitations discussed above these include attention to sample sizes to meet precision requirements and the related challenge for REDs of getting sufficient uptake of the encouraged treatment

- SEEAAction (2012): State and Local Energy Efficiency Action Network. 2012. Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations.
- LBNL (2013): Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines. Peter Cappers, Annika Todd, Michael Perry, Bernie Neenan, and Richard Boisvert. Environmental Energy Technologies Division
- UMP (2015): Chapter 17 of the UMP discusses the recommended Residential Behavior Protocol, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures. James Stewart, Annika Todd.

SEEAAction and UMP specifically address behavior-based programs that use a range of techniques from the social sciences to affect energy consumption behavior. The CBS report describes its goal as to understand how “the incentives and information embedded in time-based rates, enabling technology, feedback strategies, and other treatments” affect energy consumption”(p. 6). With regards to discussion of randomized assignment design, there is a great deal of overlap between the three protocols. Most importantly, there is full agreement on the place of randomized assignment experimental design in the top of the hierarchy of options.

These primers share a similarity with this whitepaper in that they explore the connection between program design and estimation methodology. They discuss the range of methods that are used across randomized and quasi- experimental designs in behavior and pricing programs. Generally, randomized designs support more simple estimation approaches that produce robust results with few assumptions. In the RCT context, a simple difference in means is an unbiased estimate of program effects without any qualifying assumptions. There are a variety of methods that leverage the randomized design to offer results with greater precision but the design is the source of the internal validity of the estimates. Practical calculation of RCT or RED savings is not the focus of this whitepaper. We refer the reader to these documents to find a summary of the range of methods used to estimate savings in a randomized assignment experimental design.

Quasi-experimental designs require more complex models with greater assumptions. It is essential to discuss specific quasi-experimental design methods because it is through the methodology that these approaches address the lack of the proxy counterfactual available in a randomized assignment design. The range of methods discussed in the SEEAAction and UMP documents are the focus of the next section.



### 3.4.2 HER program evaluations

HER programs are the most common example of RCT experimental design in use in energy program evaluation today. In many ways, HER programs are responsible for the high profile of randomized assignment designs in energy program evaluation today. The RCT design addressed the challenge of accurately estimating small, variable consumption reductions motivated by information and behavioral cues.

HER programs work within a pre-defined eligible population using randomized assignment designs to produce treatment and control groups. Reports are mailed to only the treatment group. The combination of the randomized assignment design and the large size of the treatment and control groups support unbiased estimates of treatment effects with a high degree of precision.

Similar to the protocols discussed above, HER program impact evaluations offer examples of the range of methods that can be used in an RCT context. Many reports build from simple mean difference-in-difference estimates to monthly fixed-effect methods that replicate the difference-in-difference structure in a regression form. HER program designs have also evolved over time in the recognition that randomizing with strata or matched pairs can decrease both the potential of unbalanced samples and reduce variation.

## 4 QUASI-EXPERIMENTAL METHODS: THE SPECIAL CASE FOR TOP-DOWN MODELING

As indicated in Section 2, true RCT is possible only for certain program types and must be built into the program design. More commonly, programs attempt to attract customers, and customers have the option to participate or not. Indeed, strategies to attract customers are an essential part of most program designs. Put another way, self-selection is inherent to the design of most programs.

Absent randomized assignment, the assumption that the participants and nonparticipants are equivalent in factors affecting consumption or changes in consumption, apart from the program effect itself, is generally not supported. Within the general context of comparison group approaches, alternative methods are needed to attempt to isolate net program impacts (i.e., controlling for factors that contribute to underlying differences between participants and non-participants). The challenge of isolating net impacts becomes even greater when attempting to capture comprehensive effects of a program portfolio, accounting for free ridership, spillover, interactions, and take-back.

In this section, we first provide a brief overview of common quasi-experimental methods that may be applied at the program level. We then describe an alternative approach, top-down modeling, and discuss how top-down models are and are not similar to the more traditional quasi-experimental methods. We also review the associated advantages and limitations of top-down modeling for the current portfolio-level analysis. To this end, we review key findings from recent top-down studies, focusing primarily on the California work, and also drawing on recent investigations in Massachusetts. Finally, we discuss a number of key issues that would arise if we were to extend the general top-down framework to the hourly level.


### 4.1 Overview of quasi-experimental methods for program evaluation

#### 4.1.1 Methods summary

The SEEACTION Report and UMP Chapter 17 provide a comprehensive summary of alternative comparison group methods for EE program evaluation. These methods focus on post-hoc identifying a comparison group that is as similar as possible to the participant group of interest. These methods include the following:

- **Pre-post analysis:** In the absence of a comparison group, pre-post analysis relies on each participant as its own comparison case, implicitly assuming all changes in consumption are associated with the program. When combined with a comparison group, Pre-Post analysis becomes the **difference-in-differences** method.

In terms of the counterfactual scenario discussed in Section 2, the prior period snapshot of the program participants replaces the counterfactual as the basis of comparison. The non-time-varying group characteristics are the same between the pre- and post-program periods but disentangling the program and non-program changes is problematic. In addition to these potential method uncertainties, pre-post analysis provides a gross savings estimate; natural adopters may be present but are unaddressed. Adding the comparison group generally improves the ability to separate program and non-program effects but, in turn, risks bias associated with poorly matched observable and unobservable characteristics between the treatment and control groups. Free-ridership will only be addressed to the extent the comparison group has comparable levels of natural EE adopters. This is not feasible given the



difficulty of identifying natural adopters and finding sufficient number outside the program to populate the comparison group.

- **Variation in adoption:** Later program participants serve as a comparison group for current program participants. This approach is reasonable if the program is stable over time, and there are unlikely to be major non-program factors that affect both the decision to participate in one period rather than another and energy consumption. This approach is often used for low-income weatherization programs, where participants in different years are likely to have similar characteristics, and are unlikely to be undertaking EE on their own. This approach is most effective when subsequent participants are included in the regression who did not participate until after the evaluation time frame. This is commensurate with adding a comparison group. This approach relies on the households who have either already participated or are yet to participate to inform a time-series effect that captures exogenous trends. The addition of subsequent participants beyond the evaluation time frame addresses that concern more directly.
- **Regression discontinuity:** Participants who are just beyond an eligibility threshold serve as comparison group for participants just below the threshold. The threshold may be a geographic boundary or a quantitative factor such as consumption level. The method can work well if there are not a lot of other differences associated with being on one side or another of the threshold, such as prices or access to other services. In terms of the counterfactual scenario discussed in Section 2, regression discontinuity posits a subset of the widget installing population that, because of the eligibility threshold, looks approximately like the remainder with respect to natural EE adopters but does not have the program.
- **Matching on observable characteristics:** For each participating unit, one or more comparison cases are selected from the eligible non-participants, matching on available characteristics. These characteristics may include consumption or demand level, geography, indicators of space heating or water heating fuel, indicators of air conditioning, dwelling unit type, and participation in other programs. Neighborhood socio-economic factors may also be included based on census data, or these may be implicitly included in the geographic mapping.

In terms of the counterfactual scenario discussed in Section 2, this would involve locating similar households outside of the program but preferably within the widget installing population to populate the comparison group. Matching can locate households with quite similar consumption characteristics but will not match by unobservable characteristics like natural EE adopter of widgets. However, free-ridership will only be appropriately accounted for, thus giving a net savings estimate, if the exact same proportion of natural adoption is present in the comparison group as opted into the program.

- **Propensity scoring:** The propensity to participate is modeled as a function of the observable characteristics, fitting the model across the participants and non-participants. The resulting model is used to calculate a propensity score for each participant and non-participant. Non-participants are matched to participants based on this participation propensity score. This approach essentially condenses the observable characteristics into a single dimension that reflects the self-selection into the program inclination, to the extent that inclination is associated with the observable characteristics. Propensity scoring is similar to matching on observables with respect to its inability to identify natural EE adopters.




Additional methods not described in SEE Action include:

- **Comparison region:** A geographic region where a program is not offered is used as a comparison case for a region where the program is offered. Variations on this approach are common for assessing the effects of upstream programs or other types of programs where there is no explicit tracking of which individuals participated or did not. In this type of analysis, the unit of analysis is typically a geographic region as a whole. In terms of the counterfactual scenario discussed in Section 2, a comparison has the potential to be as close to the ideal counterfactual as any other method. It represents a parallel reality without programs. The challenge is controlling for all of the differences in population characteristics.
- **Regression modeling of individual customers:** Individual customers' consumption is modeled as a function of program participation and other observable characteristics. The model combines participants and nonparticipants, or participants at different times. In effect, the regression model creates an implicit "match" for each participant representing what the participant would have consumed absent the program, rather than finding the closest available match among actual non-participants. One advantage is greater flexibility compared to explicit matching, particularly in situations where there are few close matches for some types of customers. Another is the availability of statistical diagnostics to indicate which characteristics are and are not valuable in explaining variations across customers. A disadvantage is the dependence on the model structure to produce the no-program estimate for the participants.
- **Aggregate regression modeling:** Aggregate consumption for a geographic region is modeled as a function of the level of program activity and other observable characteristics of the region. The model is fit across regions with varying levels of program activity and other characteristics. Three broad forms exist:
  - Cross-sectional aggregate models are fit across geographic units within a single point in time or for a single pre-post change. Such models rely on variation across units in the mix of levels of program activity and other characteristics to separate the effects of program activity.
  - Time series aggregate models are fit within a single geographic unit across time periods. Such models rely on variation over time in the mix of levels of program activity and other characteristics to separate the effects of program activity.
  - Time series cross sectional or Panel models are fit across time periods and geographic units. Such models have the advantage of variation over both time and geography in the mix of levels of program activity and other characteristics.

Various combinations of these methods also are used. Pre-post analysis is often used in combination with comparison groups defined by various means. When individual regressions are used with participants and comparison cases, it is common to restrict the comparison group to cases similar to participants with respect to some key features. Propensity scoring may be used not just as a basis for matched comparison, but also as an explanatory variable in a regression model.

As summarized in Table 1, these methods can be characterized by several key features:

1. The unit of analysis: individual customer or geographic region.
2. The time frame of analysis: a single period, single difference of periods, or multiple time periods.
3. The form of matching: explicit assignment of individual comparison cases to individual participants; implicit construction of individual no-program cases via regression; or comparison of the participant group as a whole with a "similar" non-participant group.

- 
4. The basis for the comparison that provides the savings estimate, with the associated implications regarding what factors are and are not well controlled for by the comparison group construction.

**Table 1. Summary of common quasi-experimental methods for program evaluation**

Method	Unit of Analysis	Typical Time Frame	Matching Form	Basic Comparison	Key Non-Program Factors Controlled for	Key Factors Not Controlled for	Conditions where Useful
<b>Pre-Post</b>	Customer	Single difference	Explicit (self)	Pre- vs post-participation	Participant prior condition	General trends Non-program changes associated with participation decisions Free ridership	No meaningful comparison group, non-program changes minimal or separately calculated
<b>Variation in Adoption</b>	Customer	Single year or single difference	Group	Customers who have already participated vs. those who are about to	General trends General self-selection factors	Non-program changes associated with participation decisions Free ridership	Stable program, minimal association between participation timing and timing of other major changes at the premise
<b>Regression Discontinuity</b>	Customer	Single year or single difference	Group	Customers just within vs. just beyond eligibility	Factors that are similar across thresholds	Non-program factors that differ across thresholds	Minimal differences across thresholds; ability to estimate overall program effect from the edge cases
<b>Matching on Observable</b>	Customer	Single year or single difference	Explicit	Participants vs. customers who are similar in other observable characteristics	Changes associated with the match variables	Self-selection factors not associated with the observable match variables	
<b>Propensity Matching</b>	Customer	Single year or single difference	Explicit	Participants vs. customers who would have had a similar propensity to participate, based on other observable characteristics	Underlying propensity to participate	Self-selection factors not associated with the observable variables used for the propensity modeling	
<b>Comparison Region</b>	Geographic area	Single year or single difference	Explicit	Region with an active program vs. one without	General trends that are similar between the regions	General trends that are different between the regions	Upstream programs and market effects
<b>Customer-level regression</b>	Customer	Single year, single difference, or multi-year	Implicit	Incremental effect of each factor in the model, all others held constant	General trends and factors accounted for in the model	Factors omitted from the model or mis-specified	
<b>Aggregate Regression</b>	Geographic area	Multi-year	Implicit	Incremental effect of each factor in the model, all others held constant	General trends and factors accounted for in the model	Factors omitted from the model or mis-specified	



#### 4.1.1.1 Needs of comprehensive portfolio impact estimation

For portfolio impact estimation, the goal is to capture a number of effects. These effects include:

- Free ridership, or naturally occurring savings among participants: A free rider is a program participant who would have implemented the program measure or practice in the absence of the program (p. 422)<sup>6</sup>
- Spillover: Reductions in energy consumption and/or demand in a utility's service area caused by the presence of the DSM program, beyond program related gross savings of participants. These effects could result from: (a) additional EE actions that program participants take outside the program as a result of having participated; (b) changes in the array of energy-using equipment that manufacturers, dealers, and contractors offer all customers as a result of program availability; and (c) changes in the energy use of non-participants as a result of utility programs, whether direct (e.g., utility program advertising) or indirect (e.g., stocking practices such as (b) above, or changes in consumer buying habits). (p. 442)
- Market effects: A change in the structure or functioning of a market or the behavior of participants in a market that result from one or more program efforts. Typically, these efforts are designed to increase in the adoption of energy efficient products, services, or practices and are causally related to market interventions. (p. 430)
- Take-back effects: A change in energy using behavior that yields an increased level of service and that occurs as a result of taking an EE action. (p. 438)
- Interactive effects on other participant equipment or systems: The interactions between the measure and a non-measure-related end use (e.g., efficient lighting generally reduces cooling loads) and interactions between packages of measures that can cause the sum of the measure package savings to be less than the sum individual measure savings. (p. 164)
- Upstream program impacts: Impacts from programs that provide information and/or financial assistance to entities in the delivery chain of high-efficiency products at the retail, wholesale, or manufacturing level. (p. 446)
- Interactive effects of multiple programs.

All of these effects have been studied to varying degrees and by various methods for individual programs. Almost all such studies are characterized by uncertainties, and sometimes by considerable controversy, related to method limitations and assumptions.

Table 2 provides a high-level summary of the effects that quasi-experimental methods are typically designed to account.

---

<sup>6</sup> All definitions come from the The California Evaluation Framework Prepared for the California Public Utilities Commission and the Project Advisory Group June 2004 Last Revision: January 24, 2006. <http://www.cpuc.ca.gov/PUC/energy/Energy+Efficiency/EM+and+V/>

**Table 2. Effects typically designed to be captured by quasi-experimental methods**

Method	Free Ridership	Participant Spillover	Non-Participant Spillover	Take-back	Upstream Program Impacts	Interactive effects on other participant equipment	Interactive Effects of Multiple Programs
Pre-Post	O	P	O	X	O	X	O
Variation in Adoption	O	P	O	X	O	X	O
Regression Discontinuity	X	P	A	X	O	X	O
Matching on Observable	P	P	A	X	O	X	O
Propensity Matching	P	P	A	X	O	X	O
Comparison Region	X	P	P	X	X	X	O
Customer-Level Regression	P	P	A	X	O	X	O
Aggregate Regression	X	X	X/A	X	X	X	X

**X: Captured by method design**

**P: Partially captured**

**A: Anti-effect. Actual non-participant spill-over from other areas or years tends to reduce the estimated effect when this method is used**

**O: Method is not designed to capture this effect**

Free-ridership is not accounted for by pre-post methods or variation in adoption methods. Both methods include only program participants and have no way to control for naturally occurring savings that may occur among participants. Both regression discontinuity and comparison regions compare similar customers who are and are not eligible for the program; hence, the comparison group includes the no-program natural adoption rate and accounts for free ridership. Matching and regression methods account for free ridership to the extent the comparison group construction or modeling is able to mitigate self-selection effects. The challenges facing these methods were discussed in the figures in Section 2.

Participant spillover is accounted for by comparison group consumption analysis only to the extent the spillover activity is completed within the timeframe of the analysis. Longer term spillover is typically not captured unless a long-term study is conducted.

Non-participant spillover not only is not captured by most comparison group analysis, but also tends to reduce the estimated program effect when present. If there is spillover from the participant group to the

comparison group, the estimated program effect is reduced by the amount of the nonparticipant spillover, rather than being increased by that amount. When the comparison group is another region then there is no spillover from the program being studied.

Take-back and equipment interactive effects are captured in most methods based on consumption data analysis. Upstream program impacts are typically not captured via individual customer consumption analysis, because “participants” or level of program uptake is not identifiable at the customer level, by program design. Upstream impacts can be captured by appropriately designed regional comparisons.

Only regional studies are well suited to capturing interactive effects across a portfolio of programs.

Aggregate regression or top-down modeling does not automatically solve all method problems or remove all method uncertainty, but in principle has several attractive features for capturing the full set of impacts across a program portfolio.

- Analysis over multiple years provides the opportunity to capture spillover and market effects that take time to develop.
- The regression structure implicitly accounts for all the program effects and interactions, without requiring (or providing) explicit estimates of each individually.
- The estimation process yields a measure of statistical accuracy for the combined portfolio-level estimate.

In Section 4.1.4, we review some of the empirical experience with top-down approaches and some of their limitations.

## 4.1.2 Regression-based quasi-experimental methods

### 4.1.2.1 Relation between difference in difference and basic regression

A basic savings calculation using a comparison group is:

$$S_{\Delta\Delta} = \Delta T_{-} - \Delta C_{-}$$

Where

$S_{\Delta\Delta}$  = participant group savings per customer by the difference of difference method

$\Delta T_{-}$  = average pre-post change among the treatment (participant) group

$\Delta C_{-}$  = average pre-post change among the comparison group

This savings structure can also be expressed in the form of an Ordinary Least Square (OLS) regression as

$$Y_{tgj} = a + bPOST_t + cTREATED_g + d POST_t * TREATED_g + e_{tgj}$$

Where

$Y_{tgj}$  = consumption for unit j of group g in year t

$POST_t = 0$  for time t = pre, 1 for t= post

$TREATED_t = 0$  for group g = Comparison, 1 for group g = Treated (participant)

a, b, c, d = coefficients estimated by the regression

$e_{tgj}$  = residual error.

This regression formulation will yield the same estimates as the basic difference in difference calculation. That is, the regression coefficients are as indicated in Table 3.

**Table 3. Basic regression equation and difference-in-difference estimator**

Coefficient	Interpretation
$a = C1\_$	Base usage is estimated by comparison group year-1 average
$b = C2\_ - C1\_$	General trend or non-program “post” effect is estimated by the comparison group year-1-year-2 difference
$c = T1\_ - C1\_$	Initial treatment-comparison difference is given by the year-1 difference in averages
$d = (T2\_ - C2\_ ) - (T1\_ - C1\_ )$ $= (T2\_ - T1\_ ) - (C2\_ - C1\_ ) = S_{\Delta\Delta}$	Program effect is the difference of differences, the treatment change minus the comparison group change

The use of the difference of differences estimator  $S_{\Delta\Delta}$  helps control for any initial difference between the Treatment and Comparison groups, whether those differences are random or not. Without a randomized assignment however, there is the potential for systematic, non-random differences between the two groups that could affect the pre-post difference apart from treatment. That is, the participants and non-participants may differ in ways that would make the participants’ change different from nonparticipants. Section 2 describes various reasons why this may be so.

A key purpose of moving to regression-based savings estimation is to incorporate additional factors into the estimation, to control for other non-program factors that may account for pre-post differences. For example, if participants tend to be customers undertaking a major renovation, and we had a variable (likely from survey data) indicating major renovation activity for nonparticipants, we could include that in the model. If we have to rely only on existing data sources, we might use indicators such as income or house size to control for factors that could be associated with different levels of consumption change apart from the program.

#### 4.1.2.2 Performance of quasi-experimental methods

In most applications of quasi-experimental methods, there is no opportunity to cross-check the results against a more accurate source. Such an opportunity was offered in a recent study sponsored by Pacific Gas and Electric Company (PG&E) (Schellenberg et al. 2015). A HER program was analyzed using the RCT design by which the program was implemented. Three quasi-experimental approaches were applied to the same program. The results provide a rare glimpse into the reliability of these methods when they are the only methods available, and the “gold standard” estimate does not exist.

The gold standard savings estimate from this program was essentially a monthly difference-of-differences estimate. The quasi-experimental methods included propensity score matching (PSM), Bayesian Structured Time Series (BSTS), and Regression Tree with Random Effects (RE-EM Tree). The study found that in terms

of annual savings, the three alternative estimates all varied from the RCT estimate and from each other. They also exhibited high variability month to month, which the RCT estimates do not. The BSTS and RE-EM methods both involve modeling future usage based on initial usage, and include temperature terms. These methods both seemed to run into more trouble in the third evaluation year when the temperature was unusual. Results are summarized in Table 4.

**Table 4. Comparison of RCT estimates with quasi-experimental estimates, Schellenberg et al.**

Estimation Method	Percent savings			Monthly savings (kWh)	
	2012	2013	2014	Low	High
RCT	1.16%	1.58%	1.69%	5.9	12.2
PSM	2.08%	3.07%	3.21%	9.4	23.4
BSTS	-0.40%	2.21%	6.43%	-36.4	73.1
RE-EM Tree	2.03%	3.07%	6.08%	-25.8	89.3

These findings do not invalidate quasi-experimental methods. They do, however, underscore the reasons RCT was introduced into the OPower design. That is, when the effect of interest is small on average, there is a greater potential for method bias to swamp the true effects. RCT provides the only method that is rigorously unbiased and has sufficient statistical precision to yield reliable estimates in this context. This method, however, requires that the service being evaluated can be delivered based on random assignment.

BSTS and RE-EM Tree have not been used extensively for EE program evaluation. Potentially applications of these methods can be refined to improve their accuracy in this context. What’s not clear is whether in a context of higher total savings the PSM methods would be expected to yield estimates that are off by a factor of two, or estimates that are off by one to two percentage points.

### 4.1.3 Top-down modeling approaches

Top-down models attempt to measure changes in energy consumption over time that are attributable to programmatic interventions by the utilities. The goal of this modeling technique is to isolate the effect of program activity using a holistic approach, estimating program impacts across all EE programs in a given geographical region or service territory, rather than estimating savings with separate studies for each program or measure/end-use within a program. Top-down techniques use a holistic approach by estimating program impacts across all energy-efficiency programs in a given geographical region or service territory, rather than running separate studies for each program (or measure/end-use within a program).

Top-down methods estimate portfolio-level effects by fitting a regression model across time and across geographic units. The dependent variable is aggregate energy consumption that is typically normalized to measure consumption per unit (i.e., per household, per square foot, or per employee). The independent variables include one or more measures of program activity that may include aggregate program expenditures, ex ante savings, or incentive costs, measures of economic activity such as employment or GDP, and energy prices.

Thus, a very general model could be of the form:

$$Y_{tgj} = a + \sum_k b_k P_{kgt} + cI_{gt} + dA_{gt} + e_{tgj}$$

Where for geographic unit  $g$  in year  $t$

$Y_{tgj}$  = consumption in time period  $t$ , geographic region  $g$

$P_{kgt}$  = price of fuel  $k$

$I_{gt}$  = personal income

$A_{gt}$  = program activity

$e_{gt}$  = residual error

and

$a, b_k, c, d$  are coefficients estimated by the regression.


It is appropriate to fit these models separately by sector. At a minimum, separate models should be estimated for Residential and Non-residential sectors, but previous studies demonstrate that non-residential models should be separated into small commercial, large commercial and industrial models, and if possible, separate models should be run by industry (i.e., retail/office, manufacturing, healthcare, and education).

In principle, if the models provide good representations of all non-program factors affecting energy consumption over time and over geographies, the coefficients for the program activity variables give estimates of the incremental change in consumption associated with an incremental unit of program activity.

As described in Section 3.1.2, in a properly specified model, with sufficient data to provide accurate estimates, the coefficients on program variables capture the full cumulative effect attributable to the program portfolio. Included in these effects are the combined effects of all programs in the portfolio over the time frame captured in the program metrics, incorporating spill-over, cross-program interactions, physical interactions within premises, and customer take-back effects.

Some previous authors of studies have argued that properly specified top-down models also control for free-ridership, as they account the impact of price differences on the adoption of energy efficient equipment. While the price of fuel is relevant, the more relevant price for free-ridership is the price of incentives. For price effects to properly adjust for free-ridership, however, models must account for measure costs, or at a minimum, incentive, thereby reflecting the price of technology. Further, there must be sufficient variation in measure costs to reflect the condition of the un-subsidized measure cost. Because incentives do not vary substantially within the state, it is still necessary to extend the analysis to other regions to isolate free-ridership.

Therefore, in order for top-down modeling to work, the study must still include a comparison (non-program) area to serve as the counterfactual. One way that researchers have attempted to circumvent this issue is by extending the time series to a period of limited or no incentives. However, this approach is likely to introduce many exogenous factors for which the models cannot realistically control, such as structural changes to the economy and differences in the energy intensity of firms resulting from the advancement of information technology. In addition, using a long historical time series will actually reduce the level of



program activity everywhere, so the model will not be able to isolate differences in consumption resulting from differences in programmatic activity. Ideally, the models require differing levels of programmatic activity across the study region, with some being zero. This is unlikely within a single state or even region. Consequently, models are more effective in isolating free-ridership at a national level as a proxy for overall impacts, as they provide for more greater diversity in programmatic activity for a given time period, but national level models are limited in their ability to yield the refined estimates we require for state or local impact evaluations.

The key challenge then is being able to specify a model appropriately, and obtaining the data to support it. Data availability is key to the model specification as well, since the model must rely on the data available. Data issues are discussed further in Section 4.1.5.

#### 4.1.4 Results of recent top-down studies

California has investigated top-down analysis methods for portfolio-level savings analysis in a series of methodological studies and workshops. Key products of this work included three whitepapers and two pilot studies (Demand Research, 2012 and Cadmus, 2012). These two studies used many of the same elements and high-level concepts, but differed in many particulars. The table below summarizes key features of the two modeling approaches.



**Table 5. Features of two California top-down studies**

Model feature	Demand research	Cadmus
Geographic Area	Census tract	Service territory
Time unit	Year	Year
Population measure	Number of sites or premises	Residential: # Housing units Non-residential: Square-footage Utility: Per capita
Program activity measure	Total ex ante savings Total incentive costs Total measure costs	Program expenditures
Normalization	Ex ante savings, incentive costs, and measure costs all per unit of consumption	Consumption and explanatory variables per unit of population
Lag treatment	Cumulative ex ante savings, incentive costs, measure costs, and instruments	Lag program terms
Linear or log-linear	Residential: log consumption vs unlogged predictors Non-residential: log consumption vs mostly logged predictors	Mostly log-log with non-logged expenditures and new construction
Simultaneity treatment	Simultaneous equations with instrumental variables, 2Stage Least Squares	Attempted 2Stage Least Squares
Code impacts treatment	Units built 2000-04	New construction floor space per unit built in each code period
Price terms	Same fuel	Electric and gas
Weather Terms	Heating and Cooling Degree-Days	Heating and Cooling Degree-Days (logged) (service territory population weighted average)
Income	Aggregate Income	Personal Income
Residential Population characteristics	household size College Population in Group housing Median age Median rooms Mobile Homes 3-4-plexes Boat, RV, Vacant housing units	CAC saturation
Fixed effects and trends	Year fixed effects	Utility fixed effect Linear time trend
Number of observations	~30,000	~30-120

A key difference between the two modelling approaches was that the Cadmus study used data at the service territory level, while the Demand Research study used data at the account level aggregated to the census tract level. As a result, the Demand Research study had far more data points available for a similar period, and was able to incorporate a larger number of predictor variables. In addition, the Demand Research study incorporated a more complex model structure.

Each of the studies explored a variety of model fits. The Cadmus study fit models separately for the Residential and Non-residential sectors as well as for the utility as a whole, fit models across the IOUs only and across IOUs and POU's combined, and fit models with and without an autoregressive error structure. The Demand Research Study fit models separately by Residential, Commercial, and Industrial sectors, and also split the Residential sector by IOU, based on initial fits indicating the need for this separation.


Both studies accounted for the effects of building codes by including terms for the proportion of the total stock that was built since a given code went into effect.

Each of the studies provided a single estimate of overall savings with error bounds by combining the results of the preferred models. Results are summarized in the table below. (Reported results from each of the studies have been converted to estimated percent savings per year with relative precision at 90% confidence.)

Both studies indicate savings on the order of 1% per year for the IOU portfolios in total. This is good corroboration given the many differences in specifications and data between the two. On the other hand, the results all have relatively wide error bands. Results at the individual sector level appear to exhibit even greater uncertainty.

**Table 6. Electricity savings estimates from California top-down studies**

Study	Sectors(s)	Timeframe	Effect	Savings % of consumption	savings % per year	Relative precision at 90% confidence
Cadmus	Full utility	2005-10	Cumulative	5.0%	0.80%	55%
Cadmus	Full utility	2006-08	1st year	0.70%	0.70%	108%
Demand Research	Sum of Residential, Commercial, & Industrial	2006-09	Cumulative	5.40%	1.35%	47%
Demand Research	Sum of Residential, Commercial, & Industrial	2006-10	Cumulative	7.30%	1.46%	31%



A recent Massachusetts study also explored top-down impact estimation (DNV GL and NMR) One set of analyses aggregated account-level data to the town and county levels for IOUs only, and the second used data at the service territory level for IOUs and municipal utilities.

The model that used aggregated account level data was limited to the commercial-industrial sector within the Program Administrator utilities. This approach allowed for exploration of separate models impacts at the town and county level. The approach also allowed for separate models of large commercial, small commercial, and industrial sectors, and for separate metrics for upstream and downstream program activity, and for lighting and non-lighting activity. The models also employed both program expenditures and ex ante savings to measure potential program impacts. However, that analysis had only three years of data available and did not produce statistically significant results. The approach was also limited to Massachusetts, and therefore did not include a comparison area to isolate net impacts.

The latter study contrasted energy consumption and programmatic activity within Program Administrator service territories to that of municipal utilities. The municipal utilities served as a comparison area, as the level of program activity was at zero or near zero throughout the time series within those territories. The program activity variable in this approach was limited to program expenditures only, but these models did include 16 years of data. This allowed for a ten-year time series, plus up to a six year in lag programmatic activity to account for the cumulative effects of programs over time. The study estimated separate models for the Residential and commercial-industrial sectors separately, but could not separate commercial and industrial sector models, nor divide commercial models by customer size, industrial classification, or finer level of geography (i.e., town).


The Residential service-territory-level analysis produced savings estimates within 15% of the evaluated net savings estimates with one kind of formulation, and nearly twice as high with another. Some of the 90% confidence intervals included zero and some did not. In subsequent further exploration of these models (DNV GL and NMR, 2015), alternative specifications and data screening choices were found to change the results substantially. This sensitivity is not surprising, but does point to the need to make such choices carefully and transparently, as well as the potential for different modelling perspectives to lead to different conclusions.

#### 4.1.5 Guidance from recent top-down studies

The California and Massachusetts studies all concluded that the top-down methods are promising, but face substantial data hurdles. Each study recommends further research concerning data compilation and model specification. Data compilation is considered a primary limitation of this research at present, as many of the model specification concerns can only be addressed once the necessary data are made available.

To this end, data compilation work began in California that will result in a nine-year time series dataset (2006-2013) for the commercial and industrial (C&I) sectors and a five to six-year residential sector time series dataset (2009 or 2010 through 2013). “The objective ... (is) to produce a database of energy consumption, IOU program participation, weather, demographics, economic activity, and other energy service demand variables that allows time-series, panel analysis of changes in total energy consumption at the census tract-level” (Itron 2015).

Given that this process is already established for California, the present report does not detail potential data challenges and needs in an environment lacking that repository. A similar effort is underway in




Massachusetts, but this effort is focused more on developing the time series going forward, thereby adding to the existing three-year time series with each consecutive year. The Massachusetts report identifies the following criteria for successful top-down models, based on the analysis experience in that state and review of the California and other work.

- **Elements that increase signal.** These elements increase the differences in savings levels between different units, facilitating the detection of savings effects.
  - Diversity of program activity levels: Programs have to vary over time (year over year) and across geography (towns, counties, or states have different offerings). If the program levels are very similar across all the units, the incremental effect of additional program activity will not be detectable.
  - Minimal effect of one area on another (cross-area spillover): Program information and experience from one area influencing behavior in another reduces the apparent program effect. The spillover savings reduces the difference in consumption between the two areas, so that the savings estimate is reduced rather than having a positive increment due to the spillover.
  - Long enough time series to detect and isolate program impacts: Research shows many successful models have more than 10 years of program and consumption data. The Demand Research study used data from a span of only five years. The series length required depends on several factors including the number of geographic areas, (over 6,000 for the Demand Research work) the variability across time and geography, and the treatment of lag effects.
  - Structure that accounts for the persistence of program impacts: Program expenditures in prior years are expected to affect consumption in the current year through equipment survival and spillover.
- **Elements that reduce noise in estimates.** These elements improve the quality of savings estimate by reducing sources of variability.
  - Consistent reporting of data: When data are derived from consistent sources and reported with consistent definitions there is a better chance that variables have the same meaning across units of observation. Without this consistency, estimates may be distorted by differences in meaning of variables that are nominally the same.
  - Consistent relationship between program activity metric and savings: The key estimates derived from the top-down models are those determined from the program activity metric coefficients. If the relationship between program activity and consumption is inconsistent across units of observation, the savings estimate may be dampened or distorted.

Related to these factors are several issues that must be addressed in model specification. Most of these have no one right or wrong approach. The issues and considerations for each are discussed in brief below.

#### 4.1.5.1 Level of aggregation

The level of geographic resolution of the model has implications for the data availability and other cross-cutting factors. Smaller geographic units offer the advantage of more data points to explain the variation in energy consumption. However, smaller geographies diminish the ability to capture spillover—a primary motivation for use of top-down models—and are faced with data availability constraints. Public data sources, such as from the US Census Bureau, obfuscate data to preserve confidentiality. Larger geographies necessarily mean fewer data points, which generally results in less certainty around model estimates. As



noted, the data recently compiled for California are at the census tract level. Some studies have been conducted using census tract, town, or county as the geographic unit, others using service territory or state.

Use of aggregates defined by census units has the advantage that economic variables of interest are available at this level. If the analysis is to be conducted at the service territory level, starting with data at the census tract level makes it possible to aggregate the economic variables to the service territory consistently and with reasonable accuracy. Using geographic data finer than service territory or state provides more units of observation within a given time period, which can improve the model accuracy. At the same time, there are some disadvantages to conducting the analysis at a finer level of aggregation. Two key issues that have plagued bottom-up estimation are of greater concern when finer aggregates are used:

- Cross-area spillover: Spillover from adjacent unit of analysis (i.e., neighboring block-groups or towns).
- Self-selection bias: If program activity tends to be higher in areas and times where customers would have a natural tendency to adopt more (or less) EE on their own, and there is no good metric of this natural tendency that can be incorporated to control for it, the estimated program effect will be overstated (or understated). This is a special case of omitted variable bias. This challenge can be mitigated if the regression includes areas and time periods with and without program availability, or at least with substantially different programs available. To control for self-selection bias, assuming no spillover, we can also develop instrumental variables for the activity level, similar to the use of self-selection correction terms for individual customer regression models.

These constraints obscure the ability to separate the effect of more program activity from the effect of greater interest in participating in programs. By contrast, when the geographic unit is an entire service territory or state, the program activity level across the unit in a given year tends to be driven by earlier policy choices and not by current program interest. Spillover across service territory or state borders is still a concern, but less so than spillover across finer geographic units.

In addition to these structural challenges, there are different data challenges at finer levels than at higher levels of aggregation. In particular, some useful census variables are not available below the county level. In particular, NAICS-code-level employment or domestic product is not available below the county level, to protect confidentiality.

#### **4.1.5.2 Use of logarithmic terms**

Use of logged dependent and independent variables is common in econometric modeling. A primary reason for using the log transformation is to reduce heteroscedasticity, which means that the variance in a variable (consumption) increases (or decreases) as the magnitude of the variable increases (or decreases).

Consumption and many of the factors that drive consumption tends to have a right-skewed distribution, that is, a small proportion of the population with large values compared to the median. Fitting models on a log scale results in fewer extreme residuals and less potential for highly influential observations. In addition, a log vs log fit results in coefficients that are interpreted as elasticities, or percent change in consumption per percent change in the predictor.

On the other hand, some of the predictors don't make a lot of sense in log terms. Using a mix of logged and unlogged terms creates an overall model that is difficult to interpret.

### 4.1.5.3 Weather terms

Weather terms are known to affect consumption, and are typically included in aggregate models in terms of heating and cooling degree-days.

In the Massachusetts pilot study that used aggregated account level data, the model was estimated treating weather normalization two separate ways. Because the model started with account level data that was aggregated to the town or county level, models were first estimated using aggregate weather-normalized annual consumption data that was derived using a monthly degree-day analysis prior to aggregating and fitting the top-down model. This eliminated the need for using heating degree-day and cooling degree-day terms as explanatory variables. This approach takes advantage of the monthly consumption variation to determine weather dependence far more accurately than could be obtained by including only annual degree-day terms in the aggregate model. This normalization approach also avoids the inclusion of log degree-day terms that are difficult to interpret physically.

Upon review of the account level normalization results, the evaluation team found that the degree-day normalization process did not find significant heating or cooling effects for roughly half of the C&I accounts. While this result was not surprising for C&I accounts, the evaluation team also tested a set of models that used non-normalized consumption as a dependent variables and included heating and cooling terms as explanatory variables. These models did not find a statistically significant relationship between consumption and the weather dependent terms, suggesting that the account level normalization models accurately removed weather dependent consumption patterns.

Other practitioners prefer to have all the model terms jointly estimated. An argument for this approach is that consumption is the key variable being explained by the aggregate analysis, and it's less clear what's being measured if consumption itself is adjusted prior to this analysis.

### 4.1.5.4 Treatment of persistence

EE measures have expected useful life on the order of 5, 10, or 20 years, depending on the measure type. As a result, program activity in one year should be contributing to savings in many successive years. That is, consumption in the current year is affected by activity in the current year and in each of the previous several years. One way to treat this effect in the model is to include past (or lagged) activity terms as predictor variable. This was the approach taken in the Cadmus and Massachusetts studies.

A potential problem associated with including the lagged activity terms is that the model may not be able to separate the lagged effects from other factors included in the model if past programmatic activity is strongly correlated to the other factors. For example, past programmatic activity may be highly correlated with past economic activity. As a result, the lagged coefficients may not make sense as a sequence, even if they indicate a reasonable overall savings effect. If the lag terms represent the effect of surviving activity from prior years, with some additional spillover, we expect to see these terms gradually declining over further lag distance. Instead, they are frequently erratic.

An alternative approach is to consolidate the current and prior activity into a cumulative effect. This provides a somewhat less complicated model. The disadvantage is that this single cumulative activity term requires explicit assumptions about how much prior activity persists or extends into the current period, rather than allowing the model itself to reveal the combined effect of equipment survival and spillover. Since in practice the sequence of lag effects is not necessarily well behaved, this may be a reasonable trade-off.

#### 4.1.5.5 Program activity metrics

The metrics used to represent program activity are typically either program expenditures or estimated program savings. The program savings estimate could be *ex ante* or *ex post* based on existing evaluation methods.

A program portfolio has a variety of measure types with varying expected savings per program dollar. As a result, program expenditures are a relatively crude measure of anticipated savings. As noted, the model works best when the impact of an incremental unit of activity is expected to be similar across units of observation. If the units represent very different mixes of measures, or very different savings assumptions for the same measures, this will not be true. Including not just total expenditures, but separate terms for measure costs and incentives, for upstream and downstream programs, and for lighting and non-lighting programs can mitigate this problem.

Program expenditures are a useful metric where savings estimates are not consistently available, or are not considered consistently meaningful, across units of observation. In California, where savings are determined using consistent evaluation protocols, *ex post* savings may be the best available metric. To obtain results before the *ex post* savings become available, *ex ante* values would be used.

#### 4.1.5.6 Consumption normalization

In their whitepaper written as part of the California top-down investigations, Sanstad and Loudermilk, recommend the use of absolute aggregate consumption rather than consumption per unit or energy intensity. However, neither of the subsequent California top-down analyses used this approach, nor did the Massachusetts studies. Normalizing consumption, and the associated predictors, to a per-unit basis reduces overall variability and improves the ability of the model to estimate the effects of interest. Useful normalizing factors include per housing unit for residential, per square foot or employee for commercial-industrial, or per person. If absolute consumption is used, care is needed to address the resulting wide inherent variability (or heteroscedasticity) in the data.

#### 4.1.5.7 Economic factors included


Key terms to include as predictors include fuel prices, a measure of aggregate income or product, weather (perhaps via pre-aggregation normalization per 3.1.6.3), and program activity metrics. Fixed effects terms for each geographic unit and each year can help reduce inter-correlation of residuals. These terms account for each unit's being different across time and each year's being different across geographic units, apart from factors captured in the other explanatory variables. A linear trend term may be useful in place of the time period fixed effects.

To the extent additional population characteristics are available in the data, and the model can support the additional estimates, such characteristics are useful to include. Examples [for the residential sector] are indicated in Table 5 above. When choosing which additional explanatory terms to include, priority should go to terms that:

- Are likely to affect energy consumption
- Tend to have trajectories over time that differ from one geographic region to another

One term that is likely to be important in the models is an indicator of wealth or economic well-being, which varies over time and units and affects customers' inclination to make improvements to their premises, with





or without choosing high-efficiency options. Available variables include residential property values and home sizes, or commercial-industrial GDP by sector or NAICS group. As noted, some of these are harder to obtain at finer levels of aggregation.

#### **4.1.5.8 Model specification assessment**

Model specification is assessed at one level in terms of the statistical precision of the estimates of interest, and their stability under alternative reasonable specifications. Another indicator is whether the estimated coefficients make sense.

Residuals plots provide another a useful diagnostic tool. The residual of the model fit is the difference between each actual observation and its estimated value from the model. Examination of residuals plots is recommended to identify if the model is systematically high or low for certain areas or time periods, or if certain observations have strong influence on the model fit. Useful plots include:

- Residuals vs. year
- Residuals vs. predicted values, separately by year
- Residuals vs. predicted values, separately by area
- Residuals vs. bottom-up savings estimates
- Residuals vs. alternative or omitted variables


Standard statistical estimation packages also provide influence statistics. These statistics indicate if particular points strongly affect the estimated coefficients. If the model results change dramatically when a small number of points are excluded, the model validity is called into question. The points should be explored to see if they are data anomalies or if other factors should be included to account for another effect.

## **4.2 Potential applications to net savings load shape development**

### **4.2.1 Use of top-down methods to estimate net annual energy savings**

As indicated above, top-down methods have been tested in California and elsewhere for annual energy savings estimation at the portfolio level. A data compilation process is in place to provide the data needed for future analysis in the state. Results to date are promising, but not definitive. The results of alternative top-down analyses over a similar time period are similar, but different enough that no one such analysis could be considered authoritative.

In principle the top-down approach captures all elements desired in a net savings estimate, including free ridership, spillover, physical interactive effects, program interactions, and take-back. However, this is not necessarily the case, for reasons discussed above. One limitation is that program activity at the census tract level reflects self-selection, so that the program effect cannot be distinguished from the population's inclination to join the program. This is a key issue that affects free ridership estimation using many bottom-up methods, and potentially leads to overstatement of program-attributable savings. A second limitation is that cross-unit spillover is likely between census tracts, potentially leading to underestimation of total savings. Potentially, these two effects might balance each other. This is an argument that has been made for decades related to other forms of net-to-gross estimation. Reliance on available explanatory variables creates further potential for biases.



Apart from these potential biases, there are limitations to the levels of detail provided by the aggregate analysis. While it may be possible for top-down approaches to account for free-ridership, spillover, and take-back, the approach does not provide separate estimates for these factors. The method also does not provide insights into the relative contributions to the total savings of different program types, unless they are separate terms in the model. Similarly, the aggregate model determines incremental effects of incremental program activity overall across the time and geographies included in the analysis. The method will not identify that programs have been more effective in certain areas or time periods.

The ability of the top-down model to capture spillover effects is constrained by the period of the analysis. Program-attributable installations from years before the first year of program data in the analysis implicitly are included in the baseline, and spillover from program activity in the pre-analysis period implicitly is part of an underlying trend in naturally occurring savings. At the other end, spillover that occurs beyond the last year of consumption data in the analysis does not contribute to the estimated savings coefficients.

#### 4.2.2 Use of top-down methods to estimate net hourly energy savings

Extending top-down methods to produce load impact shapes is unlikely to be fruitful until the annual estimation methods are more solid. Assuming the annual estimate effort proceeds, in principle, the same or similar methods basis could be used to provide hourly savings. However, several additional specification questions would need to be addressed in such an extension.

A challenge with moving to hourly analysis is that most of the predictor variables are available only on an annual, or in some cases quarterly, basis. It doesn't make sense to fit an hourly time series with predictors that are static for a year at a time. While it would be possible to interpolate the annual variables, this would likely lead to substantial inaccuracies. It's tempting to think of modeling each hour of the year separately, but the same date and hour are different depending on the calendar.

Given the temporal structure of the predictors, a more useful approach might be to reduce each year's data to a set of load shapes or load shape parameters. For example, we might generate 36 day-type load shapes (weekday, weekend, and peak day for each month) and fit a separate top-down analysis for each hour of each day type, using a large seemingly unrelated variables regression. Alternatively, we might construct calendar-and-weather-normalized load shapes for summer, winter, and shoulder seasons, and work with those.

Apart from the model specification issues, additional data issues arise moving to an hourly analysis. The data compilation currently developed for California does not include hourly data, only the monthly billing parameters. Data that are not used routinely tend to have a much higher rate of missing and bad data compared to the consumption data used for billing. Developing the data to support an hourly top-down analysis would be a substantial additional effort.

## 5 LEVERAGING AMI DATA FOR PROGRAM EVALUATION

The availability of AMI data provides as much motivation for this whitepaper as the promise of randomized assignment methods. The dream of being able to produce an 8,760 load shape that captures savings across all hours of the year appears tantalizingly close. After all, with AMI, the data is available for every household.

The previous two sections have already discussed why the savings load shape is not a simple prospect. Randomized assignment designs offer the most rigorous evaluation results but would be difficult if not impossible to apply across the range of programs. Top-down modeling in effect provides an analytic estimate of the counterfactual as a basis for net savings estimation, but is subject to model specification uncertainties and data limitations. Furthermore, top-down modeling is not ideally suited to taking full advantage of the riches of hourly data. This section gives a high level overview of the role AMI is playing in evaluation and energy program implementation today.

### 5.1 Background


As described in the Introduction, one motivation of this whitepaper was an interest in exploring uses of AMI data as the basis for program or portfolio evaluation. Section 3.2.2 addresses how top-down analysis could be extended to provide hourly impacts, using AMI data.

Another approach to using AMI data for evaluation is to apply AMI analysis for individual program impact estimation and load shape profiling. AMI data is already in routine use for evaluation of demand response programs. AMI data provides a natural enhancement to energy program impact evaluation based on energy consumption data analysis. Where such analysis in the past would be conducted using monthly data, daily data from AMI can provide more accurate energy savings estimates (Allcott 2014), even if hourly impacts are not a consideration. California can avail itself of these benefits as it ranks number one nationally for the number of installed AMI meters (12,400,000) and fifth for overall AMI penetration rate at approximately 85%.

Monthly consumption analysis is generally recognized as an appropriate evaluation method for programs with relatively homogeneous participant groups and with savings a relatively large fraction of pre-installation load. Monthly data tends to dampen data “noise.” Data across customers at a given hour, particularly the peak hour, in general will be more variable than monthly consumption, potentially making it more difficult to detect savings in a peak-hour analysis.

When the goal is energy savings estimation, daily data, and the incorporation of supplemental premise information from existing utility or third party databases, has the potential to extend the applicability of using AMI data to detect smaller proportions of savings. AMI data also presents the possibility of determining peak impacts. However, conducting evaluation, measurement and verification (EM&V) with AMI data for whole premise consumption still requires weather normalization, variance analysis, a counterfactual, and close to a year of pre- and post-implementation data to capture all seasonal relationships.

Statistical analysis of more granular data may enhance the ability to separate heating, cooling and base-loads with increased precision, using the same degree-day model structure with daily data as has been commonly used for monthly data. In addition, daily data allow additional variables to be included in the models such as day types (weekends, holidays), peak hours, and consecutive hot or cold days. The



expanded model specifications enhance the ability to analyze and predict shell performance. Combining daily meter and temperature data shows the relationship between degree-days and HVAC loads when they occur, rather than results that are averaged over 30 days. Daily detail also reveals insight into differences in building performance in mild versus extreme conditions.

Analysis of HER programs has used daily consumption data analysis over multiple years to study the persistence of program savings (DNV KEMA, KEMA). Similar analysis can be applied for other programs amenable to consumption data analysis, with daily data extending the circumstances in which such analysis can be useful. Outside the RCT context, the validity of comparison groups over extended time may become a challenge for persistence studies. Moreover, such analysis does not directly address whether measures themselves survive, or whether attenuation of measured savings reflects other behavioral changes by both participants and the comparison group.

For analysis of hourly impacts, load research practitioners have established techniques that expand degree-day models in additional directions. One approach is to fit separate models for each hour of the day. Such models may incorporate calendar information, and lagged temperature effects. Another approach is first to model daily energy use, then to use a second kind of model to estimate shape (distributing total daily energy over hours) or peak load only.

AMI data is not without drawbacks, notably, the volume. AMI data could possibly be the single largest volume of data collected and managed by utilities. For example, a small study using two years of monthly data, analyzing six variables for 10,000 premises, produces 1.5 million data points. The same analysis using hourly data produces 1 billion data points. Due to the volume of data, many analyses require cloud based computer platforms and off-site data storage. In the face of sometimes-difficult data collection processes, collecting and preparing AMI data is laborious for both the utility staff and EM&V practitioners. Data validation will require automation, and manual reconciliation will be limited to extreme anomalies (McMenamin).

A key to making consistent use of hourly AMI data for evaluation is establishing compilation, cleaning, and screening protocols. Utilities are using these data primarily to obtain billing parameters. Data that are not of routine operational use are unlikely to be cleaned and stored consistently. More work needs to be done to determine best practices for using daily and hourly AMI data for evaluation, including dealing with data volume and cleaning challenges.

### 5.1.1 Automated measurement and verification

For many years, the private sector has embraced automated measurement and verification to measure impacts of EE measures and equipment, to support building commissioning, and as a maintenance management strategy for monitoring building operations. The application of automated M&V to utility and government EE programs is relatively new and automated M&V is still considered an emerging EM&V tool. Historically, the majority of applications have been adopted in the private commercial sector, but both commercial and residential applications are currently being deployed in the utility sector to support program delivery and customer engagement.

Using metered data as its primary input, automated M&V methods are similar to traditional billing analyses as practiced by evaluators for many years, with similar accuracy. Automated M&V software measures energy savings by creating pre-program model specifications of energy consumption from metered data (the model training period) and applies the predicted model to post-program conditions. The automated M&V vendors report energy savings by subtracting metered consumption from predicted consumption in the post period.

As has been recognized for years in the evaluation context, consumption data analysis for savings estimation is most reliable when it uses close to 12 months of consumption data for the pre-implementation (training) period and for the post-implementation (savings estimation) period. Six to nine months may suffice if all seasons are included, but accuracy will be lower than a 12 month analysis period. These guidelines have been found in the context of automated analysis (Granderson et al.) just as in earlier consumption analysis (Fels).

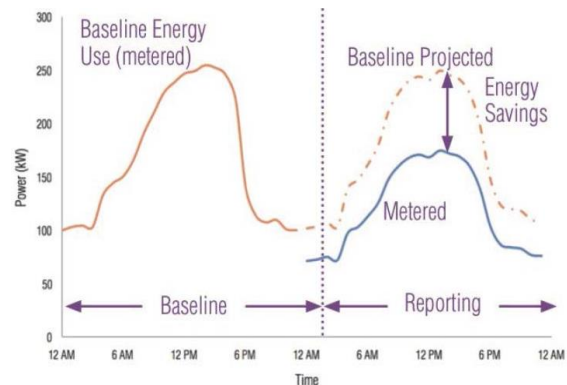
Depending on the application, the automated models sometimes incorporate occupancy, operations schedules, product throughput and third party data to bring in information about premise characteristics to support the analysis, particularly for assessing energy intensities for program targeting. Automated M&V takes advantage of both machine learning and cloud-based platforms to perform analysis on a continuous basis in near real-time. In principle, automated M&V software provides rapid and ongoing analysis of consumption data.


Automated M&V supports program delivery by providing whole building feedback at the site, project, and program level to program administrators and participants early and throughout the post-implementation period. Ongoing automated analysis could be useful in an evaluation context also. Evaluation applications of automated M&V would require agreement on the automated analysis and acceptance of the data screening process and data loss. Rapid and ongoing feedback to the program using automated M&V would be a useful evaluation function if not already provided by the program delivery.

Automated methods could reduce the turn-around time of the delivery of evaluated results if there is up front agreement on the use of the methods, with no extended exploration of the results afterward. The potential benefits of automated analyses are that emerging results could be tracked much earlier than the end of the 12th month. If savings are less than expected in successive seasons, reasons could be explored early, and possible changes to the program or to the analytic method could be considered.

Automated M&V is sometimes employed by third party evaluators for internal research and as a screening tool, but is typically offered to utilities in a software-as-service (SaaS) model to identify high potential savers, as a customer engagement tool, and to measure gross savings to provide feedback to program administrators. Most of the automated M&V methodology is proprietary but some vendors have indicated a

**Model training period (left), baseline projected model and metered consumption (right) (LBNL)**





willingness to release model specifications on an as needed basis. To date it has been deployed in limited pilots, at the site and program level, primarily with monthly data due to the availability of AMI data.

### 5.1.2 Current research on the performance of automated M&V methods

Lawrence Berkeley National Laboratory (LBNL) compared how well data-driven automated baseline models predicted commercial energy consumption from high interval metered data. Model specifications were developed from three, six, nine, and twelve months of consumption data (the training period). Predictions from each model were compared to metered data in the post period, and each other, using standard statistical tests.

In LBNL's research on automated models, performance across models was similar. The most recent study, which included proprietary and public models, found that for most buildings in a large data set (500+ buildings), the models are likely to meet ASHRAE guidelines for how well baseline models should fit the pre-measure data. In addition, the study found that for half of the buildings, annual energy use could be predicted with errors less than 5%.

The accuracies achieved in this study were for a fully automated case. In practice, errors can be further reduced with the oversight of an engineer to conduct non-routine adjustments where necessary. Errors are also further reduced when individual buildings are grouped and considered at the program level. However, these additional expert processing steps add to the total time for results to be delivered in an evaluation context (Granderson et al.).

These modeling and analytic approaches have the potential to reduce the time and cost associated with EM&V, and the calculation of savings relative to a baseline defined by the prior equipment in place. Savings relative to "standard" new equipment would not automatically be provided without additional agreements, assumptions, and embedded analytics. These methods also do not address other aspects of evaluation such as program attribution effects, including free ridership, spillover, and market effects.


Future LBNL research will focus on demonstrating the automated M&V approach in partnership with utilities and program implementers. Gross savings from sets of program data will be calculated using interval data baseline models, along with the associated uncertainty and confidence in the savings results. Where possible, the time savings with respect to more conventional EM&V methods will also be investigated.

## 5.2 Potential applications to net savings load shape development

### 5.2.1 Assessing automated M&V for program evaluation

For use of whole-premise automated analytics in program evaluation, a first step would be to validate the predictive accuracy of the tools using a protocol similar to LBNL. Ideally, such validation would be conducted within multiple regions, and over multiple years. The protocol should also address data screening rules and transparency of these rules. Once the predictive accuracy is established, evaluation should be based on a 12-month training period and 12 month post-implementation period.

Key to using these results for evaluation is the need to establish a valid comparison group. Defining a valid comparison group for this application has all the same challenges as with conventional billing analysis. (SEEAAction, Agnew, DNV GL 2014). The comparison group must be defined such that it can reasonably be assumed to represent the level of natural EE adoption among the participants absent the program.



Alternatively, regression analysis is used to control for the underlying non-program differences between participants and the comparison group. This type of analysis would be extremely challenging to automate.

To make the establishment of a comparison group as automated as other aspects of the analysis, it would be necessary to agree in advance on a protocol, and a basis for testing its validity. The application of this protocol to creating the comparison group should be overseen by an independent third-party evaluator.

As noted, even if a valid comparison group is available, pre-post analysis in general provides program impacts only in the context of retrofit measures with homogeneous participant groups. For other contexts, the post-only automated analysis could be used as part of a calibration process. While this work could, in principle, begin earlier than 12 months after installation, there may be limited value to the preliminary savings estimates beyond the benefits already provided by the automated analysis itself.

### 5.2.2 Program evaluation with net load shapes using AMI data

As indicated, most applications of AMI load data to energy program evaluation to date have been using consumption data analysis with daily data. Load impact analysis for DR programs commonly models “event” versus non-event days as well as participants versus a comparison group. Some of the basic modeling approaches used for load data analysis can be applied to estimate net savings at the hourly level, in contexts where a comparison group is well defined and meaningful.

Automated M&V at this stage is mostly not using hourly data or addressing hourly impacts. These are natural extensions, with many details yet to be worked out.

Data quality remains an issue for any applications of AMI data to evaluation, whether for annual energy or hourly impacts. Data attrition is an issue for any consumption data analysis, but potentially is worse with AMI data. Work is needed to establish whether there are any patterns to the prevalence of missing and anomalous data for a particular service territory, and what biases these might lead to in an evaluation.



## 6 SUMMARY AND CONCLUSIONS

Part of the impetus for this whitepaper was clearly motivated by the relatively recent availability of AMI data. The availability, in theory at least, of full 8760 load shapes for all customers has already made some things possible that would never have been possible ten years ago. It would not have been feasible to measure the demand reduction effects of a behavior program. Those small reductions can be estimated only because we collect hourly data from hundreds of thousands of residential households taking part in an RCT experimental design. In fact, the potential for estimating unbiased and highly precise peak load effects of a behavior program is a key drive of this whitepaper. Why, if we can do that, can we not expand the framework to cover all programs?

This whitepaper has addressed this question in three ways:

1. Exploration of the potential for applying RCT methods at the portfolio level, with hourly data
2. Exploration of quasi-experimental approximations to RCT at the portfolio level, in particular “top-down” analysis, including hourly top-down analysis
3. Exploration of other uses of AMI data for evaluation

### 6.1 Applying RCT methods at the portfolio level

With a discussion of a parallel universe counterfactual scenario, we outlined how one could in principle produce a residential portfolio net savings load shape. This approach would require AMI data but would also require access to a parallel universe where everything was identical except that programs did not exist.


While not particularly practical, this approach is informative of some of the central challenges of such an endeavor. In particular, we illustrated the relationship between self-selection and free ridership and the challenges of directly addressing those effects. Program participation, in general, is an act of self-selection. Natural EE adopters, if they exist, are a subset of those who self-select into the program. Any method that claims to address free-ridership must show how the method constructs a proxy counterfactual that includes those natural EE adopters.

We also linked this counterfactual scenario to the RCT and RED experimental designs that have become more common in the energy space in the last decade. This linkage showed that the practical randomized assignment designs derive their power from the approximation of the ideal counterfactual. This approximation is statistically unbiased with quantifiable uncertainty. As a result, the control group created by random assignment yields an estimate that is rigorous, unambiguous, and reliable, potentially even for relatively small impacts.

We also showed that central aspects of a randomized assignment design make literal application of this approach impractical at the level of a full residential portfolio net savings load shape. In particular, it is not possible to operate a full portfolio of programs using a range of delivery and outreach methods while comprehensively denying treatment to a randomly selected control group,

### 6.2 Top-down analysis and other quasi-experimental methods

When randomized assignment is not practical, quasi-experimental methods provide an alternative approach. These methods construct an approximate counterfactual using a combination of explicit and implicit



comparison group selection, including various analytic techniques to attempt to control for non-program effects.

Top-Down analysis models aggregate consumption across geographic areas and time as a function of portfolio activity along with demographic and economic factors. This modeling is designed to isolate the effects of the EE portfolio while controlling for other factors. This quasi-experimental approach has particular appeal for portfolio-level analysis because in principle it captures comprehensive portfolio effects, including spillover, market effects, cross-program interactions, free ridership, and take-back. The approach has thus far been applied only at the level of annual energy savings, but could in principle be extended to an hourly framework.

California has established a database that can be used for developing top-down models for portfolio-level impact analysis. This database addresses one of the common impediments to development of such models.

There remain several limitations to the top-down methods. One is that any such analysis will retain the potential for method bias. Even with a well-established database, the only variables that can be used in the model are those that are available from the compiled sources, and at the level of geographic and temporal detail provided by those sources. As a result some level of model mis-specification bias is inevitable, as with virtually any modeling exercise. In addition to the data limitations, self-selection bias is also an issue.


A second set of limitations has to do with the level of detail the model can provide. A top-down estimate cannot provide separate estimates of the various net-to-gross components, such as free ridership and spillover. The estimate also provides an overall net savings per unit, and does not identify how a particular program-year or area was more or less effective.

To extend the top-down methods to address hourly data would require a number of technical issues to be addressed. Most of the predictor variables are available only at the annual or at best monthly level. The model structure must take into account variation by time of day and calendar, while appropriately reflecting lag effects. Work would be required to develop and test some approaches. Availability of clean hourly data for this work also remains a challenge. Hourly AMI data are collected but not comprehensively cleaned and compiled.

For all these reasons, top-down analysis cannot provide the definitive all-in, portfolio-level net savings estimate, at either the annual or hourly level. Nonetheless this approach offers an important level of verification of portfolio-level accomplishment. If programs are operating at a non-trivial level, it should be possible to discern their effects in aggregate. The top-down analysis provides a high-level confirmation that the program portfolio effects are real.

### **6.3 Using AMI data in evaluation**

AMI data is already being used for a wide range of measurement and verification purposes. This extensive new data source offers a number of strengths and challenges for these applications. AMI offers easy access to hourly data for full program populations. Among other things, this improves the evaluators' ability to characterize consumption with respect to weather and appropriately assign savings effects on peak demand. Automated M&V, in particular, has the potential to provide almost real-time feedback on changes with respect to modeled consumption from the pre-program period. Methods are in development that will provide standardized and benchmarked results within a pre-determined range of error. While these results do not



constitute a full evaluation, as they do account for exogenous effects that may be conflated with program effects, the improved timeliness and scrutiny may support more comprehensive consideration of these confounding effects.


For the purpose of portfolio-level net impacts, AMI plays a more suggestive role. In the context of top-down modeling, AMI may provide better data for top-down modeling. Though AMI data already has greater granularity than other key data sources, these AMI data may support top-down modeling in other indirect ways such as producing calendar-and-weather-normalized load shapes to allow modelling at the hourly day type level. The search for the useful applications of AMI data across all aspects of energy analysis has just begun.

## **6.4 Where to next**

Top-down modelling is a feasible approach for developing some form of portfolio-level aggregate analysis. Furthermore, it is reasonable to expect that approaches will be developed using AMI data that will give the analysis an hourly application. These results are unlikely to have the rigor of an RCT analysis but they will provide valuable corroborative evidence of the overall impacts of the EE program portfolio. The larger goal of portfolio-level net savings load shape, however, provides a useful framework within which to understand not just the other portfolio-level options but the full range of evaluation approaches. The goal of all energy program evaluation is to produce valid results with the kind of accuracy and precision that is already possible with randomized assignment design. A top-down approach that provides corroboration for the combined claims of the many program specific evaluations would further enhance the savings claims of all EE programs.

## 7 REFERENCES

- Agnew, K & Goldberg, M. 2013. *Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol*. The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures (UMP).
- Allcott, H., & Rogers, T. 2014. *The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation*. American Economic Review, American Economic Review.
- American Society of Heating, Refrigerating and Air-Conditioning Engineers. 2014. ASHRAE GUIDELINE 14:2014, Measurement of energy, demand, and water savings. Atlanta, GA: American Society of Heating, Refrigerating, and Air-Conditioning Engineers.
- CADMUS Group. 2012. *CPUC Macro Consumption Metric Pilot Study - Final Report*. Prepared for the California Public Utilities Commission.
- The California Evaluation Framework. 2004. Prepared for the California Public Utilities Commission and the Project Advisory Group. Last Revision: January 24, 2006.  
<http://www.cpuc.ca.gov/PUC/energy/Energy+Efficiency/EM+and+V/>
- Demand Research, LLC. 2012. *Macro Consumption Metrics Pilot Study Technical Memorandum - Preliminary Findings*. Prepared for the California Public Utilities Commission.
- DNV GL and NMR. 2015. *Top-Down Modeling Methods Study - Final Report*. Prepared for the Massachusetts Electric and Gas Program Administrators.
- DNV GL. 2014. *Evaluating Opt-In Behavior Programs: Issues, Challenges, and Recommendations* California Public Utilities Commission – Energy Division Report No.: CPU0088.01, Rev. Version 01.
- DNV GL and NMR. 2015. *Top-down Modeling Methods Study—Final Report*. Prepared for the Massachusetts Electric and Gas Program Administrators.
- DNV KEMA Energy & Sustainability. 2012. *Puget Sound Energy's Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation*. Prepared for Puget Sound Energy.
- Fels, M.F. ed. 1986. *Energy and Buildings (Special Issue Devoted to Measuring Energy Savings: The Scorekeeping Approach)*, 9, no.1&2.
- Fowle, M, Greenstone, M and Wolfram, C. 2015. *Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program*. E2e Working Paper 020.
- Granderson, J, Price, PN, Jump, D, Addy, N, Sohn, M. 2015. *Automated measurement and verification: Performance of public domain whole-building electric baseline models*. Applied Energy 144: 106-113; Granderson, J, Touzani, S, Custodio, C, Fernandes, S, Sohn, M, Jump, D. 2015. *Assessment of automated measurement and verification (M&V) methods*. Lawrence Berkeley National Laboratory, LBNL#-187225.
- Itron, Inc. 2015. *2010-2012 WO081: Macro Consumption Metrics Data Development*. Prepared for the California Public Utilities Commission.
- KEMA, Inc. 2013. *Puget Sound Energy's Home Energy Reports Impact Evaluation*. Prepared for Puget Sound Energy.

- 
- Lawrence Berkeley National Laboratory et al. 2013. Quantifying the Impacts of Time-Based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines. Prepared by Peter Cappers, Annika Todd, Michael Perry, Bernie Neenan, and Richard Boisvert Prepared for the U.S. Department of Energy.
- Sanstad, A. and Loudermilk, M. 2011. *Estimating Aggregate Energy Consumption Effects of Energy-Efficiency Programs in California: Review, Proposed Approach, and Pilot Study*. Prepared for Itron, Inc. and KEMA, for the California Public Utilities Commission.
- Schellenberg, J. et al. 2015. *Comparison of Methods for Estimating Energy Savings from Home Energy Reports*. Prepared for Pacific Gas and Electric Company.
- State and Local Energy Efficiency Action Network. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations*. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. <http://behavioranalytics.lbl.gov>.
- Stewart, J., and A. Todd. 2015. *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures, Chapter 17: Residential Behavior Protocol*. Prepared for the National Renewable Energy Laboratory (NREL).
- TecMarket Works et al. 2004 (latest revision January 24, 2006). *The California Evaluation Framework*. Prepared for the California Public Utilities Commission and the Project Advisory Group.