Boston | Headquarters

617 492 1400 tel
617 497 7944 fax
800 966 1254 toll free

1000 Winter St
Waltham, MA 02451

Observations on Chapter 8 of the Uniform Methods Project:

A Discussion of Comparison Groups for Net and Gross Impacts

Pacific Gas and Electric Company

CalMAC Study ID: PGE0407.01
Date: July 12, 2017

Contributors

Katherine Randazzo, Ph.D.
Principal Data Scientist, Opinion Dynamics

Richard Ridge, Ph.D.
Ridge & Associates

Stef Wayland
Director, Opinion Dynamics

Table of Contents

1. Context and Overview

The Uniform Methods Project (UMP) was developed by the US Department of Energy (DOE) with the goal of strengthening the credibility of energy efficiency (EE) programs by improving the consistency and transparency of how gross and net energy savings are determined. Current EE EM&V practices in the United States use multiple methods for calculating energy savings. These methods were initially developed to meet the needs of individual EE program administrators and regulators. While the methods served their original objectives well, they have resulted in differing and non-comparable savings results—even for identical measures. These differences can be significant according to a study published by the State and Local Energy Efficiency Action Network.[1] This increased credibility should give electric utilities, their regulators, and other stakeholders a greater level of confidence about reported savings and reduce the risks of using EE as an electricity resource. Each chapter of the UMP was written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. DOE considers the UMP to be a living document and the website encourages feedback on any of the chapters. Here, our focus is on Chapter 8, Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol. Chapter 8 has become particularly relevant in California in light of the passage of AB802 and IOU creation of high opportunity projects and programs (HOPPS). The methods in Chapter 8 are currently being used as part of a pilot project designed to estimate gross savings for Pacific Gas and Electric Company's (PG&E's) HOPPS programs using advanced metering data.

The "Application Conditions of Protocol" section of Chapter 8 sets the limits of how the protocols that follow should be applied. The authors describe the proper situation in which to apply the protocols as programs that promote multiple measures as *retrofits (i.e., early replacement)*, and where the baseline for estimating gross program savings with consumption data is defined as the pre-program consumption. Early replacement obviates the need to make any adjustments to baselines for efficiency codes. In other words, the measures should not be those that are replaced on burned-out. Four approaches are proposed for estimating gross savings and two approaches are proposed for estimating net savings. Table 1 from Chapter 8 is presented below.

With the exception of Row 1, randomized control treatment (RCT) design, in Table 1, Chapter 8 is devoted to estimating gross savings using various statistical techniques and research designs. In Row 1, the authors begin with the randomized control trial (RCT) design that provides an unbiased estimate of net but note that such a design is rarely feasible. The remaining rows are all designed to estimate gross savings, with rows 2, 4, and 5 using different types of comparison groups to control for exogenous changes.

[1] https://www4.eere.energy.gov/seeaction/system/files/documents/emvscoping_databasefeasibility.pdf

Table 1: Program Characteristics, Comparison Group Specifications, and Consumption Data Analysis Structure and Interpretation

Program Condition	Consumption Data Analysis Form	Comparison Group	Gross or Net Savings	Unknown Biases
1. Randomized controlled trial experimental design	Two-stage or pooled	Randomly selected control group	Net	Spillover, if it exists
2. Stable program and target population over multiple years	Two-stage	Prior and future participants	Gross	Minimal
3. Participation staggered over at least one full year	Pooled	None: pooled specification with participants only	Gross	Minimal
4. Not randomized, not stable over multiple years, participants similar to general eligible population, nonparticipant spillover minimal	Two-stage or pooled	Matched comparison group	Likely between gross and net	Self-selection[6] and spillover
5. Not randomized, not stable over multiple years, participants unlike general eligible population, nonparticipant spillover minimal	Two-stage or pooled	General eligible nonparticipants	Likely between gross and net	Self-selection and spillover

The target audience for this paper is the evaluation practitioner who is evaluating whole building programs and is familiar with the UMP Chapter 8, as well as regulators who supervise and judge the studies that are conducted for whole-building programs that are designed to achieve deep savings. In particular, the audience would include those who wrote and reviewed chapters for the UMP and especially Chapter 8. We aim to raise issues for the evaluation community to consider, with the idea of broadening the possibilities for good evaluation designs specific to this type of program.

This whitepaper has four objectives:

1. *Define key terms and to suggest clarifying the meaning of key terms* such as “eligible population,” “counterfactual” and “comparison group.”

2. *Provide suggestions for further elaboration and detail in the guidance offered in the use and composition of comparison groups*.

3. *Anchor the discussion of research designs in the traditional research design literature*.

4. *Suggest other possible ways of estimating net saving*s in addition to the two outlined in Chapter 8.

Making clear and supportable recommendations for a revision of Chapter 8 requires that we come to a common understanding of a number of social science concepts as they apply to the EE field. In pursuit of that common understanding, we devote considerable space in Section 2 to laying out our understanding of those concepts.

2. Review of Pertinent Fundamental Concepts

While there is considerable consensus in the EE industry about the definitions of gross and net savings, there is less clarity about comparison groups and their composition, and how their composition affects the estimated savings. Specifically, it is not always clear what types of comparison groups might help us generate a good estimate of gross savings, and which will generate a good estimate of net savings. In this paper, we will provide the common definitions of gross and net savings, and then consider, in some detail, what types of comparison groups provide the basis for both gross and net savings, paying particular attention to their use in evaluating whole house programs. This paper will focus on the observable characteristics of comparison groups. A companion white paper (Train et al., 2017) deals with a related issue: addressing the endogeneity when estimating net savings. Endogeneity stems from both the observed and unobserved variables within models predicting consumption or change in consumption due to program participation. A primary source of endogeneity arises here because participation, and variables associated with it, can be both cause and consequence of consumption patterns. This affects the use of comparison groups as well.

Following are some definitions and descriptions of some fundamental terms used in our industry.

2.1 Gross Savings

Gross savings is defined as:

Changes in energy consumption that result directly from program-related actions taken by participants of an EE program, regardless of why they participated (Violette and Rathbun, 2014).[2]

Note that this definition only addresses the effect of the program-related actions (i.e., the installation of efficient measures or the adoption of efficient behaviors) on energy consumption. It does not address the issue of how many program-related actions would have occurred in the absence of the program, which is the focus of the next definition.

2.2 Net Savings

Net savings is defined as:

Changes in energy use that are attributable to a particular EE program. These changes may implicitly or explicitly include the effects of free ridership, spillover, and induced market effects (Violette and Rathbun, 2014).

In other words, with “attributable to the utility DSM program” we want to isolate the savings that are *caused* by the program from those that would have occurred naturally, i.e. in the absence of the program. What would have occurred naturally is the counterfactual, which we discuss next.

[2] These are the definitions provided by works and authors the industry considers authoritative. However, others note that using the word “change” introduces a perspective that implies a pre-post comparison, which isn’t necessary to the concept or its measurement. Rather, the more general language would refer to the “difference” between what happened versus what would have happened absent the installation of the efficient measure.

2.3 Counterfactual

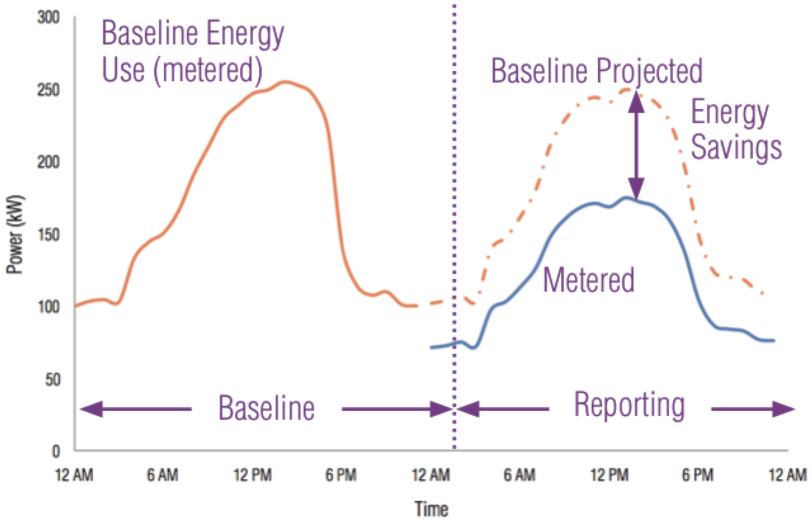
Social science methods texts such as Shadish, Cook and Campbell, 2002 provide a good description of the counterfactual:

> In an experiment, we observe what did happen when people received a treatment. The counterfactual is knowledge of what would have happened to those same people if they simultaneously had not received the treatment. An effect is the difference between what did happen and what would have happened. (p. 5)

While the definitions of the terms discussed in this section are relatively straightforward, defining the nature and purpose of comparison groups is much more complex. In a true experimental design, where sample units (people, businesses, etc.) are assigned randomly to treatment and control conditions, determining relevant behaviors (including post-program usage) for each group, and subtracting the level of behaviors of the control group from the treatment group provides an estimate of net impacts. In this situation, the control group acts as the counterfactual. Under some circumstances, discussed later, a quasi-experimental design that uses a non-randomly-assigned comparison group can provide the counterfactual such that it provides the basis for estimating program net effects. Another way that evaluators establish a counterfactual is by asking participants directly what they would have done if there had been no program. We might term this a hypothetical counterfactual (Ridge et al., 2009; 2010). Yet another approach the revealed-preference discrete choice model (Train, 1993; Goldberg and Train, 1995).

Next, we discuss how the words we choose to describe our evaluation designs, can contribute to the confusion. Some researchers have used forecasts of energy use that they refer to as a counterfactual baseline for estimating savings. Figure 1 illustrates one such example[3].

Figure 1. Example of the Use of a Forecast of Usage Based on Metered Pre-Installation Consumption



[3] Taken from a presentation by Jessica Granderson (2015) entitled, "Accuracy of Existing Use Baselines, AKA Normalized Metered Energy Use." It uses historical metered consumption as the basis for a forecasted baseline.

However, in the presentation from which Figure 1 is taken, the "baseline projected*",* a perfectly reasonable term*,* is nevertheless referred to as the *counterfactual* baseline in the presentation and is used to estimate *gross* savings. Others have referred to the *baseline projected* used to estimate gross savings as representing what would have happened in the absence of the program, words that are often used to describe the counterfactual in the traditional research design literature. Our view, and that of the larger social science community, is that the term "counterfactual" and the words used to describe it should only be used in connection with estimating a *program's net saving*s (Mohr, 1995; Pearl, 2000), i.e., the causal relationship between an intervention and the observed outcome. There can be other types of baselines that support estimates of gross impacts, but the counterfactual is, by definition, the point of comparison for estimating net program impacts, since it is meant to represent what would have happened without the program. That is how we use the term throughout this paper. As an aside, we note that there is some disagreement as to whether the difference between the *baseline-projected* energy use and the *metered* energy use is gross or net or somewhere in between (Malinick and Ridge, 2015).

2.4 Comparison Groups

One way to describe the requirements of a good comparison group in the traditional literature is provided by Rubin (1974). He introduces two relevant concepts: Stable Unit Value Treatment Assignment (SUTVA), and Ignorable Treatment Assignment (ITA). The principle of SUTVA is described:

> The outcomes of any unit are not affected by the treatment assignment of any other units. Example of a violation: non-participant spillover, where comparison group members may learn about energy saving behaviors by talking to their treatment group neighbors.

Violations of SUTVA will most frequently be illustrated by non-participant spillover, or free drivers in the EE field.

The principle of ITA is described:

> For every unit, it must be possible that that unit could have been assigned to either treatment or comparison group. Further, that treatment assignment is independent of the outcome, given the covariates. This is sometimes called 'unconfounded' or 'no hidden bias.'

Random assignment of customers to a treatment or control group would accomplish ITA, but accomplishing it in the absence of random assignment is challenging indeed. It implies a *comparison* group with matching to participants on essentially all variables relevant to the outcome variable of consumption or change in consumption, observable or not.

Standard research design texts (e.g. Campbell and Stanley, 1963; Shadish, W. R., T. D. Cook, and D. T. Campbell, (2002)) point to multiple potential functions that must be performed in quasi-experimental designs, i.e. using comparison groups where control groups are not feasible. Multiple factors could affect the outcome variable, and therefore will confound the effects of the treatment if not addressed adequately. They point to the need to control for such influences as

1. **History**: Events outside of the study/experiment or between repeated measures of the dependent variable may affect participants' responses to experimental procedures. In the EE field, history includes factors that change over time, such as weather and social/economic conditions.

2. **Selection**: This refers to the problem that, prior to participation, differences between groups may exist that may interact with the treatment variable and thus be "responsible" for the observed outcome.

Selection biases can occur due to program targeting, or due to customers self-selecting into the program. In the EE field, this can include many factors, including anything that has an impact on energy use and that differs between treatment and comparison groups. This certainly includes, e.g., the types of attitudes and motivations that are associated with self-selection into an EE program.

3. **Maturation**: Subjects change during the course of the experiment or even between measurements. For example, young children might mature and their ability to concentrate may change as they grow up, or, a person's attitude toward global warming might change slowly over time making them more predisposed to reducing their energy use. Changes in the needs of a household over time could also be categorized as Maturation

4. **Statistical Regression to the Mean**: This type of error occurs when some subjects' have extreme scores (one far away from the mean) such as high energy use. For example, when customers whose annual energy is greater than 12,000 kWh are targeted for an energy audit, reductions in energy use after participation will be at least partially due to regression toward the mean and not the program's effectiveness. On the other hand, if the extreme scores are equally distributed between extremely high and extremely low, there will not be a biasing effect.

5. **Testing**: Repeatedly measuring participants may lead to bias. Participants may remember the correct answers or may be conditioned to know that they are being tested. Repeatedly taking (the same or similar) intelligence tests usually leads to score gains, but instead of concluding that the underlying skills have changed for good, this threat to Internal Validity provides good rival hypotheses. This is unlikely to be a factor in most EE programs as the participants are generally not conscious of the data gathered by evaluators to show their responses to the intervention. One exception to this can occur in our industry when surveys are used at multiple times during the evaluation.

6. **Instrumentation**: The instrument used during the testing process can change the experiment. This also refers to observers being more concentrated or primed. If any instrumentation changes occur, the internal validity of the main conclusion is affected, as alternative explanations for apparent gains are readily available. This factor, as well, is not usually an important factor in determining internal validity for evaluations of EE programs, although this statement is subject to the same exception as noted with Testing.

These influences are characterized as threats to the internal validity, which refers to inferences about whether observed covariation between A and B reflects a causal relationship from A to B (Shadish, Cook and Campbell, 2002). Section 3 will address how these factors are dealt with in the EE program evaluation industry generally, and then specifically for gross and net savings.

3. Design and Analytic Issues in the Energy Efficiency Field

A translation of the potential confounding influences listed above into the factors that our industry recognizes as essential when conducting a consumption analysis, could look like this list:

1. Economic & political events & trends (History)

2. Weather (History & Selection)

3. Building characteristics (Selection)

4. Occupancy characteristics (Selection)

5. Geographic areas (Selection)
6. Motivations, attitudes, and behavior (Selection)
7. Changes in motivations, attitudes, and behavior over time, apart from program-related changes (Maturation, Statistical Regression to the Mean)
8. Naturally-occurring relevant installations and motivations to install (History and Selection, and perhaps Maturation)

An important aspect of Selection, as an influence, is that it can be thought of as being of at least two types: 1. program implementer selection (by design or by accident) and 2. self-selection into a program by participants themselves.

A model successfully controlling for Factors 1 through 8 above, whatever the design, would produce an estimate of net impacts. Factor 8 is quite specific to net impacts, while controlling for Factors 1 through 7 only, will generally lead to what we define as gross impacts. A simple pre/post regression with participants only will usually control for the first six factors and provide the basis for estimating gross effects without obvious bias. The exception to this is a situation where participants experienced changes in motivations, attitudes, etc. coincident with participation in the program and apart from the program effects on those factors. If changes of this kind occur, only a series of measurements over the studied period would allow the changes to be controlled for, and this is almost always impractical. So, Factor 7 muddies the waters a bit and represents a slight weakness in the pre-post design. Further, if the program was responsible for any attitude changes that occurred between the pre and the post period, the effects of those changes on usage would be attributed to the program's gross impact. However, for most programs, it is unlikely that they would have "moved the needle" on attitudes such that they would compromise the interpretation of pre- to post-program usage change to gross effects. The gross impacts would be captured by the coefficient representing the presence or absence of program-promoted equipment, unless confounded by changes in attitudes and the like. How the equipment installation is represented in the model is a subject for another paper. Representing/controlling for Factors 7 and 8, mainly falling into the influence categories of History and Selection, when adequately controlled, would yield an estimate of net impacts. Note that some comparison groups can serve to control for factors 1 through 7, and the result would be gross impacts, which is counterintuitive to many people who think of comparison groups as always producing net impacts since 1 through 7 are also threats to internal validity. Next, we turn to more detailed consideration of gross[4] and net impacts or savings.

3.1 Gross Impacts

In this section, we describe issues and designs relevant to estimating gross savings. As noted in the preceding section, controlling for factors 1 through 7, and not 8, will generally provide gross program effects. There are a number of methods that researchers in this and other fields use to produce what we would call gross savings. We present basic descriptions of the most common here.

[4] This paper focuses on statistical models for estimating gross impacts and ignores engineering methods.

3.1.1 Pre-Post Participant-Only Pooled Cross-Sectional Time-Series Design

The pre-post participant-only pooled cross-sectional time series model (Wooldridge, 2002; Kennedy, 2008) has, over the last 30 years, been the most commonly used regression approach to estimate gross savings. One basic specification of such a model is illustrated in Equation 1.

$$ADC_{it} = \alpha_i + \delta_t + \beta_1 Post_t + \beta_2 HDD_{it} + \beta_3 CDD_{it} + \sum \beta_k X_i + \varepsilon_{it} \qquad (1)$$

Where:

ADC_{it}= Average daily consumption (kWh or therms) for household i at time t

α_i= Household-specific intercept

$\delta_t = 0/1$ Indicator for each time interval t time series component that tracks systematic change over time

β_1= Coefficient for the change in consumption between pre and post periods

β_2= Coefficient for HDD

β_3= Coefficient for CDD

Post = dummy variable for pre (Post=0) and post (Post=1) participation

HDD_{it}= Sum of heating degree-days (e.g., base 65 degrees Fahrenheit)

CDD_{it}= Sum of cooling degree-days (e.g., base 75 degrees Fahrenheit)

β_k = A vector of k coefficients that reflect the energy change associated with a one unit change in the kth explanatory variable

X_i = A vector of explanatory variables (i.e., covariates), such as changes in occupancy or square footage, for the ith factor

ε_{it} = Error

This model works reasonably well as long as three conditions are met: 1) participation is reasonably well distributed across the time periods during the program year, 2) there are enough time periods (e.g., daily or monthly consumption) of pre-installation consumption and post-installation consumption observations, and 3) there is sufficient statistical power[5]. Meeting the first condition allows for some control of exogenous factors such as changes in economic and political events (history) and changes in motivations, attitudes and behavior (maturation) in the general population, both of which might over time affect energy use. Such a distribution allows for the representation in the analysis of a wider range of customers who are differently affected by time-related events. Meeting the second condition allows for estimating any seasonal effects for weather-sensitive

[5] Power is the probability that you will detect a true “effect” that is there in the population that you are studying. Put another way, the power of a statistical test of a null hypothesis is the probability that it will lead to a rejection of the null hypothesis when it is false, i.e., the probability that it will result in the conclusion that the phenomenon exists. The “effect” could be a difference between two means, a correlation between two variables (r), a regression coefficient (b), a chi-squared, etc. Power analysis is a statistical technique that can be used (among other things) to determine sample size requirements to ensure that statistical significance can be found.

measures. Meeting the third condition increases the chances that the estimated savings will be statistically significant.

An additional set of variables can also be included to provide controls for exogenous changes. That is, these variables attempt to capture the effects of economic, historical, and social conditions that can be explicitly modeled. Examples of variables that could be included are:

- Real per capita personal income provided quarterly, by Metropolitan Statistical Areas (MSAs)
- Local unemployment rate
- Local or national consumer price index
- Time period (e.g., months or days) dummies to control for factors that change over time that are not specifically modeled. If good measures of the variables, it is likely more effective to include those in the model rather than the generic time variable. There is danger in including the time variable in that it could absorb some of the treatment effect when pre- and post-participation periods are in the model.

Note that the first three types of variables listed above are only available on a monthly basis and can only be used in monthly models.

Of course, this specification works as long as all installations are early replacement, i.e., there is no need to normalize the gross savings to account for different baseline assumptions for equipment that is replaced on burn-out. However, a large number of participants whose installations represent a mix of early and normal replacement can complicate the analysis since a method for adjusting these models to account for different baselines has yet to be agreed upon.[6]

3.1.2 Cohort Design

To estimate gross savings for any program, including, and maybe especially whole building programs, it is very appealing to use future participants in the evaluated program as the source of a comparison group for the evaluated participants. This approach is sometimes called a cohort design. We describe here the structure of this approach, and then consider the pros and cons of using it.

Figure 2 illustrates the cohort design, which we have simplified to conserve space. Program Cycle 1 covers 9 months with all subjects being treated in month 5 only. Program Cycle 2 also covers 9 months with all subjects being treated in the 5th month (shown as month 14) only. For both current and future participants, we have ongoing monthly measurements covering both cycles. The measurements for program months 1 through 9, for the future participants in Program Cycle 2, serve as the comparison for months 1 through 9 of the participants in Program Cycle 1.

[6] Note that Agnew and Goldberg (2009) have developed methods for addressing this issue but only for the installation of a single measure, a central air conditioner.

Figure 2. Cohort Design

	Program Cycle 1									Program Cycle 2								
Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
Current Participants	O	O	O	O	X	O	O	O	O	O	O	O	O	O	O	O	O	O
Future Participants	O	O	O	O	O	O	O	O	O	O	O	O	O	X	O	O	O	O

O=Recorded monthly measurement
X=Program participation

This design allows us to control for exogenous factors e.g., changes in economic and political events (history) and changes in motivations, attitudes and behavior (maturation) in the general population, both of which might over time affect energy use, and for self-selection. The latter factor is particularly difficult to account for in non-randomly assigned comparison groups. Under the right conditions, this provides the basis for a good estimate of gross savings. This design will be effective in controlling for such exogenous factors and self-selection as long as these conditions hold:

1. The program design (e.g., the mix of measures promoted, the size of the rebates, etc.) remains stable,
2. The program delivery (the mix of participating contractors, their qualifications, and training, the types of customers targeted and successfully recruited, the marketing materials and channels used) have remained stable over the period of participation for both evaluated and future cohorts,
3. Future participants have not installed any program-qualified measures in the year prior to their own participation (prior to month 14 in our above example). That is, any changes to their consumption are due only to these exogenous factors, and
4. There is sufficient statistical power (as with all comparison groups).

Equation 2 illustrates one possible specification of this model.

$$ADC_{it} = \alpha_i + \delta_t + \beta_1 Post_t + \beta_2 Part + \beta_3 Post * Part + \beta_4 HDD_{it} + \beta_5 CDD_{it} + \sum \beta_k X_i + \varepsilon_{it} \qquad (2)$$

Where:

ADC_{it}= Average daily consumption (kWh or therms) for household i at time t

α_i= Household-specific intercept

δ_t = 0/1 Indicator for each time interval *t*, time series component that tracks systematic change over time

β_1= Coefficient for the change in consumption between pre and post periods

β_2= Coefficient for participation (1=participant in Cycle 1 and 0=nonparticipant in Cycle 1 (i.e., future participants in Cycle 2))

β_3= Coefficient for the interaction of Post and Part and represents the gross savings

β_4= Coefficient for HDD

β_5= Coefficient for CDD

Post = Dummy variable for pre (Post=1) and post (Post=0)

Part=Dummy variable representing participation (1=participant and 0=nonparticipant (i.e., future participants))

Post*Part=Variable representing the interaction of the post and participant variables

HDD_{it}= Sum of heating degree-days (e.g., base 65 degrees Fahrenheit)

CDD_{it}= Sum of cooling degree-days (e.g., base 75 degrees Fahrenheit)

β_k = A vector of k coefficients that reflect the energy change associated with a one unit change in the kth explanatory variable

X_i = A vector of explanatory variables (i.e., covariates), such as changes in occupancy or square footage, for the ith factor

ε_{it} = Error

If the comparison group members installed any program-qualified measures, or did any work that reduced their energy use (including installing non-program-qualified equipment), the resulting estimate would move toward net impact. That is, they would to some extent represent the potential free riders among the eligible population.

Targeting the same types of customers over time increases the chances that future participants will be very similar to the evaluated participants with respect to their demographics, attitudes, energy use, and building type etc. This is important since selection factors (program- and self-) and their correlates can be an important aspect in how comparable the two groups are, and future participants can be very helpful in allowing self-selection factors to be adequately controlled. This occurs because both current and future participants will have self-selected into the program, just at different points in time.

We note here that some have found the use of a comparison group composed of future participants confusing and we find that the assumptions regarding its use are rarely tested. Customer targeting can change dramatically from one year to the next. For instance, one whole-house program, had targeted coastal customers during the initial program roll-out. The evaluators recommended targeting more inland areas where winters are colder and summers are hotter, thus producing more savings for participants. This is a case where using future participants was not appropriate for estimating gross savings.

Sometimes it is not as obvious that future participants will not provide an adequate comparison for accurately estimating gross savings. This can happen when future participants make some upgrades in the year prior to their participation. It is highly unlikely that the customer would have installed a full complement of home upgrades in the year prior, but they might have done some upgrades so that they do not meet the third condition above. The only way the evaluator is likely to discover this is if she surveys these future participants, or a sample of them, to determine whether such installations are sufficient to question the accuracy of the resulting estimate of gross savings.

We consider it essential that evaluators who plan to use the following year's participants as a comparison group to control for exogenous factors, test the comparability of the two groups before proceeding with the plan. Taking this recommendation seriously means that the evaluator and the PA must be flexible enough to change designs if assumptions of the planned design are not met, and could mean surveying a sample of future participants to assess their comparability. Flexibility is required because such a test could not be done until the "future" cohort has been identified, which will likely be a year after the evaluated cohort was identified. Thus, if the two cohorts are not very similar, this design becomes unfeasible and a back-up plan will be needed.

We conclude this section with a discussion of one other source of confusion. Traditional research design literature presents the cohort design in the context of estimating (implicitly) the net savings of a program. As mentioned earlier, the main advantage of this design is that selection biases, introduced by adding a comparison group, are reasonably well controlled assuming the composition of the current and future participants is similar and that the design and implementation of the intervention has not substantially changed over time.

An example used by Campbell and Stanley (1963) is an officer and pilot training program, whereby participants in year one (cycle 1) are compared to participants in year two (cycle 2). The assumption is that training to be an officer is only available through the Army's training program. In other words, during the first program cycle, the future participants in the second program cycle (or any soldier eligible to participate in the second program cycle) could not have been exposed to any training that would have prepared them to be an officer or pilot. As a result, they provide an unbiased estimate of what members of the eligible population would have done in the absence of the program, i.e., the net impact of the program.

Thus, a traditional research design text would consider this design to produce the net effects of the officer and pilot training program rather than gross effects (if, indeed, they made that distinction). An illustrative example: the evaluation of the California 2005 Low Income Energy Efficiency (LIEE) Program relied on a cohort design as one way of estimating net savings. This design was appropriate since the future participants represented what the larger eligible population of low-income households would have done absent the program, which is essentially nothing since they were very likely unable to afford purchasing any new equipment. Evaluators who use the cohort design to control for exogenous factors in estimating gross savings should clearly explain that historically such a design has been used to estimate net savings but it is being used, in this particular instance, to estimate gross savings, assuming that the conditions mentioned earlier have been met.

3.1.3 Two Stage

Gross savings can also be estimated using the two-stage approach outlined in Chapter 8 of the Uniform Methods Project[7], an approach that is consistent with IPMVP[8] Option C, a site-specific, whole-building regression analysis approach that allows for an existing-conditions baseline in estimating gross savings.

Stage 1. Individual Premise Analysis

[7] Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures (http://energy.gov/eere/about-us/ump-protocols)

[8] International Performance Measurement and Verification Protocol (IPMVP) available from the Efficiency Valuation Organization at http://evo-world.org/en/

Stage 1 involves the following three steps:

1. Fit a premise-specific degree-day regression model (as described in Step 1, below) separately for the pre- and post-periods.
2. For each period (pre- and post-) use the coefficients of the fitted model with normal-year degree days to calculate the normalized annual consumption (NAC) (defined below) for that period.
3. Calculate the difference between the pre- and post-period NAC for the premise (i.e., ΔNAC).

Step 1. Fit the Basic Stage 1 Model

For each participating site, estimate the following model:

$$E_m = \mu + \beta_H H_m + \beta_C C_m + \varepsilon_m \qquad (3)$$

$E_m =$ Average consumption per day during interval m

$H_m =$ Specifically, $H_m(\tau H)$, average daily heating degree days at the base temperature(τH) during meter read interval m, based on daily average temperatures on those dates

$C_m =$ Specifically, $C_m(\tau C)$, average daily cooling degree days at the base temperature(τC) during meter read interval m, based on daily average temperatures on those dates

$\mu =$ Average daily baseload consumption estimated by the regression

$\beta_{H,}\beta_C =$ Heating and cooling coefficients estimated by the regression

$\varepsilon_m =$ Regression residual.

Step 2. Apply the Stage 1 Model

To calculate NAC for the pre- and post-installation periods for each premise and timeframe, we combine the estimated coefficients μ, β_H, and β_C with the annual normal-year or typical meteorological year (TMY) degree days H_0 and C_0 calculated at the site-specific degree-day base(s), τH and τC. Thus, for each pre- and post-period at each individual site, we use the coefficients from Equation 3 for that site and period to calculate the weather-normalized annual consumption (NAC) (see Equation 4). This example puts all premises and periods on an annual and normalized basis.

$$NAC = \mu * 365 + \beta_H H_0 + \beta_C C_0 \qquad (4)$$

The same approach can be used to put all premises on a monthly basis and/or on an actual weather basis.

Step 3. Calculate the Change in NAC

For each site, the difference between pre- and post-program NAC values (ΔNAC) represents the change in consumption under normal weather conditions.

To control for the exogenous changes mentioned earlier, a comparison group, either composed of prior participants, future participants (i.e., cohort design) or a contemporaneous group of nonparticipants, these same three steps are followed.

Stage 2. Cross-Sectional Analysis

Next, the cross-sectional model in Equation 5 is estimated incorporating both current and future participants.[9]

$$\Delta NAC_j = \beta + \gamma I_j + \varepsilon_j \qquad (5)$$

$I_j =$ 0/1 dummy variable, equal to 1 if customer j is a (current-year) participant, 0 if customer j is in the comparison group composed of future year participants.

$\beta, \gamma =$ Coefficients determined by the regression model

$\varepsilon_j =$ Regression residual.

From the fitted equation:

- The estimated coefficient γ is the estimate of mean savings.
- The estimated coefficient β is the estimate of mean change or trend unrelated to the program.

The coefficient β corresponds to the average change among the comparison group, while the coefficient γ is the difference between the comparison group change and the participant group change. That is, this regression is essentially a difference-of-differences formulation and can be accomplished outside of a regression framework as a difference of the two mean differences. More complex models that include other available premise characteristics can be included that can improve the extrapolation of the billing analysis to the full population.

3.2 Net Savings

The preceding section focused on gross impacts. This section focuses on net savings, though gross savings are sometimes necessarily mentioned by way of comparison.

3.2.1 The Counterfactual—Single-Measure Programs

[9] The sole purpose of this second stage is to control for exogenous changes through the use of prior participants, future participants or a contemporaneous group of nonparticipants. If one did not need, for whatever reasons, to control for these exogenous factors, then only Steps 1 and 2 in the first stage would be required.

All of the factors that must be controlled to produce a valid estimate of gross program savings are equally applicable to producing net savings. Designs producing net program savings are distinguishable by the need to control for additional factors.

As we have asserted before in this paper, the counterfactual is applicable only to estimating net savings. In fact, it is central to the endeavor. In the broader world of evaluation research, it is defined as what would have happened had an investment (in a program or intervention) not been made. In our industry, this translates to what a member of the eligible population/customers would have done if the program under evaluation were not there. Thus, we either need to measure directly the hypothetical situation of what participants would have done absent the program (using the self-report approach), or we must identify a comparison group that can reasonably represent the counterfactual. As this paper is about comparison groups, we discuss that alternative in some detail, next.

3.2.1.1. What Was or Might Have Been Installed

For a comparison group to support estimating net savings, it must represent the counterfactual. But measuring this concept is extremely complex. We can think of two aspects of the counterfactual: 1. Motivation (the *why*) and 2. the action taken, specifically what was installed, if anything. So, in addition to trying to represent in a comparison group, the *why* of participant installations (program-influenced or not), we also have to consider *what* was installed in the comparison group, and whether that technology is a suitable point of comparison for what participants installed. Even when considering only what equipment installations the counterfactual should represent, it is complex. In some programs, there is no alternative to the program equipment, or no non-efficient alternative. In others, there are various efficiency-rated alternatives.

Specifically, in cases of equipment such as air conditioners and furnaces, the customer could replace the existing equipment with something less efficient than what the code requires, or he could purchase code-compliant equipment, or he could choose a version that goes beyond what the code requires, with or without the program's influence, and to a greater or lesser degree. Other types of program-promoted equipment are either present or absent and do not consume energy. Examples of this type are duct sealing, wall insulation, or thermostats. So, the counterfactual question becomes: What might the participating customer have installed without the program, if anything? Table 1 provides some examples of the kinds of equipment that efficiency programs might promote and some possible installations that *could* represent the counterfactual *in terms of the type of equipment installed*. As a reminder, we are talking only about single-measure programs at this point. Also, question marks indicate greater uncertainty about the possible comparison group.

Table 1. New Equipment Installed During Program Evaluated Year

1 Participant-Installed Program Measure	2 Comparison-Group-Installed Measure
SEER 17 Air Conditioner	Any Air Conditioner
SEER 19 Air Conditioner	Any Air Conditioner
Tankless Water Heater	Any Water Heater
Duct Sealing	Any house with a working HVAC system that uses ducts that were not insulated during the evaluated period?
R30 Wall Insulation	Any Wall Insulation? No Wall Insulation?
Envelope Sealing	Any house with a working HVAC system that has not been sealed?

If we could assume that any customer who took an action in the second column of Table 1 represents what participants would have done in the absence of the program that promotes the measures in column 1, we could find a comparison group that represents the installed-equipment aspect of the counterfactual, and thus be able to estimate program net effects, provided we had also controlled statistically for the first seven factors listed earlier. This would require finding customers for the comparison group who had installed these things or, something analogous to it (efficient or not), or had the opportunity to. Finding such customers can be expensive, but not impossible, as evaluators have done this many times. However, it isn't always entirely clear what the right actions would be to constitute a good comparison group customer. What is the right comparison group member for a program that promotes duct testing and sealing? Or envelope sealing? Perhaps it is the customer who has not done that work but would benefit from it? The answer isn't apparent, but it is essential to address this issue in designing a comparison group.

3.2.1.2. Awareness, Motivation and the Size of the Eligible Population

Another central issue in finding an appropriate comparison group that will represent the counterfactual in estimating net savings can be described as the motivation and awareness of the customer making the equipment choice; in fact, this is key. The customer who is motivated to install a program-qualified measure regardless of incentive, if aware of the incentive, is highly unlikely to refuse it. (The most altruistic, committed environmentalist might do that so that the incentive could be used to motivate a less motivated installer.) Thus, environmentally-motivated customers would naturally be to some extent under-represented in a non-participant comparison group for such a program, unless the customer was unaware of the program. But an aware customer might also refuse the rebate because they perceive applying for the rebate to be a hassle. So, a comparison group pool of customers might not be completely devoid of environmental or convenience motivated customers.

The foregoing means that the only possible comparison group member for a program would be the customer who is unaware of the program or who is aware but, for whatever reasons, chose not to participate. Over time it might become more and more difficult to find such customers. If we do find them, we have to ask if those customers have the same rate of naturally choosing efficient alternatives that the participants have. Maybe the unaware customers all live in very rural areas. Would they have the same naturally-occurring rate of choosing efficient options? Do they have the same opportunity to purchase the efficient options? The answer to both is probably, No. A design that included customers that had a different set of opportunities and motivations to choose efficient equipment compared to the participant group would fail to comply with the principle of ITA. Thus, good comparison groups are unlikely to be available unless the program meets one or more of the following conditions:

1. it was relatively new,
2. it was driven by relatively few participating contractors,
3. it is only offered in a few areas,
4. the eligible population was large, and/or
5. the level of program awareness was low.

Some programs will meet one or more of these conditions and some will not. And if they do, the evaluator must address additional complexities. In any case, where there is a large pool of non-participants under these conditions, it becomes potentially feasible to find an appropriate comparison group by further matching and/or screening, *in terms of observable variables.* Of course, a core issue in estimating net effects is the set of largely unobservable factors involved in self-selection. If the program is a deep-savings oriented program such as a whole-building program, there are additional complexities that are discussed in Section 3.2.2, and in the context of Chapter 8, in Section 4.1.3.

3.2.2 The Counterfactual—Whole Building Programs

The issues in representing the counterfactual with a comparison group are compounded for whole-building programs. We find that discussions of comparison groups and counterfactuals are often carried out with the example of an air conditioner rebate program, and treated as if this represents all of the various program types. We find that there are issues unique to each program type, and that it is important to consider this specifically when deciding the right approach to estimating net program savings, including whole-building programs. In the whole building scenario, Table 1 still applies, but we have to think about the entire list of measures and how the group of measures installed under the program would be represented in the comparison group, if we wanted the comparison group to support estimating net effects. At first glance, it would seem that the comparison group measure categories (column 2, in Table 1) would have to be represented in the same proportion as their counterparts in column 1. But this is called into question when we consider customers who took some, but not all of the program-promoted measures. Is the customer who did some envelope sealing and some insulation a counterpart to the program participant who did those things plus several others all under the program? Is the customer who did the envelope sealing and some insulation a good counterfactual match for the participant who had done the same things before participating, and installed a new heat pump, and a tankless water heater, and did duct sealing all under the program?

While the issue of what mix of measures constitutes good candidates for eligible customers to be counterfactual representatives is particularly complex for whole-building program evaluators, there are some issues that make it easier for evaluating this type of program compared to single-measure programs:

1. The eligible population is likely large, this type of program is relatively new, and, because they tend to be contractor driven, there will be many customers who are not aware of the program. To participate in the program, one generally needs a contractor who is approved by the program, and there are a limited number of approved contractors.

2. Some, though not all, of the participants would be recruited into the program by a contractor who is using the program as a sales tool. Customers consulting with a non-participating contractor will not be exposed to the program and thus may be unaware of it. Because there are many contractors who are not associated with the program, there may be many customers interested in an energy-related (not necessarily energy-efficient) upgrade or renovation who are not aware of the program.

3. One could argue that any home upgrade or renovation is an opportunity to include energy-efficiency measures in it. To the extent that customers decide to do that outside of the program, this would approximate the naturally-occurring rate of such measures in this context. To the extent that they decide not to, or never think of it at all, this would represent the other part of the counterfactual. Thus, any home upgrade or renovation could be a legitimate comparison group member for a whole-house program that would yield net savings as long as self-selection factors are accounted for.

These three situations make the task of finding appropriate counterfactual representatives considerably easier, in the sense of finding potential members, and the possibility should not be dismissed lightly. It also implies a more complex set of decisions about which potential members should be included and excluded. The large pool of unaware renovators does not automatically constitute a net-effects-supporting comparison group. More matching and/or screening, at least, would be required to produce such a group. This is not to say that a perfect comparison group can be found. It will always be imperfect, but that is true of any approach short of an experimental design. But we think it is important not to make the perfect the enemy of the good. All the decisions resulting from these complexities, and their rationales should be documented. The implications of these facts for Chapter 8 are discussed in Section 4.1.3.

3.2.3 Potential Comparison Groups for Whole-Building Programs

We consider the multiple-measure, expensive upgrade approach of a whole-building program a critical factor in thinking about appropriate comparison groups, as discussed above. Any one or two measures installed outside of the program would not qualify a customer to be in a comparison group meant to represent the counterfactual for a whole-building program. To qualify a customer as a member of such a comparison group, he must have done a substantial building and equipment upgrade. Of course, he need not have installed program-approved efficiency levels because the frequency of customers doing that or not is what would provide the baseline for estimating net effects. One could make the argument that any substantial building and equipment (in combination) upgrade or renovation would qualify a customer to be a comparison group member. Contractors could make an excellent source of a whole-building program comparison group if they could be recruited.

Permits are another potential source of comparison groups, and these are reflected in public records. Standard building permits in many jurisdictions will be present only for those who increase the living areas of the homes, in the case of residential buildings. Others, like California, will require permits for a wide variety of measures as either Title 24 or title 20 (both are efficiency codes) is triggered. However, again, many DIY projects and those done by contractors that aren't vigilant about code compliance will be missed as well. Some discussion/consideration of whether these possible missing projects would unduly compromise a net savings comparison group is called for. There could be arguments for and against them. But projects of the size of most whole-building participants' will, in some jurisdictions, at least, trigger permits and inspections of some kind that could be accessed by evaluators.

The most inclusive approach to generating comparison groups for this type of program should also not be dismissed lightly, and could be quite feasible. Specifically, surveying the eligible population with screening interviews could also generate customers who had completed a whole-building renovation or upgrade, some of which would have included energy-efficient measures without benefit of a program. The incidence of qualified customers would be higher for whole-building programs than for single-measure programs, since a variety of measure combinations could qualify the customer for comparison group status. The complexity of this or any approach would be to determine which and how many combinations of measures would be considered comparable to what participants did. Evaluators would have to address these issues directly and make their decisions clear.

Future participants as members of whole-building comparison groups were discussed in the Gross Impact section. But there is reason to think that at least some measures are of comparable types to whole-building participants' measures, though probably not enough to constitute a full net-effects-oriented comparison group.

A final point on this topic: A comparison group generated through contractors, permits, or surveys, with the intent to find customers to represent the counterfactual for a whole-building program, need not complete upgrades of the very same size as the participant projects. Once it is established that a customer undertook what they would define as an upgrade or a renovation of their building/home, that customer could have chosen a larger, more expensive project that would result in more energy savings. Introduction to the program might well have influenced them to increase the project size and budget in order to take advantage of the program. Thus, such a customer could well be considered an appropriate comparison group member. We suggest that it is only important to establish that the customer intended to do a substantial upgrade or renovation.

There is a lot to consider in deciding whether a particular comparison group represents the counterfactual well, incompletely, or not at all for different measures and groups of measures. In fact, there is a lot to consider in just deciding what a suitable counterfactual would be for many of the individual measures, not to mention groups of measures. Of course, all of the issues addressed in earlier sections also come into play in deciding on a comparison group design, including how many customers who are unaware of the whole-house program can be found. In addition, what constitutes an “eligible” customer must be defined. As already mentioned, being a homeowner is essential, and very likely a certain minimum income level is also important. The customer income factor may be affected by the presence of efficiency-based financing programs. All of these factors must be considered in deciding on a design and on how close the design comes to producing net or gross program savings.

The foregoing discussion forms the background from which to consider some critiques of Chapter 8 of the Uniform Methods Project.

4. Some Critiques of Chapter 8

Our reading of Chapter 8 of the Uniform Methods Project is that it is unnecessarily limited in several ways, which we will describe in this section. We also have a few differences of opinion regarding decisions and statements made in the document. Finally, we find some sections of the chapter confusing. Based on these observations, and on our discussions in Sections 2 and 3, we make specific critiques and suggestions for a Chapter 8 that we propose be revised along these lines. We divide our critiques into three categories: Differences of Opinion, Omissions, and Lack of Clarity and Inconsistencies.

4.1 Differences of Opinion

4.1.1 Recommending the Two-Stage Model for Consumption Analysis

The two-stage design for estimating gross savings essentially means running individual regression models for each participant and comparison group building to adjust usage by the weather values during the modeled period. In stage 1, *the weather-normalized annual consumption (NAC) is estimated separately for each building* based on a regression for 12 months of energy use in the pre-period and another regression for 12 months of energy use in the post period. For participants, the difference between the building's weather-normalized pre- and post-program NAC represents the program-related change in consumption plus exogenous change. For non-participants, the weather-normalized pre-post difference represents only exogenous change. Stage 2 takes as its input the output of stage 1, savings for each building, and completes a cross-sectional analysis of the participant and comparison groups.

A central problem with this method is that in the first stage, for both the participant and comparison group buildings, the only variable included in the pre- model and the post- model is weather. There is no ability to control for other exogenous factors. To the extent that this is the case, the models are very likely misspecified (omitted variables), leading to a biased estimate of the change in energy use from the pre- to the post- period, i.e., the dependent variable in stage 2. If the models are biased for one or both groups, the resulting weather-normalized estimate of gross savings is biased. We think there are other methods that would deal more effectively with the problem of misspecification. In particular, multi-level modeling would work well here. We discuss that in Sub-section 4.2.2.2 as part of the Omissions section.

4.1.2 Making the Two-Stage Model the Default Approach

In Chapter 8, the two-stage approach is clearly the implicit preferred choice. However, given the serious problems, discussed in Section 4.1.1, why not recommend the approach in Row 3 of Table 1, a participant-only pooled-time-series cross-sectional model (described in Section 3.1.1) as the preferred or default approach? Except in rare situations where there is insufficient pre or post monthly/daily data or insufficient statistical power, we have never seen a situation in which the conditions repeated below from Chapter 8 cannot be met, at least any situation where a billing analysis is appropriate.

- ***A balance of participant installation intervals across at least three billing intervals,*** preferably more. Having a balanced participation across three intervals would ensure that two-thirds of the participants provide a steady-state comparison during each interval of change. In the extreme, with only a single start date (as with a program that starts mailing comparative usage reports to homes at the same time), the model fails to control for exogenous change across the change point. This explains the more stringent requirement for these programs of a randomly assigned experimental design.

- ***A balance of data between pre- and post-installation periods with respect to the number of data points per household and the seasonal coverage.*** Similar seasonal coverage in the pre- and post-installation is particularly important if measure savings are temperature sensitive. For gas heat modeling, the model should include at least one full winter in both the pre- and post-periods *and* some non-heating months. A full year of pre- and post-installation data removes concerns regarding imbalanced data. (p. 8-24)

We suggest that Chapter 8 say something about how rare such a situation is. In the end, the conditions that must be met to use the pre-post analysis of participants approach seem far fewer than those for the two-stage approach. The pre-post method also has the advantage of not including any part of net effects in the gross impact estimate, which can be a problem with the future-participants approach.

There are additional designs and analysis methods that we believe would be preferable to the two-stage model approach, and they seem not to have been considered in Chapter 8. We describe these possibilities in Section 4.2, along with other issues that seem to us to have been omitted from consideration.

4.1.3 The Possibility of Finding Adequate Comparison Groups for Net Savings

As we noted in Section 3.2.2, good comparison groups for estimating net savings, based on “eligible populations” are unlikely to be available unless the program meets one or more of the following conditions:

1. it was relatively new,
2. it was driven by relatively few participating contractors,
3. it is only offered in a few areas,
4. the eligible population was large, or
5. the level of program awareness was low.

Chapter 8 assumes that these conditions can never be met, a position that we examine next.

After discussing the RCT design in Row 1, Chapter 8 notes that:

> Where a program is not designed as an RCT, a comparison group is developed after the fact in a quasi-experimental design framework. For that design framework, the term “comparison group” denotes groups that are not randomly assigned, but still function as experimental control groups (p.8-6).

The chapter goes on to say:

> Customers and contractors inclined toward EE have little reason not to take advantage of the rebates. This is likely to lead to an over-representation of natural adopters in the participant population, as compared to the general incidence in the population. This, then, affects in multiple ways the level of savings and freeridership that will be measured by the consumption data analysis.
>
> - First, any comparison group developed after the fact from those who chose not to participate will tend to have a lower percentage of energy-efficient furnace installers (in this example) than would a true control group. To the extent that this is the case, the comparison group will not control for the full extent of natural energy-efficient furnace installations had the program not been in place.
> - Second, the treatment group includes a higher proportion of natural EE adopters than the general population, due to self-selection into the program. These households increase the freeridership rate beyond the natural level of natural adopters in the eligible population.
> - Finally, the more general concerns regarding self-selection are still present. Because of their natural inclination to adopt EE, the participants are likely to exhibit different energy-consumption characteristics than the general population.
>
> These are the key factors that make it difficult to define fully the measured differences in consumption for the participant and comparison groups. As a result, when comparison group change is netted out of the participant change, the netting will control for some but not all of the naturally occurring measure implementation leaving an unknown amount of free ridership in the final savings estimate. The resulting estimate is thus a mix of net and gross savings. (p. 8-7, 8)

These claims rest on at least five assumptions: 1) the program is mature, 2) if the program is contractor-driven, most or nearly all contractors are participating, 3) the number of households/buildings that are in the market for the measure(s) being incented by the program is relatively small, 4) the effort to make every eligible household aware of the program is highly effective, and 5) that all or nearly all of the aware households would upgrade the home/building through the program, and not on their own.

When thinking about comparison group issues in terms of air conditioner replacements, the case is often made that virtually everyone who purchases this appliance learns about relevant rebate programs and is offered (and takes) the rebate, including those who would have installed efficient equipment regardless. Thus, any comparison group would consist of customers who were quite out of touch, or who were extremely altruistic because they refused the rebate, in addition to customers who chose equipment that just met code. This would not be an appropriate comparison group for either net or gross savings. In contrast, whole-building programs are newer and are largely contractor driven and contractors who do upgrades or renovations that are energy related are in the majority, but those who participate in whole-building programs are a small minority. One implication of this is that it is entirely feasible for a customer to do an energy-related upgrade without any knowledge of the program. There will also be a subgroup who learn about the program but decide not to participate for a variety of reasons, including that the contractor sold against the program so that both could avoid the hassle. Customers might also be out of touch simply because the program is new and/or offered only in certain regions due to budget constraints.

For a single-measure program like an air conditioner replacement program, the eligible population is the population of customers who have purchased a new air conditioner. If the evaluator were able to overcome the problem of finding customers who had not been exposed to the program, that part of the eligible population appropriate to a net effects comparison group would be those who purchased and installed some air conditioner, whether efficient or not. Similarly, the eligible population for a whole building program would be those who have completed an upgrade, perhaps one with energy-related elements. This is a broader group than those who simply needed to replace one type of equipment. Specifically, while a home upgrade could be triggered by the failure of some piece of equipment like an air conditioner, it need not be, and other add-ons and replacements can be included as part of the upgrade process whether there is existing working equipment or not. Thus, the pool of potential comparison group members is broader than for single-measure programs. It would be the customers who completed an upgrade during the program year. An income minimum might be considered as a qualification for comparison group membership. But we don't think the evaluator should be ready to assume that there is no potential comparison group available for assessing the program savings of a whole house or whole building program on the same basis that it is assumed for single-measure programs. Such customers could be found by traditional survey screening or by other sources such as building permits, or, with greater difficulty, from contractors. Wherever they might be found, we believe that Chapter 8 should require that evaluators be specific about who the eligible population is, and the chapter itself might be more explicit about defining it for the whole building type of program.

As discussed earlier in the paper, these features add complexity to the thinking about methods of estimating program savings, but they also make generating contemporaneous comparison groups more feasible. Anyone who has completed a home renovation project that impacted energy use or that could have impacted it, could be considered a potential member of a comparison group for a whole-house program. Almost any home upgrade or renovation could be turned into an energy-related upgrade and therefore an energy-efficient upgrade. Renovators who choose a contractor that doesn't participate in the local whole building program will be much less likely to choose EE measures in their renovation, and that is a condition that could be argued to represent the counterfactual. For those customers, the whole-building program doesn't exist, so their choices about the renovation components may reflect the naturally-occurring rate of choosing energy-efficient versions.

Below, we focus on two of the more prominent whole-house programs to illustrate another way to think about the size of the eligible population.

Consider the evaluation of the 2010-2012 California Whole House Retrofit Program (known as EUC)[10] in which about 3,750 residential customers participated each year from 2011-2012. The 2011-2012 General Households Population Study in California[11], conducted in January/February, 2012, found that over 23% of the 7.5 million owner-occupied homes in California had actually completed a comprehensive home upgrade[12] since January 1, 2010, covering a two-year period or about 862,500 annually. Of these, only about 20% were aware of EUC. Thus, approximately 855,000 (862,500 - 7,500) residential customers performed a comprehensive upgrade of their homes each year outside of the Program and of these only about 171,000 were aware of EUC, leaving a large number who would eligible for comparison group membership. For this Program, the size of the eligible market is large and the level of awareness of the EUC is small making it highly likely that a suitable comparison group could be formed that could avoid the problems identified in Chapter 8.

Or, consider the California Energy Savings Assistance (ESA) Program. The number of low income households in California that qualify for ESA and the California Alternate Rates for Energy (CARE) Program[13] has been estimated to be about 4.1 million households (32% of all California households). Fifty-nine percent of 2012 eligible California IOU households have been treated by ESA during the period of 2002-2012, leaving 41 percent or 1.7 million untreated California IOU households (Evergreen Economics, p. iv, 2013). Evergreen (2013) also asked telephone survey respondents who were on the CARE Program whether they are aware of ESA after providing them with a general description of the program. Since the program had recently changed its name from the Low Income Energy Efficiency (LIEE) Program to the Energy Savings Assistance Program, the surveyors did not ask an unprompted question about awareness of “Energy Savings Assistance Program”. While awareness was reasonably high with two-thirds of respondents reporting that they were aware of ESA, unprompted awareness is always lower. Again, while penetration is higher today than in 2012, it seems that the pool of nonparticipating eligible homes is sufficiently large that a reasonable comparison group could be formed thus avoiding the problems identified in Chapter 8.

In the end, little evidence is presented to support the claims that the differences between participants and a matched group of eligible nonparticipants remains large and that the available statistical methods to control for these observed and unobserved differences are largely ineffective. If these claims turn out to be true, the arguments by many (including The E2e Project and the CPUC’s Office of Ratepayer Advocates) for a greater reliance on comparison group designs for estimating net savings would be made moot. Note that the whitepaper being written in parallel with this whitepaper should shed some light on recent methods to control for self-selection. On the other hand, if these claims turn out to be at least partly false, Chapter 8 should suggest that readers determine whether, for these types of programs, these two claims are true and, if not,

[10] Available at www.calmac.org, Study ID CPU0093.01

[11] Available at www.calmac.org, Study ID SCE0321.01

[12] A comprehensive energy upgrade includes the following: sealing areas around windows and doors, insulating walls and attic, replacing windows, roofs, and ducts, and if replacing appliances, installing high-efficiency appliances, including air conditioners, heat pumps, water heaters, and furnaces. In other words, it includes a whole package of upgrades of this kind.

[13] CARE, the California Alternate Rates for Energy program, provides a monthly discount on energy bills for income-qualified households and housing facilities. Qualifications are based on the number of persons living in the home and the total annual household income. FERA, the Family Electric Rate Assistance program, provides a monthly discount on electric bills for income-qualified households of three or more persons.

refer readers to Chapter 17 of the UMP (*Estimating the Net Savings: Common Practices*) and the companion white paper where best practices in the use of such designs are discussed.

These factors together suggest that it is much more feasible to find an appropriate comparison group for this type of program than for some other program types, or than is assumed in Chapter 8. We recommend that these differences be analyzed and highlighted in a revised Chapter 8.

4.2 Omissions

4.2.1 Incomplete Consideration of the Unique Features of Whole-Building Programs

Whole-house or whole-building programs, designed for producing deep savings, have a number of features that distinguish this class of programs from traditional single-measure rebate programs. We frequently use air conditioner rebate programs as a concrete example to help us think through comparison group issues. This example has the advantage of being simple and common; it allows consideration of efficiency levels and naturally-occurring rates of adopting program-qualified measures, as well as adopting measures that just meet code requirements. What it does not do is help us think about the unique features of whole-building programs and their implications for estimating gross or net savings. Similarly, the current version of Chapter 8 gives very little attention to these unique features and their implications. Below, we describe the two major issues that we suggest be more fully considered in a revised Chapter 8.

4.2.1.1. Size of Investment. To qualify for a whole-house program, customers must make multiple upgrades to their home or building that will improve its energy efficiency. This is an expensive proposition. The high expense of participating in this kind of program is particularly an issue for future-participant designs. A customer who participates in such a program is unlikely to have done very much in a prior year to install measures because if they had, they probably could not do enough in the following year to qualify for a whole-building program. This fact has a tendency to limit estimation of savings to gross savings since little or nothing would have been done by the Cycle 2 participants during the evaluated program year, for which they are serving as members of a comparison group. On the other hand, as we noted in Section 3.1.1.3, this is not entirely true. Some relatively inexpensive or DIY things, or even one expensive item could have been done by the future participants during the evaluation period (i.e., prior to their participation in Cycle 2) which could move savings estimates toward net. At a minimum, we recommend that this issue be addressed explicitly both in the chapter and as a requirement of evaluators of this kind of program.

4.2.1.2. Multiple Measures. As discussed 3.2.2, the fact that whole-building programs will involve multiple measures, some of which are efficiency rated, and others being add-ons that are generally done or not done, raises the issue of what would constitute a good comparison group for either gross or net impact estimates. The issue is complex. There are many possible combinations of measures that are installed in this type of program, so careful consideration of what measures should be included in a comparison group is needed. We recommend that this issue be addressed in Chapter 8 and that some requirement for addressing the issue be stated for evaluators.

4.2.2 Omission of Some Possibly Legitimate Designs for Gross and Net Savings

4.2.2.1. Pre-Post Participant-Nonparticipant Pooled Cross-Sectional Time-Series Design

The pre-post participant-nonparticipant pooled cross-sectional time-series design has been used in the past to estimate the net savings for residential programs and should be at least considered as one possible quasi-experimental approach. It is often referred to as the non-equivalent comparison group design and involves participants who have self-selected into the program and a group of nonparticipants. Equation 7 illustrates one possible specification of this model.

$$ADC_{it} = \alpha_i + \delta_t + \beta_1 Post_t + \ \beta_2 Part + \beta_3 Post * Part + \beta_4 HDD_{it} + \beta_5 CDD_{it} + \sum \beta_k \, X_i + \varepsilon_{it} \qquad (7)$$

Where:

ADC_{it}= Average daily consumption (kWh or therms) for household i at time t

α_i= Household-specific intercept

δ_t $= 0/1$ Indicator for each time interval t, time series component that tracks systematic change over time

β_1= Coefficient for the change in consumption between pre and post periods

β_2= Coefficient for participation (1=participant in Cycle 1 and 0=nonparticipant in Cycle 1 (i.e., future participants in Cycle 2))

β_3= Coefficient for the interaction of Post and Part and represents the gross savings

β_4= Coefficient for HDD

β_5= Coefficient for CDD

Post = dummy variable for pre (Post=0) and post (Post=1) participation in P4P

HDD_{it}= Sum of heating degree-days (e.g., base 65 degrees Fahrenheit)

CDD_{it}= Sum of cooling degree-days (e.g., base 75 degrees Fahrenheit)

β_k = A vector of k coefficients that reflect the energy change associated with a one unit change in the kth explanatory variable

X_i = A vector of explanatory variables (i.e., covariates), such as changes in occupancy or square footage, for the ith factor

ε_{it} = Error

Of course, the major threat to internal validity is self-selection, an issue that has been a challenge for such designs. An update on the approaches for controlling for self-selection is described in the companion white paper by Train et al. (2017).

4.2.2.1 Engineering Designs—Gross

There might be situations in which none of the first four designs in Table 1 of Chapter 8 are possible and so few variables are available for the eligible population that matching or controlling statistically for selection is impossible. That leaves us with the method described in Row 5 in which the comparison group is composed of the general eligible population for which there is little or no data on which to match the two groups or to control statistically for self-selection. Instead, when faced with such a situation, why not rely on Option D: Calibrated Simulation[14] as outlined in *International Performance Measurement and Verification Protocol* (IPMVP) as an option for estimating gross savings that could then be adjusted using a self-report net-to-gross ratio (NTGR)? Such an approach could at least control for some of the exogenous effects in the form of routine[15] and non-routine[16] adjustments and has the additional advantage of being relatively more transparent. Of course, one could conclude that both regression-based and engineering-based approaches contain too much error and rely on deemed savings for each installed measure.

4.2.2.2 Multi-Level Modeling—Gross or Net

Another approach for estimating gross savings could be the multi-level modeling method, with or without a comparison group. This is a statistical technique that allows variables to be controlled at multiple levels of aggregation, starting with climate zone, weather, jurisdiction, neighborhood, individual building, and time period (e.g. pre- versus post-period). Economic conditions that change over time can be included in the model. It does require more skill than average regression modeling, so it may not be the first choice, or the default, but it should be seriously considered. Used without a comparison group, the outcome would be gross savings. With a comparison group, depending on its composition, the result could be gross or net. Either way, its advantages include the ability to generate savings by individual buildings while simultaneously controlling for variables at higher levels of aggregation, such as weather or economic variables. While we do not generally use this method for standard consumption analysis for estimating measure or program savings, as it is generally not necessary, but we do think it is statistically preferable to the two-stage model being treated preferentially in Chapter 8.

4.2.3 Importance of Meeting and Demonstrating Compliance with Assumptions of Past and Future Participants as Comparison Groups

The stated benefits of using future (or past) participants as a comparison group, particularly the elimination of program selection/self-selection biases, are based the assumption that the program has remained stable over time with respect to the types of customers who are targeted or who choose to participate, which is a function of a number of factors including the design, marketing and implementation of the program. If the program and the environment in which it operates are not stable from one year to the next, the claim that self-selection is controlled through this design is weakened considerably.

Rather than assume that this is the case, each program should be examined to assess its stability. For example, one could interview the program staff to assess the extent to which the mix of technologies that were promoted and the customers who were targeted changed from Cycle 1 to Cycle 2. One could also compare the two groups with respect to annual *and* seasonal energy use and a variety of demographic variables. If there are some important differences, where does an evaluator draw the line? How large do the differences in the design and delivery of the program and the demographic characteristics of participants have to be before an evaluator explores other designs?

[14] Computer simulation that is calibrated to some actual performance data for the system or facility being modeled. One example of computer simulation is DOE-2 analysis for buildings.

[15] Routine Adjustments: Any energy-governing factors, expected to change routinely during the reporting period, such as weather or production volume.

[16] Non-Routine Adjustments – for those energy-governing factors which are not usually expected to change, such as: the facility size, the design and operation of installed equipment, the number of weekly production shifts, or the type of occupants.

Finally, one could interview a sample of both groups to determine the extent to which they adopted any of the program-promoted measures and practices before participating. Again, at what point does an evaluator become concerned that the estimated savings are contaminated by pre-participation installations by the Cycle 2 participants such that the gross savings are migrating toward net?

It is relatively straightforward to establish the extent to which future participants adopted any of the program-promoted measures and practices before participating in single-measure programs, but is more difficult in multiple-measure programs such as whole-building programs. This is explored in more detail in Section 3.2 (net savings), but it is important to address this issue in estimating gross savings as well. The primary determinants of whether program-qualified measures will or will not have been done in that period are: 1. Whether the measure is energy-rated or an add-on, 2. How expensive or difficult to install it is (the more expense or needful of outside help it is, the more likely it would have triggered program participation earlier than it did occur, or was expensive enough that during the following year there would be insufficient costly work to be done under a whole-building program), and 3. Whether its function is essential such that it would have to be installed or replaced if prior versions have failed or functioned poorly (e.g., central air conditioners failing in summer).

Table 2 explores how likely it is that any given measure might have been installed by future participants in their pre-participation period, i.e. the evaluated period, based on these factors. To the extent that such measures were completed in what is now the pre-participation year of the future participants, the assumption that none of the program measures were installed during the comparison period is unjustified.

Table 2. Analysis of Measures and Measure Categories as They May Be Represented in Future-Participant Comparison Groups

Col 1 Measure Category	Col 2 Measure	Col 3 Will Measure in Col 2 Be Done by a Future Participants CG in Their Pre-Participation Period (aka Cycle 1)?
Measures done/not done,Ŧ inexpensive, possibly DIY—function not essential	Air Sealing	May be done
	Weatherization	May be done
Measures done/not done, and expensive, usually w/outside help—function not essential	Duct Sealing	Probably not done
	Duct Replacement	Probably not done
	Attic Insulation	Probably not done
	Wall Insulation	Probably not done
	Floor Insulation	Probably not done
	Duct Insulation	Probably not done
	Windows	Probably not done
	Radiant Barriers	Probably not done
	On-Demand Gas Water Heater	Probably not done
Measure done/not done, and expensive, usually with outside help, function essential	Roof	Probably not done
	On-Demand Gas Water Heater (i.e. heating is essential)	Probably not done
	Central Air Conditioner	Probably not done
Measures come in degrees of efficiency, relatively inexpensive, function essential	Central Gas Furnace	May be done
	Wall Heater	May be done
	Gas Storage Water Heater	May be done
	Room air conditioner	May be done
	Electric Storage Water Heater	May be done

Ŧ Done/Not Done refers to measures that are either installed or not, i.e. they don't vary in degree of efficiency, or if they do, an upgrade to higher efficiency is unlikely if a lower level efficiency version is already installed.

4.2.4 Limited Application in the Real World of Energy Efficiency Programs

Chapter 8 begins with a description of the conditions under which its recommendations apply. Perhaps the most limiting one is that it is focused only on retrofits; i.e. early replacement of equipment. Does this mean that *none* of the many measures supported by any whole-building program can be replacements of burned out or low-functioning equipment? If so, this would be extremely limiting. In a program with many qualifying measures, any one or several of the building measures could have failed. In addition, some such programs in the country may not require that any of the equipment be retrofits in order to qualify for the program. We recommend that a revised chapter 8 be written, if possible, to include consideration of at least partial retrofits. Absent that, UMP should consider commissioning another chapter that explores acceptable methods for addressing the installation of retrofit measures in whole-building programs.

4.3 Lack of Clarity and Inconsistencies

4.3.1 Gross vs Net in the Use of Future-Participant (Cohort) Designs

As mentioned earlier, the main advantage of this design is that selection biases, introduced by adding a comparison group, are reasonably well controlled assuming that *the composition of the current and future participants is similar and that the design and implementation of the intervention has not substantially changed over time.* It also assumes that the future participants were not exposed to the treatment in their pre-period. An example used by Campbell and Stanley (1963) is an officer and pilot training program, whereby participants in year one (cycle 1) are compared to participants in year two (cycle 2). The assumption is that training to be an officer is only available through the Army's training program. In other words, during the first program cycle, the future participants in the second program cycle (nor any soldier eligible to participate in the second program cycle) could not have been exposed to any training that would have prepared them to be an officer or pilot. Thus, a traditional design text would consider this design to produce the net effects of the officer and pilot training program rather than gross effects.

However, in a program such as California's Energy Upgrade California (EUC) Program, the future participants might not completely represent what the evaluated participants would have done absent the program. To estimate net savings, what would be needed is a comparison group that represents what the members of the *eligible nonparticipant population* would have done absent the program (the population like the cycle 1 participants). Future participants don't necessarily represent that; they do represent what Cycle 1 participants would have done absent the program, which is not necessarily the same thing. Consider these three scenarios:

Scenario 1. Cycle 2 participants did not engage in the behaviors promoted by the program during Cycle 1 but some members of the larger eligible nonparticipant population did engage in those behaviors during that period. In this scenario, the effects produced by analysis would be gross. That is, there are no natural adopters among the future participants and they represent exogenous changes only.

> Example: a whole building program in which Cycle 2 participants did not engage in the behaviors promoted by the program during Cycle 1 and are, therefore, capturing only the effects of Factors 1 through 7 listed in Section 3.

Scenario 2. Cycle 2 participants did not engage in the behaviors promoted by the program during Cycle 1 and the members of the eligible nonparticipant population also did not engage in those behaviors during Cycle 1, i.e. Cycle 2 participants had the same rate of relevant installations as the larger eligible population, which is none. Here, the future participants represent not only Factors 1-7 but also 8 since they also represent what members of the broader eligible population would have done absent the program. Therefore, the modeled effects are net.

> Example: a low-income program, in which future participants in Cycle 2 are unlikely to have installed any EE measures during Cycle 1 prior to participating in the program themselves. It is also very likely that members of the eligible nonparticipant population during Cycle 1 did not install any EE measures due to cost.

Scenario 3. Cycle 2 participants did, to some extent, install equipment promoted by the program during Cycle 1 but the members of the eligible nonparticipant population did so to a greater extent during Cycle 1. In this scenario, the estimated savings would be somewhere between net and gross. (i.e., there are more natural adopters among the eligible nonparticipant population than the among the Cycle 2 participants during Cycle 1).

> Example: a whole-building program in which future participants did, to some extent, engage in the behaviors promoted by the program during Cycle 1 and are, therefore, capturing, to some extent, the naturally-occurring rate of installing program-promoted equipment so that factors 1 through 7, and to some extent Factor 8, listed in Section 3, would be statistically controlled.

Under what conditions would a member of the eligible population engage in program-promoted behaviors? If a member of the eligible population had purchased quite expensive equipment during the pre-participation period, they might find it difficult to qualify for a whole-building program in the next year, and thus might not participate. So, this type of customer would be missed by a future-participants design. Under this condition, the use of future participants would produce an estimate of gross savings. On the other hand, some members of the eligible population might do various kinds of envelope sealing as a DIY project during one year, and learn about the program the next year, and find it easy to qualify for the whole-building program during the next year. So, this type of customer might well be captured as a member of the comparison group for the evaluated period producing a savings estimate that is somewhere between gross and net.

It is important that Chapter 8 clearly distinguish among these three scenarios and place the cohort design in the context of the traditional research design literature and how its use in the estimation of gross savings is based on a different set of assumptions that cannot be blindly accepted.

4.3.2 Lack of Clarity on What Constitutes the "Eligible Population" and the Pool of Potential Comparison Group Members for Net Savings Analysis

Our experience is that chapter 8 is sometimes confusing, and a part of this confusion is the lack of definition of the "eligible population" and on what characteristics they are defined. We agree that a matched comparison group represented in Row 4 of Table 1 in Chapter 8 yields savings that are between gross and net. But this raises the critical issue of what constitutes the eligible population from which the matched comparison group was selected (the same issue pertains to Row 5 as well). If one were interested in estimating gross savings, the ideal comparison group would be one that did not engage in any of the promoted behaviors and, therefore, represent only the exogenous effect due to non-program factors. However, thinking of home upgrade programs, a random sample of eligible nonparticipants would contain a mix of households that 1) did no home upgrades, 2) did standard home upgrades, 3) did some efficient home upgrades, or 4) did only efficient home upgrades. Some mix of groups 2 through 4 represent the naturally-occurring rate of doing efficient upgrades. To some extent then, the resulting savings would be between gross and net, but closer to net. On the other hand, if the comparison group were composed only of households that completed home upgrades of similar size (i.e., budgets), the estimated savings would be net. Of course, the same logic applies to non-residential building upgrades. These factors should be included and addressed explicitly in Chapter 8. What are the characteristics of the "eligible nonparticipant population?" And what types of customers should be included or excluded in order to produce the best estimates of gross or net savings? These are the questions that we suggest be addressed more clearly than they currently are.

4.3.3 Inconsistency in Describing Estimates as Partially Net, but Treating Them as Gross

Chapter 8 recognizes that some designs in Rows 4 and 5 will probably yield estimates that are somewhere between gross and net. The chapter goes on to recommend that these estimates be adjusted by a self-report NTGR while recognizing that this will produce a conservative estimate of net savings since there will be to some extent a double counting of free riders. Setting aside the question of why one should unquestionably settle for a conservative estimate, this recommendation fails to describe the errors associated with this approach relative to the errors associated with the other quasi-experimental designs discussed in Chapter 23 of the UMP. That is, what are the potential errors associated with the use of a traditional non-equivalent comparison group design compared to the potential errors associated with the self-report approach combined with the potential errors with the use of the two-stage approach, or the future-participant design to estimating the gross savings? A thorough discussion of these tradeoffs seems appropriate. Also, it would be useful to reference best-practice documents in the use of the self-report approach (Ridge et al., 2007; Ridge et al., 2013; Illinois Commerce Commission, 2016) to increase the chances that the reliability and validity of NTGRs are acceptable and consistent across jurisdictions.

4.4 A Larger Question

We think the descriptions and discussions of various designs and conditions under which they are appropriate demonstrate strongly that the issues are complex and are not subject to uniform approaches to evaluation. It is simply not practical to require that evaluators of all programs under all conditions adhere to a rank-ordered short list of evaluation designs and analytic approaches. Programs vary enormously in what they recommend, incent, budget, and target, among other things. The same programs change over time. Some have been around for decades, some are pilots, some are transitioning. These factors and more require flexibility in selecting and implementing evaluation designs. High-quality evaluations require the flexibility to take into account all of the common as well as unique factors that describe the evaluated program, and they should be required to describe these things in detail to support design choices. None of this lends itself to imposition of uniformity in evaluation design.

5. References

Agnew, K and M. Goldberg. (2009). *Getting to the Right Delta: Adjustment and Decomposition of Billing Analysis Results.* Presented at the International Energy Program Evaluation Conference.

Agnew, K. and M. Goldberg. (2013). Whole Building Retrofit with Consumption Data Analysis Evaluation Protocol: Chapter 8 of the Uniform Methods Project, National Renewable Energy Laboratory.

Campbell, D.T. and J.C. Stanley. (1963). *Experimental and Quasi-Experimental Designs for Research*. Chicago, Ill.: Rand McNally College Publishing Company.

Cook, T.; M. Scriven; C.L. Coryn; and S.D.H. Evergreen (2010). “Contemporary Thinking About Causation in Evaluation.” American Journal of Evaluation 31:105. http://aje.sagepub.com/content/31/1/105

Evergreen Economics. (2013). *Needs Assessment for the Energy Savings Assistance and the California Alternate Rates for Energy Programs: Volume 1: Summary Report*. Prepared for: Southern California Edison, Pacific Gas and Electric, Southern California Gas, San Diego Gas and Electric and the California Public Utilities Commission.

Granderson, J. (2015). “Accuracy of Existing Use Baselines, AKA Normalized Metered Energy Use.” (a presentation on Implementation of AB802, California Public Utilities Commission, January 26-27, 2016).

Goldberg, M. and K. Train (1995). Net Savings Estimation: An Analysis of Regression and Discrete Choice Approaches. Submitted to California Demand Side Management Advisory Committee, Subcommittee on Base Efficiency.

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-experiments, and Single-Case Studies*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Illinois Commerce Commission. (2016). *Illinois Statewide Technical Reference Manual for Energy Efficiency Version 5.0: Volume 4: Cross-Cutting Measures and Attachments: Attachment A: IL-NTG Methodologies*. Prepared for the Illinois Commerce Commission by the Illinois Evaluation Teams (ADM Associates, Cadmus Group, Itron, Navigant Consulting, Opinion Dynamics, and Ridge & Associates) and the Illinois Stakeholder Advisory Group.

Kennedy, P. (2008). *A Guide to Econometrics*. Malden, MA: Blackwell Publishing.

Mohr, L. B. (1995). Impact Analysis for Program Evaluation. Thousand Oaks, CA: SAGE Publications.

Ridge, R., K. Keating, L. Megdal, and N. Hall. (2007). *Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach*. Prepared for the California Public Utilities Commission.

Ridge, R., N. Hall, R. Prahl, G. Peach, and P. Horowitz. (2013). “Guidelines for Estimating Net-To-Gross Ratios Using the Self Report Approach.” Prepared for the New York Department of Public Service.

Rossi, P., H., M. W. Lipsey, and, H.E. Freeman (2004). *Evaluation: A Systematic Approach*. Thousand Oaks, CA: Sage Publications.

Rubin, D. (1974). Estimating Causal Effects of treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, Vol. 56, 5, 488-701.

Shadish, W. R., T. D. Cook, and D. T. Campbell. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Company.

The TecMarket Works Team. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Directed by the CPUC’s Energy Division, and with guidance from Joint Staff.

Train, K. (1993). Estimation of Net Savings from Energy Efficiency Programs. Submitted to Southern California Edison Company.

Train, K., M. Goldberg and K. Agnew (2017) *Compared to What? Practical Tools for Consumption Data Analysis Mitigating Self-Selection Bias*. Submitted to Pacific Gas and Electric Company.

Violette, D. and P. Rathbun. (2014). Estimating Net Savings: Common Practices, Chapter 23 of the Uniform Methods Project, National Renewable Energy Laboratory.

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: The MIT Press.

For more information, please contact:

Katherine V. Randazzo, Ph.D.
Principal Data Scientist

510.214.0194 tel
krandazzo@opiniondynamics.com

Opinion **Dynamics**

Boston | Headquarters

617 492 1400 tel
617 497 7944 fax
800 966 1254 toll free

1000 Winter St
Waltham, MA 02451

San Francisco Bay

510 444 5050 tel
510 444 5222 fax

1999 Harrison St
Suite 1420
Oakland, CA 94612

Madison, WI

608 819 8828 tel
608 819 8825 fax

2979 Triverton Pike
Suite 102
Fitchburg, WI 53711

Orem, UT

510 444 5050 tel
510 444 5222 fax

206 North Orem Blvd
Orem, UT 84057