THIRD DRAFT (11-4-2013)– NOT TO BE CITED

Notes to Editor:

- Need to Add Appendix A to the TOC
- Since the TOC contains references, shouldn't this be placed before the bibliography?
- Need to Add Placeholder for the Acknowledgments
- Appendix A is entirely new and needs a read through

Estimating Net Energy Savings

Daniel M. Violette, Ph.D., Navigant Consulting, Inc.
Pam Rathbun, Tetra Tech

Table of Contents

1 Universality of the Net Impacts Challenge 3

2 Defining Gross and Net Savings for Practical Evaluation 4

2.1 Definition of Gross and Net Savings 4

2.2 Definitions of Factors Used in Net Savings Calculations 4

2.3 Uses of Net Savings Estimates in the EE Industry 8

2.4 The Net Savings Estimation Challenge – Establishing the Baseline 9

3 Methods for Net Savings Estimation 10

3.1 Randomized Controlled Trials and Quasi-Experimental Designs 12

3.1.1 Randomized Control Trials 12

3.1.2 Quasi-Experimental Designs 16

3.2.1 Program Participant Surveys 23

3.2.2 Surveys of Program Nonparticipants 28

3.2.3 Market Actor Surveys 28

3.2.4 Case Studies for Estimating Net Savings Using Survey Approaches 29

3.3 Common Practice Baseline Approaches 35

3.4 Market Sales Data Analyses (Cross-Sectional Studies) 39

3.5 Top-Down Evaluations (Macroeconomic Models) 41

3.6 Structured Expert Judgment Approaches 47

3.7 Deemed or Stipulated NTG Ratios 49

3.8 Historical Tracing (or Case Study) Method 50

4 Conclusions and Recommendations 51

4.1 A Layered Evaluation Approach 51

4.2 Selecting the Primary Estimation Method 53

4.3 Methods Applicable for Different Conditions 55

4.4 Planning Net Savings Evaluations – Issues to be Considered 57

4.5 Trends and Recommendations in Estimating Net Savings 58

ESTIMATING NET ENERGY SAVINGS

This chapter focuses on methods that can be used to estimate net savings in evaluation, measurement, and verification (EM&V) studies for energy efficiency program evaluations and, where appropriate, how the conceptual view of net savings can influence the choice of methods. In this context, the purpose of evaluation is to provide decision-makers[1] with the information needed to make good investment decisions in energy efficiency (EE).

Estimating net savings is central to many EE evaluation efforts, and is broad in scope with its focus on defining baselines and savings levels. The intent of this chapter is not to prescribe the use of specific methods; instead, it is meant to describe methods used in EE evaluation and their tradeoffs such that each jurisdiction can make good decisions about the net savings methods to be used.

This chapter summarizes the rationale for net savings estimation, and provides guidelines and issues to be aware of when selecting a particular method. A detailed description of the design, sampling, implementation, and analysis of each of the methods is beyond the scope of this chapter. Readers are encouraged to consult articles cited in References, as well as other sources that cover the specific methods in greater depth.

1 Universality of the Net Impacts Challenge

Any investment decision results in allocating resources to achieve an objective. Regardless of the investment, once this investment has been made, it is generally impossible to undo the resource allocation to see what would have happened in the absence of this investment. What would have happened in the absence of the investment is termed the "counterfactual" scenario.

A review of publications in journals and books dedicated to evaluation illustrates that virtually every issue arising from the estimation of net impacts of EE investments has a parallel in the evaluation of other investments made in either the private or public sector. Examples include:

- Healthcare: What would have been the effects on health if, for example, an investment in the fluoridation of water had not been made?
- Tax subsidies for economic development: Would the project—or a variant of the project—have been built anyway?
- Education subsidies: What would have been the effects if investments in school lunch programs or low-interest loans for higher education had not been made?
- Military expenditures: What would have happened if an investment in a specific military technology had not been made?

[1] Decision-makers that influence EE investments include regulators, utilities, program administrators, legislators and even the implementation contractors that may do much of the field work in program delivery. The specific audience for the evaluation effort can influence the methods used, the aspects of the evaluation that are emphasized in presentation, and the depth of the presentation of the work.

The challenge faced by EE practitioners is how to approximate the appropriate counterfactual scenario. For EE programs, the counterfactual may include the assumption that some participants in the program would have installed some of the EE measures promoted by the program, even if the program had not existed.

2 Defining Gross and Net Savings for Practical Evaluation

This section discusses the estimation of net savings as an assessment of attribution.[2] It defines key terms related to estimating net savings, summarizes the different uses of net savings measurement in the industry, and describes how many of the issues in the appropriate estimation of net savings are tied to the development of an appropriate baseline.

2.1 Definition of Gross and Net Savings

The Uniform Methods Project (Haeri, 2013) provides the following definitions of gross and net savings.

> ***Gross Savings:*** Changes in energy consumption that result directly from program-related actions taken by participants in an energy efficiency program, regardless of why they participated.
>
> ***Net Savings:*** Changes in energy use attributable to a particular energy efficiency program. These changes may implicitly or explicitly include the effects of factors such as freeridership, participant and nonparticipant spillover, and induced market effects.

A term that has become almost synonymous with estimating net savings is the net-to-gross (NTG) ratio. A NTG ratio is commonly defined as the ratio of net and gross savings. Therefore, the NTG ratio (a NTG ratio obtained from a survey method, for example) is multiplied by the gross savings to estimate net savings.

2.2 Definitions of Factors Used in Net Savings Calculations

The factors most often used in calculating net savings are: freeridership, spillover (both participant and nonparticipant), and market effects. These factors are consistent with those contained in the Energy Efficiency Program Impact Evaluation Guide (SEE Action, 2012b).

> ***Freeridership (FR)***: FR is the percentage of program savings attributed to freeriders, or those program participants who would have implemented the program's measure or practice in the absence of the program. It is important to recognize that there are different levels of FR that should be measured:
>
> - "Total" freeriders would have completely replicated the program measure(s) or practice(s) on their own and at the same time, in the absence of the program.

[2] This view of net savings as a problem of attribution has also been viewed as assessing causality. The term attribution is used in this chapter as it is more descriptive of the problem being addressed, where causality can have a wider range of interpretations in the literature that even extends to metaphysical constructs. See discussion in Cook T. et al. (2010) on Causality in Contemporary Evaluation.

- "Partial" freeriders would have partially replicated the program measure(s) (i.e., quantity or efficiency level).
- "Deferred freeriders" would have completely or partially replicated the program measure(s) or practice, but at a future time beyond the program's time frame.

Spillover (SO): SO refers to additional reductions in energy consumption and/or demand due to program influences beyond the savings directly tied to program participation, without financial or technical assistance offered through the program. There are generally two types of spillover:

- "Participant SO" represents the additional energy savings that occur when a program participant—as a result of the program's influence—installs EE measures or practices *outside* of the program after having participated in the efficiency program.
- Other evaluators have added definition to the broad category of participant SO. Commonly recognized sub-categories of participant spillover are:
    1. Inside spillover (ISO) occurs when additional program-induced actions are taken at the participating site; and
    2. Outside spillover (OSO) occurs when an actor participating in the program initiates actions that reduce energy use at other sites that are not participating in the program
    3. Like spillover (like SO) refers to program-induced actions of the same type as installed through the program
    4. Unlike spillover (unlike SO) refers to energy efficiency actions unlike those installed through the program.
- "Nonparticipant SO" represents the energy savings that occur when a program nonparticipant implements EE measures or practices as a result of the program's influence (i.e., through exposure to the program) but outside of the program accounting.

Market effects (ME): ME refers to "a change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy efficiency products, services, or practices and is causally related to market intervention(s)" (Eto et al., 1996). For example, programs can influence: design professionals, vendors, and the market (through product availability, practices, and prices); product or practice acceptance; customer expectations; and, other market-related effects. All of these elements may induce consumers to adopt EE measures or actions. As a result, an evaluation might conclude that some participants are freeriders in the current-year evaluation when, in fact, they could not or may not have adopted the EE measure, practice or service because it would not have existed in the marketplace or been available at the same price had it not

> been for prior utility programs.[3] These freeriders may represent savings that have resulted from programs in prior years due to market effects and are not usually included as attributable impacts in the evaluation of a current-year program; however, it is important to recognize these on-going effects may not have been counted in prior years' evaluations and this should be considered in developing the policy context for viewing the evaluation results from current-year programs.

There is debate around the difference between SO and ME. Prahl et al. (2013) suggest that market effects may be "best viewed as SO savings that reflect significant program-induced changes in the structure or functioning of energy efficiency markets."[4,5] While SO and ME are related, the methods to quantify these two factors generally differ, and they are addressed separately in this chapter for that reason.

Different factors to estimate net savings are used in various programs and jurisdictions, depending on how equity and responsibility are viewed within a jurisdiction (NMR et al., 2010). For example, some jurisdictions allow only FR to be included in the calculation of net savings, while others allow for individual estimates of both FR and SO, and still others allow net savings estimates where FR and SO are not individually measured (e.g., market-level estimates of net savings).[6]

From the viewpoint of a practitioner who is trying to develop methods for estimating values for these factors, the definitions provided above are useful. However, the evaluator must work with the information available and that starts with the tracking system.[7] The data in the tracking system are generally viewed as the initial estimate of gross savings. FR, SO, and ME are untracked values that must be estimated or accounted for outside of the tracking system.[8] A

---

3 Sebold et al. (2001) sets out an expansive framework for assessing EE programs.

4 Prahl et al. (2013) also suggest that market transformation is a subset of market effects, i.e., those effects that are substantive and long lasting. This view implies that market effects are a subset of SO, and market transformation is a subset of market effects.

5 Reviewers have indicated that some frameworks view spillover as a subset of ME with ME including participant and nonparticipant spillover that is long lasting. This is in addition to program-induced market effects. The terminology does vary across some applications, but the estimation issues are the same.

6 Differences across jurisdictions in the reporting of gross and net savings is also discussed in Messenger et al. (2010, p. 19-21).

7 The definitions for FR, SO, and ME must be integrated with both how the utility tracks actual data on program participation and how information on expected program impacts is recorded in the program tracking system. In general, the initial estimate of gross savings comes from the tracking system in terms of expected energy savings by participant by program or measure. Some of these data may include "deemed values" negotiated by the stakeholders. In some cases, these deemed values may include factors that lower the savings of a measure, based on assessments of current practice, codes and standards, and other factors that may directly or indirectly influence how the gross savings estimated in the tracking system are adjusted to estimates of net savings. As a result, it is important to understand how the gross savings are estimated by project and by participant. In fact, the first recommendation of NMR/Research into Action (2010) is that a process leading to the development of a consistent definition of *adjusted gross savings* is needed for the Northeast Region.

8 There are direct estimation methods that can be argued to address FR, SO, and ME without the need to estimate each separately. See sections in this chapter on randomized control trials, quasi-experimental designs, and

practical way to view the estimation problem is to consider SO and ME as savings that are attributable to the program, but not included in the program tracking system, whereas FR are savings that are included in the program tracking system but not attributable to the program.

Estimating these values and then making appropriate adjustments to the values in the tracking database (or validated tracking database[9]) results in an estimate of net savings.[10]

> ***Net Savings*** = Gross Savings – FR + SO + ME (that is not already captured in the SO value) eq. 1

In much of the literature, the approach involves a NTG ratio for which FR, SO, and ME are expressed as a ratio to gross savings. The NTG construct is widely used, as it fits easily into some types of evaluation efforts, such as survey-based estimations.

> ***NTGR*** = 1 – FR ratio + SO ratio + ME ratio (where the denominator in each ratio is gross savings) eq. 2

When using the NTGR defined by the specific FR, SO and ME factors (or ratios), net savings is calculated as:

> ***Net Savings*** = NTGR * Gross Savings eq. 3

These definitions are essentially standard in the evaluation literature.[11] However, a given jurisdiction may decide not to include one of the net savings factors in the estimation of net savings. For example, while FR is almost always included, SO and ME may not fully be considered (see NMR et al., 2010 and NEEP, 2012). This is a policy choice made by that jurisdiction. Most agree that SO and ME exist and are positive values, but it can be difficult to determine the magnitudes of these factors. Increasingly, the trend has been to include estimates

---

common practice baselines. However, each method can be viewed as an approach that essentially adjusts the savings estimates in the program tracking system.

9 A *validated tracking database* is simply a review of the program tracking database. Programs that are equipment based, either rebate or custom design, have a program tracking database that contains an estimate of the savings expected to be achieved by that installation. A review of this tracking data base is generally made to determine if there are obvious errors, if adjustments can be made to make the ex-ante savings entries more accurate, and to determine if any deemed savings values may already include adjustments that account for net savings factors (e.g., an adjusted baseline that captures market trends). The validated tracking system then contains the best information on ex ante savings for each participating site or project. The benefits of improved information in the tracking system are discussed in Violette et al. (1993).

10 Keating (2009) discusses issues concerning the math underlying net savings and NTG calculations with examples of an inappropriate multiplication algorithm using FR and SO expressed as ratios.

11 There are other factors that are generally considered as adjustments to gross impact estimates that are sometimes addressed as net-impact factors. These include rebound, snapback, and persistence of savings. These factors are addressed in UMP Chapter 13 – Persistence and Other Evaluation Issues (Violette, 2012). As is the case with other NTG factors, these factors are not always treated consistently as being components internal to gross impact calculation, or as factors that are used to adjust program gross impacts to get to a final net impacts number.

of SO in net savings evaluations; including ME is also increasing. Methods are available to address both SO and ME and, since there is really no debate about whether they exist, these factors should be addressed when estimating net savings. The SO and ME estimates may have some uncertainty around them, but no more so than what is found in the broader evaluation literature from other fields. It is important to know if these values are potentially large or small for a given program or portfolio. It then becomes a policy decision on how to use this information in making decisions about EE investments.

2.3 Uses of Net Savings Estimates in the EE Industry

There have been discussions in many regulatory jurisdictions regarding the appropriate use of net savings estimates. This is due in part to factors such as: (1) the cost of the studies to produce these estimates, and (2) a perceived lack of confidence in the resulting estimates. However, evaluators and regulators recognize the advantages of consistent measurement of net savings over time as key metrics for program performance (Fagan et al., 2009).

There are at least five uses of net savings that evaluators generally agree upon (SEE Action, 2012b):

- Using net savings, prospectively, for program planning and design (e.g., for setting consumer incentive levels).
- Assessing the degree to which programs cause a reduction in energy use and demand—with net savings as one of numerous measures that should be given serious consideration in the assessment of program success.
- Obtaining insight into how the market is changing and transforming over time by tracking net savings across program years and determining the extent to which freeridership and spillover rates have changed during the period—and potentially using this insight for defining when, and how, to implement a program exit strategy.
- Gaining a better understanding of how the market responds to the program and how to use the information to inform modifications to program design, including how to define eligibility and target marketing.
- Informing resource supply and procurement plans which need to understand the relationship between efficiency levels imbedded in reference load forecasts and net reductions from programs.

Schiller (SEE Action, 2012b, p. 2-5) also discusses the objectives and importance of evaluation. Goals and targets for EE program performance can use estimates of savings (gross or net) that can be used to assess EE program performance. In some cases, jurisdictions link financial

incentives to the meeting of EE goals. Arguments have been made for either the use of gross savings and/or net savings as appropriate targets for measuring program success.[12]

2.4 The Net Savings Estimation Challenge – Establishing the Baseline

All of the methods discussed later in this chapter for estimating net impacts from EE programs are predicated on the development of a baseline (the assumed counterfactual) from which impacts are measured. To be able to understand and defend the selection of a particular method for estimating net savings requires that the evaluator pay attention to the implicit and explicit assumptions around the baseline comparison group that is used. If a perfect baseline could be identified, most of the issues associated with estimating net impacts would not occur.

The challenge facing the evaluator is to identify a method that produces a baseline that represents the counterfactual as accurately as possible, i.e., what the participant group would have done in the absence of the program.[13] The evaluator's effort should take into account the issues that pertain to the similarity, or matching, of the participant and nonparticipant comparison groups. It should also take into account any effects the program might have had on the comparison group (that is, any interactions between the program and the comparison group that may impact the net savings calculation). In addition to baseline estimation methodology issues described in more detail in the next section, self-selection bias, FR and SO can create baseline estimation concerns.[14]

---

[12] As more jurisdictions begin to consider the delivery of EE programs as a business process that requires an investment of resources, returns on this investment are being considered. These returns on investment (more commonly termed incentives) are typically tied to performance targets. The target could be based on reaching a certain level of gross savings or it could be based on achieving a certain level of net savings. Each has pros and cons. A gross savings target for incentives can provide a clearer incentive to the program administrator, and there generally is less controversy over whether the target has been achieved. The fact that incentives are usually based on a calculation of shared benefits that has the predominant share of benefits going to ratepayers creates an equitable incentive structure through the selection of the amount of shared benefits that goes to the program administrator, and the fact that even if attributed savings (i.e., net savings) are less than expected, ratepayers still receive the majority of the benefits. For example, under an 80-20 split of the benefits (80 percent of benefits are realized by ratepayers and 20 percent are realized by the administrator), having attributed savings reduced by 50 percent still implies that 70 percent of the benefits goes to ratepayers. Other views of aligning incentives with the outputs of program evaluation can be found in Rufo (2009).

[13] Chapter 8, "Whole-Building Retrofit with Consumption Data Analysis", of the DOE Uniform Methods Project (Agnew et. al., 2013) provides a number of choices regarding the selection of control groups that are options for use in billing analyses (i.e., comparing changes in energy use for both participants and a control group). Regression analysis is discussed in Chapter 8 as a tool for making the appropriate comparison and arriving at alternative net savings values.

[14] Self-selection, FR and SO issues are common in other settings as well. Consider a business decision made to produce net benefits, such as down-sizing. Might self-selection be important to address in designing this business initiative? If there is a choice to take an early retirement package, might those employees who are most confident and have the best chance to land new jobs take the package (a freerider concept)? Might there also be SO impacts from the down-sizing program where employees that remain on the job have their productivity impacted by having co-workers leave? These issues pose challenges for EE evaluation, but they are also part of assessing many other investment and business decisions.

Self-selection bias arises when a program is voluntary and participants select themselves into the program. The fact that some consumers selected themselves into the program suggests the potential for systematic differences between program participants and nonparticipants. This issue is not unique to EE evaluations and is present in any assessment of a policy or program where people select themselves into the program. FR is one specific variant of self-selection bias. This is a baseline issue when the actions of the comparison/control group do not accurately reflect the actions that would have been undertaken by participants in the absence of the program. Specifically, the assumption is that program participants that have self-selected themselves into the program are those customers that would have undertaken more conservation actions than are observed in a general nonparticipant comparison group.[15] While FR reduces net program savings, there are other variants of self-selection bias that might increase net savings. For example, if the customers that self-select themselves into the program are those that need the financial incentives to justify the EE investment, then, an adjustment for self-selection might increase overall net savings. The fact that participants self-selected themselves into the program does not dictate that net savings are over or under estimated.

SO is another baseline issue. Nonparticipant SO can occur, for example, when the energy consumption of the comparison group used in the analysis is not indicative of what the energy consumption of group members would have been had the program not been offered. In this case, the problem can be viewed as a *contaminated* comparison group--in other words, the behavior of the comparison group was affected by the existence of the program.

The issues discussed above all deal with establishing an appropriate baseline as an approximation of the counterfactual. Understanding that FR, SO and even ME are baseline issues can help focus the net savings estimation problem. For many, the selection of the baseline to be used in the evaluation is a core issue in having confidence in the methods used. The presentation of results from a net savings estimation method should include a description of the baseline implicit in the use of that method. Establishing baselines and related issues are discussed in more depth in Section 3 of this chapter.

3 Methods for Net Savings Estimation

This section contains a discussion of different methods for estimating net savings along with some of the key advantages and challenges associated with each method. There are a variety of

[15] In this context, freeriders can be viewed as a subset of the self-selection problem. However, freeridership is not the only possible explanation of self-selection hypothesis. There may be other factors influencing consumers to participate in a program that could result in the participant groups and nonparticipant groups behaving differently. For example, if the participants in the program are those consumers that really need the rebate to make the investment and those that choose not to participate are more well-off consumers that do not need the rebate to take EE actions, then the baseline comparison group will take more EE actions than the participant group would take. The result is an estimate of savings that is too low, rather than an estimate of program savings that is too high under the commonly assumed freerider self-selection hypothesis. Developing a better comparison group in this case would correct for the self-selection bias and increase the estimated program savings.

methods that can be used, and one method may be used to estimate FR or SO (e.g., survey-based self-reports), while different methods are used to estimate ME (e.g., structured judgment approaches or historical tracing). Methods addressed in this section include[16]:

- Randomized control trials and quasi-experimental designs
- Survey-based approaches
- Common practice baseline approaches
- Market sales data analyses
- Top-down evaluations (or macroeconomic models)
- Structured expert judgment approaches
- Deemed or stipulated NTG ratios
- Historical tracing (or case study) method

Table 1 below summarizes the applicability of each of these methods for estimating FR, SO, and ME. This is meant only to indicate the general applicability of the approaches; specific application, caveats, limitations, and other key information must be read in detail in the following sections to fully understand how each of these net savings components can be assessed.

Table 1: Applicability of approaches for estimating net savings factors

Method	FR	SO	ME
Randomized controlled trials and quasi-experimental designs	Controls for freeriders[17]	Controls for participant SO[18]	Not generally used
Survey-based approaches	Is applicable	Is applicable	In conjunction with structured expert judgment
Common practice baseline methods	Is applicable	Is applicable	Not applicable
Market sales data analysis	Is applicable	Is applicable	Is applicable
Top-down evaluations	Assesses the overall change in energy use and, therefore, there is no need to adjust for FR, SO, and ME		
Structured expert judgment[19]	Is applicable	Is applicable	Is applicable
Deemed or stipulated NTG ratios	Is applicable	Is applicable	Not generally used

[16] Some reviewers suggested that price elasticity studies discussion be included within this chapter. This approach examines how consumers respond to reductions in price for an energy efficient product. To date, these studies have examined programs that lower the costs of lighting products (e.g., CFLs) but have not expanded to other EE products. Please see Appendix A for discussion on this methodology.

[17] Does not estimate FR, but rather controls for freeriders through experimental design.

[18] Does not estimate SO, but rather controls for *participant* SO through experimental design; a separate study of control group members is required to address *nonparticipant* SO where there is reason to believe that it might be significant as the net impacts will be underestimated.

[19] Ensure experts are knowledgeable about specific market being studied.

Historical tracing	Is applicable	Is applicable	Is applicable

3.1 Randomized Controlled Trials and Quasi-Experimental Designs

This section discusses two methods for selecting a baseline against which program impacts are compared: randomized control trials (RCT) and quasi-experimental approaches. These approaches are increasingly being used to evaluate behavioral programs, information programs, and pricing programs designed to increase efficiency.[20] These programs have a large number of participants and these methods are typically focused on the residential sector.

3.1.1 Randomized Control Trials

A randomized control trial design (RCT) is viewed as the ideal method for assessing the net impacts of a program--particularly the FR and short-term SO components. If the RCT is short-term (i.e., one-year), then longer term SO and ME may not be captured. First, the study population is defined and then consumers from this study population are randomly assigned to either a treatment group (participants in the EE program) or to a control group that does not receive the treatment (nonparticipants).

Random assignment is a key feature of this method. By using random probability to assign consumers to either the program group or the control group, the influence of observable differences (location, age of home, stock of appliance, etc.) between the two groups is eliminated. Also eliminated are those differences often considered as unobservable (attitudes toward energy use, expectations about future energy prices, expertise of household members in areas that might help induce participation, etc.).[21] As a result, this method, when implemented properly, can provide a near-perfect baseline that will result in the most reliable estimates of net savings.

The net savings calculations are relatively straightforward when randomized control design is done properly. Three methods for performing the net savings calculation are generally seen in the literature:

1. Use a simple post-period comparison to determine the difference in energy use between the control group and the participant group after participating in the program. For example, if participating households are using 15,000 kWh on average and the control group is using 17,000 kWh, then the net savings estimate is 2,000 kWh.
2. Use a difference-in-differences (DiD) approach, comparing the change in energy use for the two groups between pre- and post-participation periods. For example, assume

[20] The SEE Action (2012a) report focused on information and behavioral programs was authored for the Customer Information and Behavior Working Group and the Evaluation, Measurement, and Verification Working Group. More information can be found at www.seeaction.energy.gov.

[21] References addressing the RCT and quasi-experimental designs include NMR Group, Inc. and Research into Action (2010) and two reports by SEE Action (2012a, 2012b). The SEE Action reports can be downloaded at: http://www1.eere.energy.gov/seeaction/index.html

participants used 17,500 kWh prior to program participation and 15,000 after participation, for a difference of 2,500 kWh between the pre- and post-periods. Assume also that the well-matched control group has similar pre-period energy use (approximately 17,500 kWh), but the group's post-period energy use is 17,000 (that is, slightly less, possibly due to weather), so the difference is 500 kWh. Applying the DiD method results in an estimated savings of 2,000 kWh (the 2,500 kWh change for participants minus the 500 kWh change in the control group).

3. Use a simple regression approach -- termed a linear fixed-effects regression (LFER) -- where the regression model identifies the effect of the program by comparing pre- and post-program billing data for the participants to the billing data for the control group. A key feature of the LFER analysis is the addition of a term--customer-specific fixed effect--which captures all customer-specific effects on electricity usage that do not change over time, including those that are unobservable. Examples of these fixed effects include the square footage of a residence, the number of occupants, and thermostat settings.[22, 23]

If the evaluator randomized the participant and control groups, why would the evaluator use any method other than the simple post-period comparison? The practical answer is that evaluators, in an effort to be as thorough as possible, use all available data to develop the estimate. Thus, using the DiD method helps track trends over time, and using the fixed-effects component of the LFER adds another control for differences between consumers that are constant during the period being examined. In theory, all three methods should produce the same result as these approaches are all based on the assumption that randomization has fully accounted for the differences between the participant group and the control group that influence the estimate of net savings from the program.

The RCT approach is simple in concept, but it may not be as simple to implement in the evaluator's world of messy data, timing, and program design issues. For example, suppose an evaluator selects the study population and performs the random assignment for a Home Energy Reports (HERs) program in which energy use reports are delivered by mail. This program is designed to generate energy savings by providing residential consumers with information about their energy use and energy conservation. The reports give consumers various types of information, including: a) how their recent energy use compares to their energy use in the past;

---

[22] See Provencher, B. and G. Glinsmann (2013) for an example and additional discussion of the LFER method.

[23] A number of the methods discussed in this chapter use regression approaches; some of these are more simplistic while others are quite sophisticated requiring expertise in econometrics. Citations to applied studies are provided in each section and many include descriptions of the econometric techniques employed. For example, Stuart (2010) lists econometric software and routines that can be useful in matching. Also, Chapter 8 of the DOE UMP (Agnew et al., 2013) discusses regression models in more detail, but provides a limited set of references to the literature. SEE Action (2012a) recommends Econometric Analysis by Greene (2011) as a useful reference on regression techniques. Cross-section and panel data models that are often used in evaluation are the focus of Wooldridge (2002). The Guide to Econometrics by P. Kennedy (2008) is a useful supplement to any econometrics text book. Also recommended is the book "Mostly Harmless Econometrics: An Empiricist's Companion" by Angrist and Pischke (2008).

b) tips on how to reduce energy consumption, some of which are tailored to each consumer's circumstances; and c) information on how their energy use compares to that of neighbors with similar homes. The evaluator does not have to contact the consumers assigned to the HERs control group but, when contacting members of the HERs participant group, the evaluator discovers that several thousand program households have addresses to which their mail is not delivered, so these households are dropped from the program. If those consumers are dropped from the participant group, should the consumers with undeliverable mail addresses also be dropped from the control group? This situation occurred in a randomized trial of a behavioral program (see Provencher and Glinsmann, 2013), and the samples were adjusted and tested for conformance with RCT assumptions.

It is becoming standard practice for evaluators to test the likelihood that the program groups and control groups have been appropriately randomized. Statistical methods can be used to test whether the data in the two groups are consistent with random assignment of consumers to the treatment group and control group. This type of analysis involves comparing the means of the two groups with respect to demographic variables (if available) and monthly energy use in the pre-program year. One test to see if the assumptions of an RCT hold is that the difference in mean consumption between the treatment and the control groups should not fall outside a 90% confidence bound for more than two months of the pre-program year. If mean consumption does fall outside a 90% confidence bound for more than two months, it does not prove that random assignment was *not* conducted, but it does provide a signal that the process used to perform the random assignment may need to be reviewed. [24]

To maintain a RCT over a period of time, care is needed when working with the data across the treatment and control groups. For example, a behavioral program (such as HERs) may be rolled out to 20,000 high-use residential consumers in Program Year 1. In Program Year 2, an additional 20,000 consumers of all energy use classifications may be enrolled, and another 30,000 consumers may be enrolled in Program Year 3. Additionally, some consumers in Program Year 1 may have dropped out (requested to not receive the home energy reports).

Inevitably, there are also issues with the consumer energy use data. Criteria researchers have used as indicators of problems with billing data for a consumer include:

- Having <u>less</u> than 11 or <u>more</u> than 13 bills during a program year
- Having <u>less</u> than 11 or <u>more</u> than 13 bills during the pre-program year
- Energy consumption outside a reasonable range, i.e., an outlier observation defined as observations with average daily usage less than the 1st percentile or greater than the 99th percentile
- Observations with less than 20 or more than 40 days in the billing cycle

[24] An example of an application of this test for consistency with RCT expectations can be found in Provencher and Glinsmann (2013) and other tests can be found in Stuart (2010).

Even programs that have been operating for several years are likely to have issues. These include consumers for whom there is no record of when the first report was sent and entries in customer records that raise questions (such as a note "do not include this customer in the analysis").

After addressing data issues, does the evaluator still have a good RCT? Probably yes, unless there are a large number of consumers who are affected or there are consumers disproportionately affected across the participant and control groups. [25]

The ability to disseminate information to large groups of consumers has led to an increase in RCTs in EE evaluation.[26] In general, these RCT-based evaluations have focused on residential behavior-based efficiency programs such as HERs programs. These programs lend themselves to random trials in that they: (1) provide information only, (2) can be implemented for a large number of consumers at essentially the same time, and (3) allow for a study population to be selected from which both a participant group and a control group can be randomized. These characteristics, however, are not generally present for many large-scale EE programs that tend to comprise much of the EE portfolio savings.

In summary, the RCT approach is generally viewed as the most accurate method for estimating net impacts. The RCT controls for freeriders and participant SO, which are two important factors. To the extent that the control group is affected by the program, nonparticipant SO is not addressed; however, this effect is likely to be small in many behavioral programs.

If nonparticipant SO is large, then net impacts will be underestimated as these program-induced savings will be treated incorrectly as freeriders in the net savings calculation. A separate study of control group members would be needed to address nonparticipant SO where there is reason to believe that it might be significant. Since market effects are longer-term SO effects, it is unlikely that any market effects are included in any RCT net savings approach that spans just a few years.

Although the RCT method is able to produce an accurate baseline when constructed correctly, it may not be possible to apply an RCT to evaluations of EE programs for a variety of reasons. RCT generally requires planning in advance of program implementation. As pointed out in Chapter 8 (Agnew and Goldberg, 2013) of these protocols, "… evaluation concerns have been less likely to drive program planning." Also, an RCT approach may involve denying or delaying participation for a subset of the eligible and willing population. In some cases, the random

---

[25] Chapter 9, "Metering", of the DOE UMP (Mort, 2013) presents additional criteria that can result in the exclusion of sites (see p. 26-27) and suggestions on what to do if the number of removed sites becomes large (i.e., greater than 5 percent).

[26] Some evaluations of HERs programs that used RCT include: AEP (2012), SMUD (2011), and ComED (2012). Some of these studies actually compare the RCT design to procedures which match participants to nonparticipants. Another useful study, but one focused on evaluating pricing programs, which used an RCT design is SMUD (2013). This study assesses different pricing structures in the residential sector; however, the methods used are good examples of what can also be applied in EE evaluations in an RCT context.

assignment may result in providing services to consumers who either do not want them or may not use them.

Other characteristics of programs that can make an RCT difficult to implement include:

- Programs that require significant investments, such as a Commercial/Industrial (C/I) major retrofit program in which the expenditures may be in the tens of thousands of dollars, typically are opt-in and random assignment within an eligible study population might include consumers who either do not need the equipment or services, or who do not want to make that investment. Programs that involve relatively large investments in measures and services across both residential and C/I sectors are clearly not amenable to a randomized trial design.
- Participants in some C/I programs may be relatively unique, with few similar consumers that might be candidates for a control group.
- To achieve savings targets, a program must be rolled out over an entire year, with consumers opting in every month. As a result, the participant group is self-selected, and is not known until after one year of program implementation.

3.1.2 Quasi-Experimental Designs

For most EE programs, either practical concerns or design factors will limit the use of RCT methods. In these situations, quasi-experimental designs are often a good option. The use of quasi-experimental designs is not unique to EE and has been often used in evaluations of private and public investments. Stuart (2010) reviews the evolving literature on matching and propensity scoring methods in quasi-experimental designs and states that such methods "… are gaining popularity in fields such as economics, epidemiology, medicine and political science." [27,28]

Quasi-experimental designs are similar to RCT, except for the important fact that random assignment is not used. Typically, consumers select themselves into the participant group, and the evaluation researcher must then develop the comparison group. To avoid confusion, the term

[27] Stuart (2010) also provides a guide to software for matching since software limitations made it difficult to implement many of the more advanced matching methods. However, recent advances have made these methods more and more accessible. This section lists some of the major matching procedures available. A continuously updated version is also available at Http://www.biostat.jhsph.edu/~estuart/propensityscoresoftware.html. Common statistical software packages such as STATA, SAS and R address most of the current matching approaches.

[28] The majority of attribution analyses assessing business decisions and public or private investments use quasi-experimental designs as many practical factors result in the use of this method. As an extreme example, consider a study that is designed to assess the health effects of smoking. Would it be appropriate to select a study population of 9,000 18-year-olds and assign one-third to a group that does not smoke, one-third to a group that smokes a pack of cigarettes a day, and one-third to a group that smokes a pack a day, but with some mitigating medications? Clearly, this type of RCT would pose ethical issues. As a result, natural quasi-experiments are used where existing smokers are matched with a comparison group of non-smokers that is as representative as possible. The methods of matching on observable characteristics have become quite advanced in the past decade.

"comparison group" is used for quasi-experimental designs, and the term "control group" is used with RCT designs.

The evaluator's goal is to select a comparison group that matches the participant group in terms of the actions that influence energy use. If done well, the only significant difference between the two groups will be participation in the program. Still, how well the comparison group actually matches the participant group will always be subject to some uncertainty, as there may be *unobservable* variables that affect energy use, the attribute of interest. Stuart (2010) defines the problem this way:

> *One of the key benefits of randomized experiments for estimating causal effects is that the treated and control groups are guaranteed to be only randomly different from one another on all background covariates, both observed and unobserved. Work on matching methods has examined how to replicate this as much as possible for observed covariates with observational (nonrandomized) data. .... While extensive time and effort [are] put into the careful design of randomized experiments, relatively little effort is put into the corresponding "design" of nonexperimental [quasi-experimental] studies. In fact, precisely because nonexperimental studies do not have the benefit of randomization, they require even more careful design.*

"Matching" is broadly defined in the literature to be any method that aims to equate (or balance) the distribution of covariates in the treatment group and the comparison group. This may involve methods such as 1:1 matching (in which each participant is matched to another customer that did not participate), weighting, or sub-classification.

Matching Methods

Chapter 8 of the Uniform Method Project discusses consumption data analyses including alternatives for constructing comparison groups. Also, the two SEE Action guides (2012a and 2012b) address matching. Matching methods considered include:

- ***Participants as the Comparison Group:*** SEE Action (2012b, p 3-6) states that among quasi-experimental approaches, "perhaps the most common [is] the 'pre-post' approach. With this approach, sites in the treatment group after they were enrolled in the program are compared with the same sites' historical energy use prior to program enrollment. In effect, this means that each site in the treatment group is its own nonrandom control group."

  By using the participant group as its own comparison group, the energy use of the participants during a period before they participated in the program is used as the comparison or baseline. A statistical consumption analysis is used that also includes factors that are expected to influence energy use and may vary across the pre-post time periods. Weather is the most obvious additional variable that should be controlled, but there may be other variables as well such as economic factors if the

periods cover a two-year period or greater. Agnew et al. (2013) provide a useful set of algorithms for making weather adjustments.[29]

- ***Nonparticipants in the Comparison Group:*** The trend in the literature is to move away from the simple approach of using participants as their own control group in a time-series analysis and, instead, develop cross-sectional time-series data that includes data on participants and matched nonparticipants. [30] These data sets allow for the use of panel models[31] and DiD methods.

The simplest form of matching uses data that are already available. In the early days of evaluation of residential programs, evaluators matched by ZIP codes, based on the assumption that consumers within the same ZIP code would have similar characteristics. However, this method is not very refined.

More recent approaches have focused on matching by energy-use and energy-use distributions across months and seasons. These matching methods can be simple or sophisticated, even when matching is confined to energy-use data already available (i.e., no additional surveys of nonparticipants are conducted). Matching on energy use can be as simple as stratifying participants and nonparticipants by their energy consumption (for season, year or month) and then drawing nonparticipants to match the participants' distribution of energy use.

As discussed in Stuart (2010), the literature on matching based on energy use has been expanding. Provencher et al. (2013) focuses on a comparison of the distribution of energy across both months and seasons. The analysis follows the approach advocated by Ho et al. (2007) and Stuart (2010). The procedure used by Provencher and Glinsmann involves matching each participant household to a comparison household based on a minimum distance criterion – in this case, the minimum sum of squared deviations in monthly energy consumption for the three months of the specified season in the pre-program year.[32]

---

[29] There are other approaches that can be used for weather normalization, particularly if the evaluator is interested in changes in monthly peak demand in addition to average monthly energy use. Additional weather normalization approaches are discussed in Eto (1988) and in McMenamin (2008).

[30] Practical references for matching methods include: (1) Stuart, E.A. (2010); (2) Ho, D. et al. (2007) and (3) Abadie and Ibens (2011).

[31] Panel (data) analysis is a statistical method widely used in social science, epidemiology, and econometrics which deals with two-dimensional (cross sectional/times series) panel data. The data are usually collected over time and for the same individuals.

[32] In the program evaluation literature, matching often involves matching on variables with different metrics, e.g. energy use and square footage of the household. These variables are normalized in the application of the distance criterion, usually using the full covariance matrix for the variables, or the inverse of the standard error for each variable (the Mahalanobis metric). When you only consider past energy use, such as monthly energy use, this sort of normalization isn't necessary because all measures are in the same units. The Mahalanobis metric is used frequently in most propensity scoring applications. The original reference is Mahalanobis (1936) and the use of the metric is covered in Stuart (2010). One application, among many examples, is Feng (2006) which also includes the SAS code for this method.

In the second step, a panel data set consisting of the monthly energy use by program households and their matched comparisons are constructed for the same season in the program year and used in a regression model predicting monthly energy use for the season. This matching is viewed by many as preferable to that involving the distribution of households across ZIP codes or demographic variables because the estimate of program energy savings is based on the assumption that the comparison households are "just like" treatment households in their energy use, except for the effect of the program. Energy use is then the variable of greatest concern for the non-random assignment of households into the treatment and the control groups. To the extent that additional variables (such as heat type) are available at the customer level, the evaluator's validation of the two-stage RCT can be extended to these. However, Provencher and Glinsmann state that this is not necessary:

> *Strong evidence that groups of households have the same distribution of energy use in the pre-program period is sufficient to establish that estimates of program savings will be unbiased. Differences that matter, such as heat type, would be revealed in the comparison of monthly energy use in the pre-program period.*

These matching methods tend to follow the literature reviewed in Stuart (2010). Stuart indicates that matching methods have four key steps, with the first three representing the "design" and the fourth the "analysis":

1. Defining "closeness": the distance measure used to determine whether an individual is a good match for another;
2. Implementing a matching method appropriate to the measure of closeness;
3. Assessing the quality of the resulting matched samples (and perhaps iterating Step 1 and Step 2 until well-matched samples result); and
4. Analyzing the outcome and estimating the treatment effect, given the matching done in Step 3.

Closeness in Step 1 above is often defined as a minimum distance value as was used in Provencher and Glinsmann above.

Another approach for identifying nonparticipants is "propensity scoring." The most common method used in propensity score estimation involves the estimation of a logistic regression. This model uses information on both participants and nonparticipants to estimate a dependent variable that is assigned the value of 1 if that customer is a participant or 0 if the customer is a nonparticipant. This process allows the data to identify nonparticipants who are similar to

participants in terms of a propensity score (i.e., similar attributes between participants and nonparticipants). This approach has a long history in in the EE evaluation literature.[33,34]

Regression Discontinuity Design

The regression discontinuity design (RDD) is discussed in SEE Action evaluation guides (2012a, 2012b). This method is becoming more widely used, but applies to programs where there is a cut-off point or other discontinuity that separates otherwise likely program participants into two groups. This approach to matching examines the impacts of a program by using a cutoff value that puts consumers in or out of the program through a design that does not involve their selecting themselves into the program or choosing not to participate. As a result, this approach addresses the self-selection issue.[35] By comparing observations lying closely on either side of a cut-off or threshold, it is possible to estimate the average treatment effect in environments where randomization is not possible.[36] The underlying assumption in RDD is that assignment to participant and nonparticipant groups is effectively random at the threshold for treatment. If this holds, then those who just met the threshold for participating are comparable to those who just missed the cutoff and did not participate in the program.

The SEE Action reports indicate that RDD is a good candidate for yielding unbiased estimates of energy savings. The example used by SEE Action is based on an eligibility requirement for households to participate in a program. This requirement might be that a customer whose energy consumption exceeds 900 kWh per month would be eligible to participate in a behavior-based efficiency program, while consumers who consume less than 900 kWh per month would be ineligible. Thus, the group of households immediately below the usage cutoff level might be used as the comparison group.

For participating and non-participating households near the cut-off point of 900 kWh in monthly consumption, this is likely to be an extremely good design. In the larger context, this RDD assumes that the program impact is constant across all ranges of the eligibility requirement variable, i.e., the impact is the same for households at all levels of energy usage. This assumption must be considered carefully for participating households that might consume much more than 900 kWh per month (e.g., 2,000 kWh or more for some participants). Households with greater

---

[33] The use of discrete choice methods to address self-selection bias in evaluations of EE programs has been presented in early evaluation handbooks -- See: Violette et al. (1991) and Oak Ridge National Laboratories (1991).

[34] Southern California Edison (2012), provides a recent behavioral impact application using propensity scoring.

[35] In the recent years, there has been a strong movement towards focusing on the "identification" issue evaluation, i.e., the issue that in the absence of a randomized controlled trial you do not really know if the error term in a regression is correlated with the explanatory variable of interest, so your estimate of the coefficient on that explanatory variable should be assumed to be biased in the absence of "sound" corrective action. A regression discontinuity design addresses this issue.

[36] The regression discontinuity design (RDD) has a history in evaluation dating back to the 1960s. This approach has been used to assess a wide variety of attribution analyses in the fields of education, health, and policy. Recently, this approach has been used more often. For a review of RDD see: Imbens, G. and Lemieux, T.: *Regression Discontinuity Designs: A Guide to Practice*, 2010, *Journal of Economic Literature* 48, 281-355.

consumption may have greater opportunities for energy use reductions (although, the change might be constant on a percentage level). In this example, potential concerns about program impacts being constant across different levels of household energy use makes Stuart's Step 3 important: assessing the quality of the resulting matched samples.

The example above is only one instance of a discontinuity. Another example would be a time-based cutoff point. Because utilities often have annual budgets for certain programs, it is not uncommon for a program to exhaust its budget with a number of months (sometimes as many as six months) remaining in the year. In this case, a date-based cutoff might be useful. Consumers who apply for the program after the enrollment cutoff date imposed by budget restrictions might be similar to the program participants who were accepted into the program during the first six months of the year. Also, they may have a more similar distribution of energy use per month, the focus of the impact assessment.

Random Encouragement Design

Random encouragement designs (RED) are applicable to the types of data available for EE program evaluation. Like RDD above, it is another way to incorporate randomization into the evaluation design. The issue is again non-random assignment of consumers into a program participant group. For the reasons cited above, consumers self-select themselves into the program which can bias the results of the evaluation. With non-random assignment, the participant and the comparison group may differ on the variable of interest and that observed difference may be mistakenly attributed to program participation.

The RED design is useful in estimating the effect of an intervention that cannot be randomly administered. RED involves taking a randomly selected group of participants to receive extra encouragement, which typically takes the form of additional information or incentives. A successful encouragement design allows estimation of the effect of the intervention as well as the effect of the encouragement itself (See Diamond and Haninmueller, 2007). In this case, there may be an EE program for which all consumers can decide to opt in. This might be a residential audit program, or a commercial audit or controls programs. A group of randomly selected consumers are then provided extra encouragement in terms of information and/or financial incentives. This randomization can ameliorate the effects of self-selection.[37]

An application of RED to a residential weatherization program is outlined in Fowlie and Wolfram (2009) where they address the design of the study. They point out that:

> *REDs are particularly useful when:*

[37] The underlying estimation concept in RED is explained in U.S. Department of Energy (2010): "In RED, researchers indirectly manipulate program participation using an encouragement "instrument" so as to generate the exogenous variation in program participation that is so essential for causal inference. This exogenous variation can then be used to identify the effect of the program on those households whose participation was contingent upon the encouragement." Other useful references to RED are Bradlow (1998) and West (2008) as well as two documents by Fowlie and Wolfram (undated); links are included in the Bibliography.

- *Randomization of access or mandatory participation is not practical or desirable.*
- *No need to ration available services (i.e., demand does not exceed supply).*
- *The effects of both participation and outreach are of policy interest.*

Rather than randomize over the intervention itself, we randomly manipulate encouragement to participate.

This allows the effect of the encouragement to produce exogenous variation in program participation which can be used to identify the effect of the program on participants (U.S. Department of Energy, 2010).

There are practical issues which must be taken into account in any research design and RED is no exception. The sample sizes needed for a RED study are typically larger than for a pure RCT, and those receiving the encouragement do, in fact, need to show different participation rates.[38] This research design should be given consideration in estimating net savings as it aligns well with many standard EE program implementation plans. The random variation is designed not by excluding participants, but simply through the provision of enhanced information and/or incentives offered to consumers. Work that is ongoing using RED should provide useful information for practitioners, but few examples exist in the EE evaluation literature to date.

Summary Quasi-Experimental Designs – Matching and Randomized Designs

While it is impossible to determine definitively whether the matching, RDD or RED designs discussed above provide an appropriate comparison group, there are tests that can provide evidence that either supports or discounts the validity of the RDD design and other quasi-experimental designs. Additionally, Fowlie et al. (2009) point out that there have been studies comparing these designs to the ideal RCT and with comparison studies that do not address systematic bias between the participant and control groups. The finding is that randomized designs (either RDD or RED) are an improvement over simple comparison approaches. RDD depends on the program having a cutoff point for participation that which allows random selection. RED may be a good fit with many EE programs that have a large number of participants, but appropriate design in the types of information and incentives are required. These methods should be viewed as options whenever a program contains a large number of participants, preferably 500 or more.

Importantly, these methods must be considered in advance of program implementation to allow for the appropriate data, or the design of the information or incentives that will be offered potential participants. It has always been important to consider evaluation when designing or revising EE programs, but the consideration of these randomized overlays to assist in evaluation makes this even more critical.

[38] This can be one of the challenges in the design of RED approach. The design of the encouragement given to a random sample of participants must be effective – that is, produce higher acceptance rates than for the balance of the participant group.

3.2 Survey-Based Approaches

This section describes the survey-based approach and the analytic use of the data obtained. Surveys are commonly conducted to collect NTG-related data and, despite the many drawbacks discussed within this section, this approach is typically the most cost-effective, transparent, and flexible method for estimating NTG. Consequently, it is the most frequently employed NTG methodology.

Surveys may target up to three types of respondents: (1) program participants, (2) program nonparticipants, and (3) market actors.[39] While surveys with these three types of respondents are described individually within this section, the best-practice approach is to triangulate and use multiple survey approaches (e.g., enhanced self-report) or multiple net savings estimation approaches.

The methods discussed in the preceding section provide estimates of net savings directly. That is, those approaches either compare a participant group to a random control group (as part of an RCT) or to a comparison group from a well-designed, quasi-experimental application, and those approaches do not require a separate effort to estimate FR, SO, or market effects.[40]

Survey based approaches are used in evaluations that start with gross estimates and then adjusts for NTG factors – FR, SO, and market effects. This approach can be a cost-efficient means to estimate NTG factors, but it is not without issues as discussed in the following sub-sections. The reader is also encouraged to refer to Chapter 12, "Survey Design and Implementation", of the Uniform Methods Project (Baumgartner, 2013), that discusses many of the issues involved in using surveys to estimate NTG.

3.2.1 Program Participant Surveys

Survey-based methods for estimating net savings from program participants who are aware of the program incentives/services use questions about the program's influence on the participants' actions and decision-making. Participants answer a series of closed-ended and open-ended questions on these topics:

- Why they installed the program-eligible equipment.
- What they would have done in the absence of the program incentive and other services.
- What further actions they took on their own because of their experiences with the program.

[39] Note that a Delphi panel, which also uses surveys of a panel of experts, is discussed under Section 4.6 of this chapter.

[40] Market effects can be viewed as longer-term SO effects; therefore, it is unlikely that any market effects are included in a RCT net savings approach spanning just a few years.

As noted in Chapter "12, Survey Design and Implementation", of the Uniform Methods Project (Baumgartner, 2013), best practice survey design for attitudes and behavior measurement use multiple-item scales to better represent the construct. Since participant decision-making is complex, the survey must ask a carefully designed series of questions rather than a single-question, as that could result in misleading findings.[41]

The primary benefits of a survey-based approach are as follows:

- Implementing a survey typically costs less than many other approaches, particularly if the effort is combined with data collection activities already planned for process and impact evaluations.
- The evaluator has the flexibility to tailor questions based on variations in program design or implementation methods.
- It can yield estimates of FR and SO without the need for a nonparticipant control group (NMR et al., 2010). Note however, that participant surveys only capture a subset of market effects[42], a key piece of NTG.

Despite these benefits, the self-report approach has been one of the most maligned approaches for measuring net savings (Ridge et al., 2009). The main concerns related to relying on a survey-based self-report approach are these (Peters et al., 2008):

- There is a potential bias related to respondents' giving socially desirable answers.[43]
- The inability of consumers to know what they would have done in a hypothetical alternative situation, especially in current program designs that use multiple methods to influence behavior.
- The tendency of respondents to rationalize past decisions on choices.
- There is a potential for arbitrariness in the scoring methods that translate responses into freerider estimates.
- Consumers may fail to recognize the influence of the program on other parties who influenced their decisions. For example, a program having market effects may have influenced contractor practices, which, in turn, may have indirectly impacted the participants' (and nonparticipants') decisions.

---

[41] Discussions of the sequencing of a series of questions can be found in SEE Action (2012b), Megdal et al. (2009), Haeri and Khawaja (2012), as well as the recent evaluation standards adopted in New York (New York Department of Public Service. July, 2013)

[42] Participant surveys can, in theory, capture end user market effects, e.g., changes in end-user awareness, knowledge, efficiency-related procurement practices, etc..

[43] Participants may also have a bias toward overstating program impacts due to the desire to retain incentives, although this has not been widely documented.

While these concerns are valid, it is important to note that all methodologies have their inherent biases. For example, market sales analysis[44] which is based on objective sales data can be biased if the market actors who provide data for the analysis operate differently from those not participating in the study or if the comparison area is systematically non-comparable.

Ridge et al. (2009) point out that it does not make sense to compare all self-report approaches equally, as some conform to best practice, while others do not. Keating (2009) adds that many of the criticisms of the self-report approach can be alleviated through careful research design, sampling, survey timing, and question-wording.

In Chapter 12 of the Uniform Methods Project, Baumgartner (2013) presents guidelines for selecting appropriate survey designs and recommends procedures for administering best-practice surveys. The literature also contains a number of best-practice elements for survey design, data collection, and analytic methods specific to estimating net savings (NYS DPS, 2013) (Tetra Tech et al., 2011) (Ridge et al. 2009). This literature notes the importance of making the entire process transparent so that stakeholders can understand how each question and its responses impact the final estimate. Thus, the report should contain details of critical elements such as the question sequence, scoring algorithms, and the handling of inconsistent and/or missing data.

Presented below are some of the best practices regarding survey design, data collection, and analytic elements related to net savings estimation (Tetra Tech et al. 2011).

Survey Design Elements[45]

There are a number of design elements that need to be considered when developing surveys. A list of best practices is shown below:

- Identify the key decision-maker(s) for the specific energy efficient project. For downstream programs, a key decision-maker in the household or business is likely to have been responsible for making the final decision, although they may assert that their vendor was the most influential in their decision. Although consumers ultimately decide on what they will purchase, consumers may not be aware of the influence of the interventions for upstream programs where trade ally decisions are driving change (e.g., original equipment manufacturers determining equipment energy efficiency levels, and retailers determining what equipment to stock, market, or advertise as a result of upstream program incentives).

[44] Market sales analysis captures the total net effect of a program. Ideally, this method involves obtaining comprehensive pre- and post-market sales data in both the area of interest and an appropriate comparison area and examining the change in the program area compared with the change in the non-program area (Tetra Tech et al., 2011).

[45] Comments received from chapter reviewers and, in particular Mr. Michael Rufo, Itron Inc, provided additional contribution to this section.

- Use set-up or warm-up questions to help the decision-maker(s) recall the sequence of past events and how these events affected their decision to adopt the measure.
- Use multiple questions to limit the potential for misunderstanding or the influence of individual anomalous responses.
- Use questions that rule out rival hypotheses for installing the efficient equipment.
- Test the questions for validity and reliability.
- Use consistency checks when conducting the survey so that inconsistent responses can be clarified immediately.
- Use measure-specific questions to improve the respondent's ability to provide concrete answers and to recognize that different measures installed at a project may have had different motivations.
- Use questions that capture partial efficiency improvement (accounting for savings above baseline but less than program eligible), quantity purchased, and timing of the purchase (where applicable for a measure) to estimate partial freeridership.
- Use neutral language that does not lead the respondent to an "expected" answer.
- Use combinations of open- and close-ended questions to balance hearing from the end user in their own words and creating an efficient, structured, and internally consistent data set.

Data Collection Elements

Even if stellar survey design elements are present, best-practice data collection practices are key requirements for collecting reliable information and calculating valid estimates. These data collection practices entail the following:

- Pre-test the survey instrument to ensure that questions are understandable, skip patterns are correct, and the interview flows smoothly;
- Use techniques to minimize nonresponse bias, such as advance letters on utility or program administrator letterhead (the organization for which the participant will most likely associate the program) and multiple follow-ups over a number of weeks;
- Follow professional standards for conducting surveys, which include training and monitoring interviewers[46];
- Determine the necessary expertise of the interviewer based on the complexity and value of the interview (e.g., the largest, most complex projects in custom programs

[46] Data collections surveys can be conducted via telephone, web (including smart phones), postal mail, and in-person. For large complex C&I projects, an energy engineer knowledgeable with the type of project and technology should conduct the interviews.

are often better addressed by trained evaluation professionals rather than general telephone surveyors);

- Time the data collection so it occurs as soon as possible after installation, as this minimizes recall bias and provides timely feedback on program design. Recognize, however, that timely data collection for estimating freeridership will underestimate participant spillover, as little time may have passed since program participation. Although this can be alleviated by conducting a separate spillover survey at a later date with these same participants, having a separate survey will increase data collection costs (which may be warranted if spillover effects are likely to have occurred);
- Sample a census of (or oversample) the largest savers and, depending upon program participation, end-uses with few installations to ensure these are sufficiently represented in the survey sample.

Analytic Elements

In addition to discussing survey design and data collection elements, much of the literature discusses best practices for analysis, such as:

- Treat acceleration of the installation of the EE measures appropriately to produce lifetime net savings rather than first-year net savings (this requires understanding the program's influence on the timing of the project)[47];
- Incorporate the influence of previous participation in the program;
- Establish a priori rules for treatment of missing/don't knows in the scoring algorithm;
- Weight the estimates by annual savings to account for the size of the savings impacts for each customer;
- Sample, calculate, and report the precision[48] of the estimate for the design element of interest (measure, project type, or end-use);
- Conduct sensitivity testing of the scoring algorithm;
- Define what the SO measurement is—and is not—attempting to estimate and explain why that approach was used;

[47] A chapter review comment provided by Mr. Michael Rufo, Itron, notes that "A focus on program induced early replacement versus the effect on efficiency level is gaining attention in the evaluation field. In cases where there is early replacement, two net savings components may be needed to appropriately characterize overall net savings: (1) the early replacement period that uses an in situ baseline; and, (2) the efficiency increment above minimum or standard practice at the end of the early adoption period (i.e., one for the RUL (remaining useful life) period and one for the remainder of the EUL."

[48] Appendix H of the Evaluation Plan Guidance for EEPS Program Administrators in New York (New York DPS, 2013) presents guidelines for calculating the relative precision of program net savings estimates for different types of estimates, including NTG ratio based on the self-report method and for spillover savings.

- Employ, where feasible, a preponderance of evidence (or triangulation of results) approach that uses data from multiple sources[49], especially for large savers and complex decision-making cases. Potential data sources could include project file reviews, program staff and account manager interviews, vendor interviews, and site visit observations.

Additional guidelines specific to the estimation of spillover savings were developed to address recurring methodological limitations that the New York Department of Public Service staff and its contractor team had observed in the estimation of spillover in New York and the industry as a whole (New York Department of Public Service, 2012). Prahl et al. (2013) summarizes this work and the critical decisions that evaluators must make before deciding whether and how to estimate spillover. That paper also discusses how the estimation of per-unit gross savings, estimation of program influence, and documentation of causal mechanisms varies for different levels of rigor.

3.2.2 Surveys of Program Nonparticipants

Self-report surveys with nonparticipants are most commonly used to triangulate customer self-report responses and to collect data for calculating nonparticipant spillover or market effects. These surveys are used to understand what EE actions nonparticipants have taken and whether those actions were undertaken because of program influences (for nonparticipant SO). These estimates can then be compared with the estimates calculated from participant surveys.

Conducting surveys with nonparticipants poses its own unique challenges. First, as there is no record of the equipment purchase, identifying a group of nonparticipants who have installed energy-efficient equipment on their own can be time-consuming and costly.[50] Second, establishing causality entails estimating gross unit savings (often with limited evidence other than the consumer self-report), and establishing the role that program influences might have had on the consumer's decision. The consumer may not have been aware, for example, of the influence the program had on the equipment's availability or the market actor's stocking practices.

3.2.3 Market Actor Surveys

When estimating net savings, it is important to consider all the various points of program influence. In addition to targeting consumers, upstream and midstream programs often target program services and/or funding to market actors (such as contractors, auditors, and design specialists) with the goal of influencing their design, specification, recommendation, and installation practices. Thus, in upstream and midstream programs, consumers may not be aware of program influences on sales, stocking practices or prices (as discussed in Appendix A). [51] As a

---

[49] See, for example, Itron, February, 2010. 2006-2008 Evaluation Report for PG&E Fabrication, Process and Manufacturing Contract Group, for the California Public Utilities Commission, Energy Division.

[50] One approach to mitigating the efficiency and cost of this is to use one nonparticipant survey that asks about a variety of program eligible measures and use the results across multiple programs.

[51] There are studies that focus on examining how a change in the price of an energy efficient product influences consumer purchases. Two approaches have been used – 1) Stated Preference experiments that systematically

result, it is not appropriate to use only participant self-reports when estimating net savings. In these cases, evaluators use market actor self-report surveys to examine the effect of these upstream influences.

These market actor self-report surveys can be designed as qualitative in-depth interviews or as structured surveys with a statistically designed sample of contractors. The format of the survey is predicated on the use and application of the data. For example:

- Qualitative, open-ended data based on a small sample of market actors may be used to contextualize market actors' practices and, as such, are best used for triangulation purposes.
- Quantitative market actor data may be used to calculate FR and SO rates specifically related to the practices of those market actors. Those calculated rates can then be directly integrated with participant self-report results, triangulated with participant self-report results, and/or used as the sole source for FR and SO rates.

Market actor survey data can also be used to estimate nonparticipant spillover and market effects. An important issue related to the quantification of nonparticipant spillover savings using only surveys of consumers is valuing the savings of measures installed outside the program. As previously noted, during telephone interviews, consumers often cannot provide adequate equipment-specific data on new equipment that was installed either through or outside a program. Although they are usually able to report what type of equipment was installed, they typically cannot provide sufficient information about the quantity, size, efficiency, and/or operation of that equipment to make a determination about its program eligibility.

One approach to estimating nonparticipant spillover and market effects via market actors is to ask market actors questions such as: (1) what percentage of their sales meets or exceeds the program standards for each program measure category installed through the program(s); and (2) what percentage of these sales did not receive an incentive. The market actors are then asked several questions about the program's impact on their decision to recommend and/or install this efficient equipment outside the program.

3.2.4 Case Studies for Estimating Net Savings Using Survey Approaches

Three examples of estimating net savings with self-report surveys are presented in this section.

---

ask potential consumers what they would choose from a set of options with different features and prices; and, 2) Revealed Preference studies observe the actual choices consumers make from true choices available to them when making purchases. To obtain accurate revealed preference information, it is usually necessary to observe the items purchased. Consumers cannot reliably report the efficiency levels of recently purchased equipment. Direct observation can be accomplished via store intercepts for small items such as light bulbs, or via onsite visits for large items such as refrigerators. The remaining challenge for this method is the potential nonresponse bias; that is, potential differences between consumers willing to have their purchases observed and those who decline. An example of a study that focuses on how changes in price influence consumer purchases of energy efficient products is Cadmus Group (2012). See Appendix A for additional information.

- The first demonstrates how the participant self-reports method has been used to calculate freeridership of nonresidential programs in California.
- The second demonstrates how a sample set of survey questions can be used in conjunction with a matrix to estimate freeridership.
- The final example summarizes an approach used by the Energy Trust of Oregon (Castor, 2012) that calculates low, mid, and high scenario NTGRs to account for "don't know" responses to certain questions. This example addresses the best practice of conducting sensitivity analysis on the algorithm used to estimate NTG.

EXAMPLE 1: Nonresidential Programs Freeridership Assessment

The Large Nonresidential Freeridership Approach, developed by the nonresidential Net-to-Gross Ratio Working Group for the Energy Division of the California Public Utilities Commission (2012), was developed to address the unique needs of Large Nonresidential customer projects developed through energy efficiency programs offered by the four California investor-owned utilities and third-parties. As described in the Framework, the method relies exclusively on the Self-Report Approach (SRA) to estimate project and program-level Net-to-Gross Ratios (NTGRs), as the working group notes that other available methods and research designs are generally not feasible for large nonresidential customer programs. This methodology provides a standard framework, including decision rules, for integrating findings from both quantitative and qualitative information in the calculation of the net-to-gross ratio in a systematic and consistent manner.

The approach describes three levels of freeridership analysis. The most detailed level of analysis, the **Standard – Very Large Project** NTGR, is applied to the largest and most complex projects (representing 10 to 20% of the total) with the greatest expected levels of gross savings. The **Standard** NTGR, involving a somewhat less detailed level of analysis, is applied to projects with moderately high levels of gross savings. The **Basic** NTGR, is applied to all remaining projects.

There are five potential sources of freeridership information in this study. Each level of analysis relies on information from one or more of these sources:

1. **Program files**, which might include various pieces of information relevant to the analysis of freeridership, such as letters written by the utility's customer representatives that document what the customer had planned to do in the absence of the rebate and explain the customer's motivation for implementing the efficiency measure, as well as information on the measure payback with and without the rebate.

2. **Decision-maker surveys**, conducted with the person involved in the decision-making process which led to the implementation of measures under the program. This survey obtains highly structured responses concerning the probability that the customer would have implemented the same measure in the absence of the program. First, participants are

asked about the timing of their program awareness relative to their decision to purchase or implement the energy efficiency measure. Next, they are asked to rate the importance of the program versus non-program influences in their decision making. Third, they are asked to rate the significance of various factors and events that may have led to their decision to implement the energy efficiency measure at the time that they did (e.g., age or condition of the equipment, information from a facility audit, standard business practices, prior experience with the program or measure).

In addition, the survey obtains a description of what the customer would have done in the absence of the program, beginning with whether the implementation was an early replacement action. If it was not, the decision maker is asked to provide a description of what equipment would have been implemented in the absence of the program, including both the efficiency level and quantities of these alternative measures. This is used to adjust the gross engineering savings estimate for partial freeridership.

This survey contains a core set of questions for **Basic** NTGR sites, and several supplemental questions for both **Standard** and **Standard – Very Large** NTGR sites. For example, if a **Standard** or **Standard-Very Large** respondent indicates that a financial calculation entered highly into their decision, they are asked additional questions about their *financial criteria* for investments and their rationale for the current project in light of them. These questions are intended to provide a deeper understanding of the decision making process and the likely level of program influence versus these internal policies and procedures. Responses to these questions also serve as a basis for consistency checks to investigate conflicting answers regarding the relative importance of the program and other elements in influencing the decision. In addition, **Standard – Very Large** sites may receive additional detailed probing on various aspects of their installation decision based on industry- or technology-specific issues, as determined by review of other information sources. For **Standard-Very Large** sites all these data are used to construct an internally consistent "story" that supports the NTGR calculated based on the overall information given.

3. **Vendor Surveys**. A Vendor Survey is completed for all **Standard** and **Standard- Very Large** NTGR sites that utilized vendors, and for **Basic** NTGR sites that indicate a high level of vendor influence in the decision to implement the energy efficient measure. For those sites that indicate the vendor was very influential in decision making, the vendor survey results enter directly into the NTGR scoring.

4. **Utility and Program Staff Interviews**. For the **Standard** and **Standard-Very Large** NTGR analyses, interviews with utility staff and program staff are also conducted gather information on the historical background of the customer's decision to install the efficient equipment, the role of the utility and program staff in this decision, and the name and contact information of vendors who were involved in the specification and installation of the equipment.

5. **Other information**. For **Standard – Very Large Project** NTGR sites, secondary research of other pertinent data sources is performed. For example, this could include a review of standard and best practices through industry associations, industry experts, and information from secondary sources (such as the U.S. Department of Energy's Industrial Technologies Program, Best Practices website[52]). In addition, the **Standard- Very Large** NTGR analysis calls for interviews with other employees at the participant's firm, sometimes in other states, and equipment vendor experts from other states where the rebated equipment is being installed (some without rebates), to provide further input on standard practice within each company.

The next table shows the data sources used in each of the three levels of freeridership analysis. Although more than one level of analysis may share the same source, the amount of information that is utilized in the analysis may vary. For example, all three levels of analysis obtain core question data from the Decision Maker survey.

Information Sources for Three Levels of NTGR Analysis

	Program File	Decision Maker Survey Core Question	Vendor Surveys	Decision Maker Survey Supplemental Questions	Utility & Program Staff Interviews	Other Research Findings
Basic NTGR	√	√	√[1]		√[2]	
Standard NTGR	√	√	√[1]	√	√	
Standard NTGR-Very Large Projects	√	√	√[3]	√	√	√

[1]Only performed for sites that indicate a vendor influence score greater than maximum of the other program element scores.
[2]Only performed for sites that have a utility account representative
[3]Only performed if significant vendor influence reported or if secondary research indicates the installed measure may be becoming standard practice.

EXAMPLE 2: Freeridership Assessment for an Equipment Rebate Program

This example of how an NTG ratio can be calculated shows how a sample set of survey questions can be used in conjunction with a matrix to estimate freeridership. (This is from Chapter 5 of the Energy Efficiency Program Impact Evaluation Guide (SEE Action, 2012b.) In

[52] This website can be found at: http://www1.eere.energy.gov/industry/bestpractices/ .

this case, the assignment of the FR score is assigned by the evaluation team based on a participant's response to six questions.

FR SCORE	ALREADY ORDERED OR INSTALLED	WOULD HAVE INSTALLED WITHOUT PROGRAM	SAME EFFICIENCY	WOULD HAVE INSTALLED ALL THE MEASURES	PLANNING TO INSTALL SOON	ALREADY IN BUDGET
100%	Yes	Yes	—	—	—	—
0%	No	No	—	—	—	—
0%	No	Yes	No	—	—	—
50%	No	Yes	Yes	Yes	Yes	Yes
25%	No	Yes	Yes	Yes	No	Yes
25%	No	Yes	Yes	Yes	Yes	No
0%	No	Yes	Yes	Yes	No	No
25%	No	Yes	Yes	No	Yes	Yes
12.5%	No	Yes	Yes	No	No	Yes
12.5%	No	Yes	Yes	No	Yes	No
0%	No	Yes	Yes	No	No	No

Source: SEE Action (2012b) based on example provided by Cadmus.

One issue with this method is the somewhat arbitrary nature of the assigned FR scores associated with sets of question responses as they are dependent upon the judgment of the particular evaluator and different researchers may assign different FR scores to different sets of respondent answers. To address this, it is recommended that sensitivity analyses around these scores FR scores be used based on the judgments of people familiar with the program.[53] An example of increasing the robustness of this method can be found in an assessment of residential heating and cooling equipment for the Electric and Gas Program Administrators of Massachusetts.[54]

Other approaches use upper and lower bounds on FR that are developed directly from survey respondents.[55]

---

[53] Issues may arise if these FR scores are viewed as categories rather than as continuous variables. A 50% score may imply a higher level of FR than does a 25% score, but it may not denote that the 50% score implies that FR is, in fact, twice as high when compared to respondents placed in 25% FR score category. It is possible to perform arithmetic on these numbers and use the values to generate a mean value and even a variance, but this may not be appropriate. The lack of an accurate "distance" factor in these numbers makes the calculated variance hard to interpret. For variables that are meant to represent categories rather than continuous numeric values, frequencies are the more often used descriptive statistic.

[54] This work was conducted by a consortium of consultants under a prime contract led by The Cadmus Group, supported by Navigant, and Opinion Dynamics Corporation. (Cited as Cadmus, 2012).

[55] Violette et al. (2005) discuss approaches used in the net savings and attribution assessment for a large-scale C/I retrofit program. FR was assessed using a series of survey questions asked of various actors, including participating end-use consumers and vendors/contractors/consultants. FR was asked in both direct FR questions and in supporting, or influencing, questions. Participating owners and ESCOs/contractors in a large-scale C/I retrofit program were each asked for direct estimates of: 1) The "proportion" of the savings or measures that

EXAMPLE 3: Commercial, Industrial and Residential Scenario Analysis

An approach used by the Energy Trust of Oregon (Castor. 2012) calculates low, mid, and high scenario NTG ratios to account for the "Don't Know" responses to certain questions. This approach is described in the Appendix of that report. The project's FR score is composed of two elements: a project change score and an influence score.

- The project change score is based on the respondent's answer to the question, "Which of the following statements describe the actions you would have taken if Energy Trust incentives and information were not available?"

  Possible answer choices are assigned a number between 0 and 0.5, with 0 indicating no FR and 0.5 indicating that the project was a full freerider. Since respondents can select multiple responses to the question, their answer choice with the lowest score is selected. If the respondent selects "Don't Know," two scores are created to account for the range of possible answers (0 and 0.5).

  For commercial projects, respondents are asked this follow-up question when they report they would not have done anything differently in the absence of the program: "If your firm had not received the incentive, would it have made available the funds needed to cover the entire cost of the project?" If the respondents select "Yes," their project change score is 0.5. If the respondents select "No," their project change score is 0. However, if the respondents select "Don't Know," they are given two scores for project change, as described above.

- The influence score is based on respondents' answers to questions about the influence of Energy Trust incentives, program representatives, contractor/salesperson, studies, and other program elements. The answer choices are given a value between 0 (element's influence was a 5, extremely influential) and 0.5 (element's influence was a 1, not at all influential). The score for the most influential element is taken as the influence score. If respondents answer "Don't Know" for all elements, they are given two influence scores to account for the range of possible answers (0 and 0.5).

To generate the FR score for each project, the project change and influence scores are added. For respondents who do not provide "Don't Know" answers, this score will be a single number between 0 (no freeridership) and 1 (full freeridership). For those who gave a "Don't know" answer to one of the questions, there are two FR scores—one high and one low. For those who

---

would have been installed without the program; and 2) The "likelihood" that the measures would have been installed without the program. A three-step approach was used. Step 1 focused on whether the respondent believed that FR existed at all; if the respondent believed it existed in this project, Step 2 was to establish bounds on the FR effect, i.e., what was the smallest value that seemed reasonable and what might have been the highest reasonable FR value. Step 3 used questions to obtain where within this range the FR value was likely to fall. Appendices to Violette et al. (2005) discuss alternative approaches. This program had some unique characteristics that made this approach more tractable. It involved large-scale C/I projects and the survey respondents were provided with summaries of the technologies and measures installed.

answered "Don't know" to *both* the project change and influence questions, no score is calculated.

FR scores are averaged for all respondents in each program/measure group and the result is shown as a percentage, rather than a decimal.

- "Low Scenario" is the average of the FR scores where the low score is used for those who answered "Don't know" to a question;
- "High Scenario" is the average where the high score is used for those who answered "Don't know" to a question.
- "Mid Scenario" is the average of the Low and High Scenarios. In the case of C&I projects, individual scores are weighted by their share in the electric or gas savings of all respondents of their group before the scores are averaged for scenarios.

3.3 Common Practice Baseline Approaches

The common practice baseline approach[56] is relatively new in the broader evaluation literature, although the concept has been evolving and is being applied by the Northwest Power and Conservation Council (NW Council) in the Pacific Northwest for a number of years.[57] SEE Action describes the common practice baseline as:

> *Common practice baselines are estimates of what a typical consumer would have done at the time of the project implementation. Essentially, what is "commonly done" becomes the basis for baseline energy consumption* (SEE Action, 2012b, p 7-2)[58,59]

[56] The Common Practice Baseline section gave rise to a number of comments. Some reviewers did not see this method as parallel to the other methods presented in this chapter, as it focuses on ex ante values of the mean of market behavior and does not look at ex post information on actions or program participants. In this context, this approach was viewed as more of an ex ante deemed net savings approach (see section on deemed NTG values below). After considering these comments, the Common Practice Baseline approach was viewed as warranting a separate section due, in part, to the recent attention given this approach to net savings.

[57] Comments provided by Mr. Tom Eckman of the Northwest Power and Conservation Council (NW Council) indicated that this general approach has been applied in setting deemed savings since the 1980s, and it was designed to fit with the NW Council integrated planning process, i.e., it is meant to provide an estimate of the increment of savings beyond what system planners assume for naturally (or currently) occurring efficiency in their demand models. Additional information on this can be found at the Regional Technical Forum website of the NW Council -- http://rtf.nwcouncil.org .

[58] SEE Action (2012b) illustrates this "commonly done" baseline using an appliance example. "For example, if the program involves incenting consumers to buy high-efficiency refrigerators that use 20% less energy than the minimum requirements for ENERGY STAR® refrigerators, the common practice baseline would be refrigerators that consumers typically buy. This might be non-ENERGY STAR refrigerators, or ENERGY STAR refrigerators, or, on average, something in between."

[59] The SEE Action report (2012b) defines common practice baselines in its glossary as "The predominant technology(ies) implemented or practice(s) undertaken in a particular region or sector." (p, A-4).

This approach is based on the development of a market-based common practice baseline (sometimes called "standard industry practice"). This baseline includes a "consideration of what typically would have been done in the absence of the efficiency action" (SEE Action, 2012b).

As with other net savings approaches, the common practice baseline approach is designed to assess the savings attributable to EE program activities. One advantage claimed for the common baseline approach is that it avoids double counting of FR The concern is that the two-step approach where gross savings is estimated first and then a second step estimates FR and SO may double count at least some FR (Ridge et al. 2013 and Hall et al. 2013). The argument is that the estimated ex ante gross savings may be closer to net savings than are the estimates of net savings calculated by adjusting the gross savings estimates by FR, SO and ME because some of these factors are already contained in the process used to produce the gross savings estimates. Hall et al. (2013) state that "if a baseline approach already has freeriders in it, there is often no need to readjust the savings calculation to account for freeriders a second time." This emphasizes the need to: 1) understand the derivation of gross estimates as part of the EE evaluation process, and 2) to explicitly set out the assumed counterfactual in the net savings method being used. If these two steps are taken, double counting resulting in higher than appropriate freeridership estimates can be avoided.[60] Ridge et al. (2013) indicates that, in addition to being used by the NW Council,[61] the common practice baseline approach has also been adopted by three other jurisdictions -- the Northwest Energy Efficiency Alliance (NEEA), Indiana, and Delaware. Examples from the guidelines are shown here:

- ***NW Council's Guidelines Savings Estimation Methods (RTF 2012)***: The current practice baseline defines directly the conditions that would prevail in the absence of the program (the counterfactual), as dictated by codes and standards or the current practices of the market. (p. 2).
- ***Indiana's Evaluation Framework*** discusses the use of the standard market practice to estimate net savings (TecMarket Works 2012): The standard market practice (SMP) approach is a way to set energy impact analysis baselines so that the baseline already incorporates the influence of freeriders. In this approach, a freerider assessment is not needed because the use of a standard market practice baseline is already what the market is doing without the program's direct influence. The SMP baseline is typically set at the mean of the level of energy efficiency being installed across the market being targeted by the program. (p. 55)

---

60 Some reviewers indicated that this double counting problem may be the result of inconsistent program rules as set out by the program administrators and regulators, and was not an estimation issue. Further, a number of reviewers indicated that rather than over-estimating freeriders, this approach underestimates freeriders due to selection bias (discussed in the main body text below).

61 Comments provided by Mr. Tom Eckman of the NW Council's Regional Technical Forum (RTF) has evolved over since its initial introduction in the 1980s and that it is part of an integrated planning process, including program design, setting deemed savings values for measures, and the production of demand forecasts for integrated planning; and, this process has evolved over a number of years.

Similar excerpts from the NEEA and Delaware guidelines for net energy savings estimation can be found in Ridge et al. (2013).

Gross impact estimation itself is a value that requires a baseline. In other words, the gross savings from an energy-efficiency measure is the difference between the energy use of the installed high-efficiency equipment and an alternative equipment specification. The baseline for the gross impacts estimate may be any of the following: (1) the energy use of the equipment that was replaced during a retrofit; (2) the energy use of standard-efficiency technology that likely would have been installed by the customer; or (3) the energy use of the equipment required by codes and standards (assuming stringent enforcement of the codes and standards). In fact, Ridge et al. (2013) point out that the actual equipment baseline used to estimate gross impacts may not be clear cut and that "there are gradations in the way baselines are established in the energy-efficiency industry."

The case for the use of a common practice baseline appears to stem from two issues:

- The definition of gross savings may actually include factors that are more appropriately viewed as components of net savings, and additional adjustments are not needed to these original estimates. This is essentially an ex ante estimate of net savings using current practice as the baseline and net savings is the reduction in energy use resulting from the change to more efficient technologies. [62,63]
- Program evaluations that report net savings may do so inconsistently. Unfortunately, the components of the net savings calculation differ between jurisdictions, and those components are often based on what is viewed as appropriate and measureable by the jurisdiction's stakeholders (See NEEP, 2012). Although there is wide recognition that SO exists and can be significant, there has been resistance in a number of jurisdictions to estimating SO values and including them in the net savings calculations. ME values have faced similar challenges. [64]

---

[62] Mr. Tom Eckman of the NW Council expands on this point in personal communication stating that: "*What is occurring prior to program launch is a better measure of what would have occurred absent the program (i.e., the counterfactual) than a determination made after the program has influenced the market*." Essentially, the NW Council performs an ex ante net analysis when they developed deemed savings estimates that are by design viewed as net savings. For the NW Council's purposes, this is viewed as being as accurate as performing complex studies after the program has been implemented. More information on the NW Council approach can be found at http://rtf.nwcouncil.org/.

[63] The common practice approach as applied by the NW Council works best when the forecasts are made at the measure level. Covering all the measures which combine to make a program can be time consuming and expensive to update. Also, this is short term in that over time. The control group (i.e., nonparticipants) would likely have evolved their actions from one year to the next as conditions change and that accounting for these effects is also important in determining net savings. As with all approaches discussed in this section, there are pros and cons and the selection of the approach to use has to recognize the context in which this choice is made. For example, there are no financial incentives tied to net savings value among members of the NW Council.

[64] To further illustrate, net savings as presented in the findings of EE evaluations is always presented as "net" of something; however, it may be gross savings net FR, or it may be gross savings net FR and SO, or, in some

SEE Action (2012b, p. 7) indicates that appropriate common practice baselines can be estimated by surveys of participants, nonparticipants, and/or analysis of market data. The process of developing a working definition of common practice baselines may pose some challenges. There has not been wide-spread experience in developing common practice baselines allowing for a determination of best practices. The RTF of the NW Council has the most experience in developing these baselines with its methods emphasizing the use of market data,[65] and the RTF has produced guidelines for the development and maintenance of saving estimation methods based on common practice baseline approach (RTF, 2012).

A significant concern is that self-section bias may still be an issue with common practice baselines. An EE program that allows consumers to select themselves into the program may attract those consumers among the common practice baseline who would have taken the high-efficiency actions anyway. If an EE program only attracted those consumers who were pre-disposed to install the high-efficiency equipment promoted by the program, then net savings could be overestimated by not fully accounting for all FR. Additionally, to the extent that the program results in nonparticipant spillover, it is not clear how the common practice baseline approach would capture those savings.

Another point made in Ridge et al. (2013) is that prior EE programs have affected the markets for EE equipment through SO and ME. This results in current standard practice baselines that are more efficient than what would have been the case if these EE programs had not been offered. In this case, using market average can contain a fair number of past participants (e.g., end users, installers and distributors) who have already been influenced by the program. The effect of these past programs is to lower the annual energy use of the measures that constitute the current practice. This argument seems to be partly analytical and partly a policy consideration. Ideally, past evaluations of EE programs should have included all the impacts attributable to the programs but, since SO and ME were generally omitted from past evaluations, they have not been counted. The annual energy use that is represented by current practice is lower than it would have been if these past programs had not been offered. From this perspective, the use of unadjusted current practice baselines as estimates of net savings seems to be an effort to make up for mistakes in past evaluations (i.e., the omission of SO and ME that impact the overall market).

A jurisdiction may view this as a reasonable estimate of EE program impacts over time which best represents the overall investment in EE; or, it may take the position that each EE program should be evaluated as an incremental investment, i.e., a program implemented in 2014 should be evaluated against what is attributable to that investment only and all impacts from prior years' programs are

---

cases, market effects may be included in the defined net savings estimates. Navigant (2013) found that the majority of jurisdictions defined net savings as "gross savings adjusted only for FR." (The review of net savings methodologies in Navigant (2013) focused only on C/I programs. Out of thirty-eight C/I program evaluations reviewed, twenty-eight estimated net savings as gross savings adjusted for FR only. Three estimated net savings as gross adjusted for FR plus participant SO, and seven studies adjusted for FR and both participant and nonparticipant SO. None of the studies attempted to address market effects in addition to the spillover values.)

[65] The RTF of the NW Council believes that the emphasis on market research for developing common practice baselines will also help produce better program designs.

essentially sunk costs that are not to be considered. This is an example of where policy and analytic views of net savings estimation are linked. It is not possible to definitively recommend a net savings approach across program types and jurisdictions without considering the appropriateness of the decision to include impacts from prior programs on the current practice baseline.

The bottom line for assessing the common practice baseline approach is the same process as is used in all other methods: 1) understand the construction of the baseline used in the evaluation; and, 2) analyze the implications of this baseline against what is an appropriate counterfactual for that program. Based on this standard approach, decisions can be made regarding the net savings estimation method that is most appropriate for the evaluation of an EE program.

In summary, the use of common practice baselines has recently been adopted by several jurisdictions in their EE evaluation guidelines. This approach is evolving in these jurisdictions, and it is expected that issues in its application raised above will be addressed by stakeholders and evaluators. As with all methods, there are pros and cons. A potential strength of the common practice baseline approach is its use in upstream and market transformation EE programs. It can be applied market-wide and, unlike randomized trials and quasi-experimental designs, it does not require participants to be identified. However, one of the challenges is controlling for self-selection (i.e., the average consumer may not be the type of consumer who participates in the program).

Another factor is that the common practice baseline is essentially a snapshot in time. The common practice baseline will change over time and there will be a need for periodic updates. The complexity of the update will depend on the program type. If it is essentially a one technology program (e.g., refrigerator recycling), then the update may be straight forward. Updating common practice baselines for a large C&I custom program where many technologies and end-uses are impacted may be more difficult. Hall et al. (2013) state that they "are not suggesting that direct net analysis approaches (i.e., common practice baselines – *definition added*) should be used in all evaluations or that they can be applied to all types of program configurations or target markets." As a result, the common practice baseline approach is another technique in the toolkit that practitioners can use to address net savings, based on an analysis of the market and the appropriate counterfactual.

3.4 Market Sales Data Analyses (Cross-Sectional Studies)

A market sales data method can capture the total net effect of the program, including both freeridership and participant and nonparticipant "like" spillover. As described in a residential freeridership and spillover methodology study prepared for the Massachusetts Administrators (NMR Group, Inc. and Tetra Tech. 2011), the total net effects of a program can be estimated through an analysis of market sales data.

The most common is a cross-sectional, comparison area approach, in which post-program data are compared with data from a non-program comparison area (or multiple comparison areas) for the same point in time. Thus, a comparison could be made between the change in the program

area from the pre-program period to the post-program period *and* the change in the non-program area over the same period.

The methodology study lists three important factors to consider in deciding whether this approach is appropriate to use for a particular program:

- ***Does an appropriate comparison area(s) exist?*** Comparison area(s) must represent a credible baseline for the area of interest. This may entail using a set of systematic adjustments to control for differences in total size of, or demographics for, the areas. As EE programs become more prevalent, it is becoming more difficult to find comparison areas that do not have similar program activities.
- ***Is the market data available and complete?*** Market data analysis requires comprehensive market data for both the area of interest and an appropriate comparison area(s). The complication here is that comprehensive sales/shipment tracking systems have not been available for most markets. Absent comprehensive sales data, a general picture of market coverage can be obtained by conducting surveys or in-depth interviews. These are typically conducted with vendors and contractors about sales volumes and efficient equipment sales shares for conditions with and without the program, or for in-territory and comparison area sales. In some cases, the self-reported purchases of participating end-users' self-reported purchases can provide market data if the sample is sufficiently large and representative of the market. Also, it can be expensive to gather the market sales and shipment data, and even with a diligent data collection effort, there may be gaps in the data.
- ***What are the features of the program?*** Market data analysis is usually appropriate for programs that promote large numbers of homogenous measures and that have substantial influence "upstream" from the end-user.

As an example of this approach, Cadmus et al. (2012) tracked ENERGY STAR® appliances, lighting, and home electronics product sales in New York and then compared those sales to sales of the same products in Washington D.C., Houston, Texas, and the state of Ohio. All of these baseline areas were areas without significant utility efforts to promote ENERGY STAR products. The market data were used to estimate both the market share and the energy savings attributable to the New York Energy $mart℠ Products Initiative program administered by the New York State Energy and Research Authority.[66]

[66] Discussion with Mr. Scott Dimetrosky indicated that this study developed savings from product sales and installations. These savings were derived by first estimating the market share for ENERGY STAR products through estimates of total market size and sales of ENERGY STAR products. Next, portions of the market share were allocated to exogenous, non-NYE$P Program effects, including the impact of the national Environmental Protection Agency/Department of Energy ENERGY STAR Program, naturally occurring adoption (including the impact of higher energy prices and interest generated by programs in neighboring states), and the impacts of other NYSERDA residential programs. The remaining market share, after netting out these other effects, was considered attributable to the NYE$P Program.

Another example of a market sales approach entails interviewing or surveying a panel of trade allies who are either program participants or nonparticipants. This could include contactors, retailers, builders and installers. These trade allies are offered monetary compensation for information on projects or sales completed within a specified time period. The types of information requested might include manufacturer, efficiency levels, size, price, installation date, installation ZIP code, types of incentives received, and an assessment of the program's impact on incented and non-incented efficiency actions. With annual updates, this method could provide context for tracking longer term on-going program impacts or market effects. This method could also work in tandem with other approaches for estimating net savings and provide a market context for estimates that may otherwise focus only on short-term impacts.

3.5 Top-Down Evaluations (Macroeconomic Models)

"Top-down evaluations" use macro-data on energy consumption in a model that relates changes in energy consumption to a measure of EE effort (usually expressed as expenditures on EE). Top-down evaluation produced what has been termed "macro-consumption metrics" (MCMs) in two recent pilot applications in California (See Cadmus et al., 2012, and Demand Research, LLC, 2012). The broader literature on these methods is referred to as top-down methods, but the MCM notation adopted in the recent California pilot studies refers to the same set of methods and cites top-down studies as background for its pilot work.

The top-down approach has much appeal, as it directly addresses net savings. Since the dependent variable is overall energy use (often expressed as energy use per capita), this method simply examines the change in energy use due to EE efforts. There is no need to adjust for FR and SO, or even ME. In addition, the regression analyses provide confidence and precision levels around these estimates.

There are challenges in estimating the relationship between EE efforts and changes in overall energy consumption, such as the size of the impact to be isolated by the model. Developing a model that can measure a 1% to 2% change in total energy use annually that is attributable to EE programs requires a reasonably sophisticated structure. For example, the model must have an appropriate lag structure because the impacts from one year's expenditures will occur over a number of years.[67] In addition, the number of observations and quality of data needed to identify a small effect can be challenging. The data platform needed to support this top-down or MCM model approach requires the following:

- A measure of EE expenditures (or another metric of EE effort for different cross-sections, such as utilities or program administrators).

[67] BC Hydro (2012) demonstrates the importance of the relationship between current expenditures on EE and future savings. It also shows the importance of letting the data determine the most appropriate lag structure as opposed to implementing a fixed structure that acts as a constraint. How lagged effects are handled in the regression model influences the estimated energy savings.

- The number of observations needed to identify the effect of EE over a number of years, taking into account the lag structure of EE impacts.
- Matching demographic and macro-economic data to utility service areas, or sub-areas of utilities that are used as observations in the analyses.
- High-quality data regarding energy consumption for each cross-section being analyzed.

As a result, most top-down studies have been multi-utility or multi-state efforts that can provide a reasonably large number of cross-sectional areas for the analyses. The California pilot study was an effort at using top-down methods to estimate overall EE impacts for the state, and possibly utilities within the state.

Questions that evaluators should consider when deciding on the appropriateness or applicability of top-down models are: what information will be produced by these top-down models, if they are successfully estimated; and, how does this information compare to what is produced by other methods? For example, top-down models may be useful for:

- Estimating overall average change in energy usage die the EE programs for a region. A top-down model that provides a good fit, meets reasonable assumptions, and has acceptable levels of statistical significance levels can provide information on the average change in overall energy use (or energy use per capita) from overall EE efforts.
- Estimating regional environmental impacts. Aggregate models can be useful in assessing state and regional environmental impacts such as the impact on carbon emissions.
- Evidence of estimated energy-savings at a regional level. The model can confirm—at an aggregate level —whether the expected energy savings are actually reflected in the macro consumption data.
- Estimate overall cost savings due to EE programs. Top-down models can also be used to estimate an overall cost savings per kWh saved and confirm the efficacy of the overall EE effort.

Top-down models, however, are not able to provide information on:

- Savings produced by specific measures or programs.
- Where to best make additional investments in EE at the program or measure level.
- Insights into how existing programs might be improved.
- Estimates of FR and SO that might be used to suggest improvements in programs.
- Quality assurance/quality control information needed for regulatory oversight.

The relative importance jurisdictions and stakeholders place on program-level information versus information at the aggregate level will influence decisions to implement these different types of evaluation frameworks. Top-down approaches seem complementary to results produced by

program-level evaluations; however, there may be concerns about using these top-down methods as a replacement for program-level evaluations. The program-level research may be viewed as essential in that it helps ensure that the right set of programs comprise the EE portfolio and it can be useful in addressing program and portfolio specific questions regarding implementation. Top-down methods and program-level evaluation both provide useful, but different, perspectives on the accomplishments of EE efforts.

Cadmus (2012) reviewed a number of the leading top-down studies, each of which expresses energy consumption as a function of a metric meant to measure EE effort. The following studies are noteworthy in this review:

- Parmark and Lave (1996) used a panel data set of 39 utilities from 1970 to 1993. The claimed savings by utilities for their C/I programs was used as a proxy for the level of EE effort. The regression analysis was similar to a realization rate regression analysis model, where the coefficient on the claimed utility savings indicated what fraction of those savings were able to be found in the data. The study authors estimated the realization rate for the utility's claimed savings at 99%.
- Auffhammer, Blumstein, and Fowlie (2008)—working with data developed by Loughran and Kulick (2004) —used what has become the more traditional formulation. Here, EE effort was expressed in the econometric model as program expenditures reported to the U.S. Energy Information Administration (EIA). The study authors found that average utility reported savings (of 2% to 3%) fell within the 95% confidence interval for estimated savings. The cost of saved energy was approximately $0.06 kWh.
- Arimura, Li, Newell, and Palmer (2011) also used the EIA data on program expenditures across 307 U.S. utilities to examine the impact of investments in EE on overall energy consumption[68] Using utility Energy Information Administration (EIA) data from 1989 to 2006, the study authors found electricity savings of 1.8% annually and estimated the cost of saved energy at approximately $0.05/kWh.

The California Pilot Project on top-down methods involved two efforts, Cadmus et al. (2012) and Demand Research, LLC (2012).

EXAMPLE 1: The Cadmus Group California Top-Down Pilot Study

Cadmus used expenditures on EE programs as the level of EE effort in its models. The models were estimated at the utility level for residential and non-residential energy savings. Cadmus worked with data at the utility level using information from the three investor-owned utilities (IOUs) and from large public utilities in California such as Los Angeles Department of Water

[68] Arimura et al. (2011) also advance the state of the practice by modeling energy prices and utility energy-efficiency program expenditures as endogenous and allowing consumption to depend on program expenditures in a flexible way. The literature on Top-Down models represents sophisticated applications of econometric methods. Problems of endogeniety and autocorrelation with flexible lag structures have become common issues that are addressed by these models.

and Power (LADWP) and the Sacramento Municipal Utility District (SMUD). Data were also collected from some small public utilities, but this information was generally inconsistent.

A number of different models estimated the relationship between utility energy consumption for residential and non-residential customer segments and expenditures on EE.[69] Overall, it was difficult to obtain significant results across the group of models that were estimated. The best model did produce significant coefficients on the EE expenditures variable using data only from the three IOUs. To demonstrate the information that can be produced by top-down models, Cadmus developed estimates of savings from EE efforts over a 6 year period, and calculated the cost of energy saved. Savings from EE spending for the 2005 to 2010 period were estimated at 8%, and the cost per kWh saved was estimated at $0.05. The results of the Cadmus study indicated savings within 10% of the net savings reported by California IOUs for the 2006-2008 program cycle. The estimates of both energy savings and cost per kWh saved had large confidence intervals: ±66% on the energy savings estimate and over ±100% on cost per kWh saved. The number of observations (48 total observations) in the top-down IOU model resulted in lower precision than would have been obtained by studies with much larger sample sizes.

Cadmus did look into disaggregating the data beyond the IOU level to gain more cross-sections for the analysis; however, there was concern about the ability to allocate EE program expenditures to smaller geographic areas. One specific concern was the savings from CFLs. Over 50% of the expected savings were from CFLs and these sales were tracked at point of sale instead of the location where they were used making it difficult to get alignment between energy consumption and the impact of EE expenditures for smaller geographic areas.

EXAMPLE 2: Demand Research, LLC California Top-Down Pilot Study

Demand Research (2012) developed an MCM model working with California utilities and program contractors that disaggregated residential energy use and estimates of residential sector EE efforts into a data-base of cross-sectional observations at the census tract level. Commercial and industrial sector energy use and metrics for EE efforts were disaggregated down to the county level. Instead of using energy expenditures, the Demand Research, LLC study used the utilities' *ex ante* estimates of energy saved by census tract as the metric of residential EE effort.[70] For the commercial and industrial sectors, county-level data were developed. The independent variable for the EE level of effort in the commercial sector model was metric related to

[69] Cadmus (2012) did not try to estimate separate models for commercial and industrial consumers as the time-series was inconsistent. In some years, commercial sector consumption would increase and industrial consumption would decrease by approximately the same amount. This suggested that there was some switching in the definition on the commercial and industrial rate classes. As a result, the two classes were modeled together.

[70] This approach is similar to that used by Parmark and Lave (1996).

incentives paid; however, ex ante energy savings was used as the metric for EE effort by county for the industrial sector just similar to what was done in the residential sector model.[71, 72]

- The residential models estimated by Demand Research, LLC (2012) showed that higher levels of the EE effort variable resulted in reduced energy use with estimates significant at a 95% confidence interval.
- The commercial sector model produced the expected sign on the EE effort variable, but the results were not statistically significant.
- The industrial sector model did produce statistically significant results for the EE effort variable.

The findings from the Demand Research, LLC study were:

- The residential, commercial and industrial sector models produced state-wide savings estimates of 7.3% for the five-year period from 2006 to 2010.
- The relative precision for the aggregate savings estimate was ±31% (or a 90% confidence interval of 5.0% to 9.5%).
- The estimated state-wide savings of 7.3% exceeded the utility ex ante estimates of 4.8%.

The aggregate state-wide estimate of energy savings across all three sectors was forecasted with reasonable confidence and precision. Looking at the results at one level of disaggregation lower, i.e., at the sector level results, these results show a high degree of variability.

- The estimated industrial energy savings (all three utilities combined) were much higher than the utilities' ex ante values, about 745% higher (see Demand Research, LLC, 2012, p36).
- The commercial sector kWh savings estimates (all three IOUs combined) were much lower than the utility ex ante estimates (i.e., about 27% of the ex ante savings), and

---

[71] The reason why different metrics for EE level of effort were used in the commercial and industrial sector model was due to the method selected to address endogeniety in the commercial sector model, i.e., ensuring that the EE level of effort variables uncorrelated with the error term.

[72] Considerable work went into creating these sector databases. The details can be found in the full study, but as an overview of the effort -- key energy consumption and program tracking data by fuel and segment were inspected prior to modeling for missing values, seemingly erroneous data or outliers, and high and low end values that might skew the sample statistics or suggest multi-modal distributions. Other adjustments to the data sets were made including the use of a "restricted" commercial sector data set that included only counties with high ex ante energy savings values in this pilot test. Dropping sites from statistical analyses that likely provide no information because the expected savings from those sites are so small is not uncommon. The usual justification is that the total savings number is not likely to be influenced by their exclusion since the expected savings were so small.

- The residential sector savings estimates from the estimated MCM model for PG&E and SDG&E (SCE was not estimated) were substantially higher than the utility's ex ante values.

When these sector level results are aggregated up to a state-wide number, the wide discrepancies at the sector level tend to off-set each other. It is important to recognize that this was a pilot effort and different people will have different views on the overall robustness of this Pilot Study's findings at the sector and state-wide level.

Developing Top-Down Models

Cadmus (2012) and Demand Research (2012) took different paths to developing a top-down MCM model for this California Pilot Study. Both study teams concluded that the work to date indicated this was a potentially useful research path for developing state-wide estimates of energy savings attributable to EE policies. In its discussion of findings, Cadmus discussed the potential applications of these methods:

- Top-down macro-consumption methods could yield inexpensive[73] estimates of energy savings from utility energy-efficiency programs and building codes at an aggregate level.
- These methods are attractive because it is possible to produce confidence and precision levels for the net energy savings estimates, something that is not easily accomplished in bottom-up evaluation studies.
- Top-down studies can be used to verify state-wide energy-efficiency program savings estimates based on bottom-up evaluation by looking at aggregate energy consumption data.
- These methods can be useful in tracking a state's progress in reducing greenhouse gas emissions and to develop forecasts of energy savings from future program spending at an aggregate level.

The next steps might be to: (1) replicate the results of Cadmus and Demand Research, LLC using the datasets already developed, and (2) continue improving the data platform[74] used for these analyses -- both studies contained recommendations on how the data could be improved. Other questions might pertain to the sensitivity of the results to model specification, i.e., the robustness

[73] Both pilot studies ran into data problems that would have to be overcome in future work and there would be a decent price tag associated with this work. If the alternative were to build up state-wide estimates by doing measure-specific engineering analyses, this aggregate approach would be cheaper; however, bottom-up methods performed cost-effectively are probably needed for program support, design and verification of savings at the program level. The issue is whether the incremental information provided by these aggregate studies has a value greater than its cost. That may vary by jurisdiction.

[74] Violette et al. (2012) discusses the importance of the data platform on which these top-down models are estimated.

of the results under a designed set of alternative specifications also consistent with the theory and appropriate econometric methods.[75]

It seems unlikely that bottom-up studies would be entirely replaced by these top-down methods. As discussed earlier, there is likely a need to have program-level (and some measure-level) assessments to ensure that a program's design will result in a program meeting its specified targets. As a result, the question may be, "Does the incremental value of the information produced by the top-down methods exceed the cost of the work?" At the national level, data on an adequate level of observations is more easily available. For state level studies, more work will be involved in setting up the databases and disaggregating the data into the number of needed cross-sections which may introduce a certain amount of error into these observations.[76]

3.6 Structured Expert Judgment Approaches

Structured expert judgment approaches involve assembling a panel of experts who have a good working knowledge of the technology, infrastructure systems, markets, and political environments. This approach is one alternative for addressing ME in different end-use markets. These experts are asked to estimate baseline market share for a measure or behavior. In some cases, they are also asked to forecast market share with and without the program in place. Structured expert judgment processes use a variety of specific techniques to ensure that the panel of experts specify and take into account key known facts about the program, the technologies supported, and the development of other influence factors over time (Tetra Tech, Inc., KEMA, NMR Group, Inc., 2011).

The Delphi process is the most widely known technique (NMR et al., 2010). Using this process, each panelist is asked to make a judgment on the topic—based on the provided information and on their experience—and submit the information back to the evaluators. The evaluators compile the information from the panelists and resend it to the panelists for another review. The panelists are asked whether they stand by their original judgments or whether the assessments of their peers have caused them to alter their judgments. At least two rounds of judgment are required for a Delphi panel, although more rounds can be used.

Some of the advantages of the structured expert judgment approach are:

- The estimate is based on feedback from a group of experts, which can be particularly useful for programs with complex end-uses.
- It is a useful tool for consolidating results from multiple methods to develop a consensus estimate (see Example 2 below).

---

[75] This sensitivity analysis might examine the stability of the estimates under alternative functional forms, inclusion of one or two variables, testing of interaction terms, and tests on subsets of the data.

[76] Violette (2012) in Chapter 13 of these DOE Uniform Method Protocols discusses attenuation bias where the coefficients on independent variable can be biased towards zero due to errors in the measurement of variables. A similar effect is shown in Ridge (1997).

As with other approaches (such as market sales data analysis), the structured expert judgment method relies on high-quality data to inform the panel, so a lack of these data can result in inaccurate estimates of net savings (NMR Group, Inc. and Research Into Action, Inc. 2010).

Two examples of using the structured expert judgment approach to estimate net savings are presented below. The first is a discussion of how Delphi panels were used to estimate net savings for a residential new construction program in California. The second example shows the development a final estimate through the use of a Delphi panel's review of estimates derived through other methods.[77]

EXAMPLE 1: Residential New Construction Delphi Panel

A report prepared for the California Public Utilities Commission Energy Division describes in detail how evaluators used two Delphi panels of Title 24 consultants and building industry experts to convert the gross savings estimates from investor-owned utility programs targeting the residential new construction sector to net savings estimates. (The report was prepared by KEMA, NMR Group, Inc., Itron, Inc., and The Cadmus Group, Inc. 2010.)

The panelists received detailed data pertaining to code compliance, compliance margins, and estimates of annual gross energy savings in non-program homes at the state level and by climate region. After reviewing these data, panelists were asked to:

- Estimate the proportion of the electricity and natural gas savings attributable to the investor-owned utility (IOU) programs targeting the residential new construction sector and other factors (e.g., non-IOU RNC programs, the economy/housing market, energy prices, and climate change).
- Estimate the percentage of net savings in non-program homes attributable to different IOU program elements (e.g., builder trainings, incentives, and design assistance)
- Assess the extent to which the market effects are likely to persist in the absence or reduction of the IOU programs.
- Estimate the percentage of homes that would have been below-code in the absence of the IOUs' programs and other factors and estimate the compliance margin of the below-code homes in the absence of each factor.

[77] An application of the Delphi technique as applied outside of EE may be informative. Navigant (2013) conducted an evaluation of the Wind Power America program. The goal was to assess the impacts attributable to the program. The unique aspect of this Delphi exercise was that the use of range estimates, i.e., experts were asked about lower and upper bounds to the effects as well as a best estimate. This approach allowed for the experts to provide their own insights into the uncertainty of the estimates. Gauging uncertainty and then using that in probabilistic and scenario analyses is consistent with other utility resource planning activities. Adapting these methods to EE resource assessment may increase the usefulness of the information.

Each panelist completed two rounds of detailed surveys and, in the second round, they were provided a comparison with other panelists' responses and logic and allowed to change their answers.

EXAMPLE 2: Lighting Program Delphi Panel
Another way to use a Delphi panel is to have the panel review estimates derived through other methods to develop a final estimate. As part of the evaluation of the Massachusetts ENERGY STAR Lighting Program (NMR Group, Inc., KEMA, Cadmus Group, Inc., Tetra Tech, 2011), the evaluation team used a Delphi panel of lighting and EE experts across the country and Canada. The goal for the panelists' efforts was to integrate results from five methodologies that yielded NTG estimates (conjoint analysis, multistate modeling, revealed preference study, supplier interviews, and a willingness-to-pay study). The evaluation team then used the Delphi panel's review results in developing recommendations for the final NTG estimate.

3.7 Deemed or Stipulated NTG Ratios

Deemed or stipulated NTG ratios are pre-determined values and do not rely on a calculation-based approach. Deemed values are often based on previous NTG research that has been conducted using at least one of the other methods described above.

NTG ratios are often stipulated when the expense of conducting NTG ratio analyses cannot be justified or when the uncertainty of the potential results is too great to warrant a study. A recent review of 42 jurisdictions in the United States and Canada (which represented the vast majority of jurisdictions with ratepayer-funded energy efficiency programs) found that only 14% use a deemed approach to NTG (Navigant, 2013).

Deemed or stipulated NTG ratios are typically either set by a regulatory agency or negotiated between regulators and program administrators. These ratios may be determined at the portfolio level (e.g., Michigan and Arkansas)[78] or on a measure-by-measure basis (e.g., California and Vermont).[79] Typically, the ratios are based on NTG studies from past evaluations and/or reviews of other similar programs in which an NTG ratio was estimated. For example, it is not unusual in a multiyear portfolio cycle to estimate an NTG ratio for an initial year (or possibly every other year) with deemed values used in the subsequent or intervening years.

In other cases, evaluators use historical data or other information from a wide range of sources to develop a "weight of evidence" conclusion regarding the program's influence (SEE Action,

[78] Arkansas, NTG deemed at 0.8 - http://www.apscservices.info/pdf/07/07-085-tf_286_44.pdf; Michigan - NTG is deemed at 0.9 for all programs except pilot, education, and low-income programs which are deemed at 1.0. http://efile.mpsc.state.mi.us/efile/docs/17138/0009.pdf. Note that most low-income programs are not subject to NTG analysis (i.e., are deemed at 1.0).

[79] California - http://www.energy.ca.gov/deer/; Vermont - http://www.efficiencyvermont.com/docs/about_efficiency_vermont/annual_reports/2011_Gross_to_Net_Report_EfficiencyVermont.pdf

2012b). As discussed earlier, one common approach for developing a stipulated value is to use a panel of experts who have the relevant experience to make that judgment (e.g., Delphi panel).

While using deemed or stipulated values is a relatively simple and low-cost approach, there are a number of disadvantages. NTG values are variable across time and space, and strongly linked to program design/implementation making deemed values or assumptions potentially unreliable when transferred from a program in one jurisdiction to a program in another jurisdiction.[80] NTG values based on primary research efforts can produce estimates that are based on program-specific information (NMR et al., 2010). As a result, these values provide useful information for the future program design and implementation of programs[81] and may mitigate the risk to rate-payers from utilities receiving performance incentive payments on savings not actually attributable to the program (as well as the risk to ratepayers of making performance incentive payments that are too large). NTG values are also critical from a resource planning perspective and having better data on the actual energy savings achieved from energy efficiency programs can help the planning process (Navigant Consulting, 2013). Deemed or stipulated NTG values do not provide these benefits.

EXAMPLE 1: California Public Utilities Commission DEER database
The California Public Utilities Commission uses deemed savings (listed in its Database for Energy Efficient Resources) for planning purposes and interim savings estimates for its programs. These deemed savings are updated based on results of NTG studies.

3.8 Historical Tracing (or Case Study) Method

This method involves reconstructing the events (such as the launch of a product or the passage of legislation) that led to the outcome of interest. An example of this would be developing a "weight of evidence" conclusion regarding the specific influence or role of the program in question on the outcome.

Historical tracing relies on logical devices typically found in historical studies, journalism, and legal arguments (Rosenberg and Hoefgen, 2009). These include:

- Compiling, comparing, and weighing the merits of narratives of the same set of events provided by individuals who have different points of view and interests in the outcome;
- Compiling detailed chronological narratives of the events in question to validate hypotheses regarding patterns of influence;
- Positing a number of alternative causal hypotheses and examining their consistency with the narrative fact pattern;

[80] Another issue raised by a reviewer was that the use of deemed NTG values can remove the incentive for the program administrator to reduce FR and maximize SO and ME to yield greater net savings values.

[81] For example, freeridership can inform decisions to discontinue incenting certain measures, increase incentive amounts, or increase the efficiency level being incented.

- Assessing the consistency of the observed fact pattern with linkages predicted by the program logic model; and
- Using information from a wide range of sources (including public and private documents, personal interviews, and surveys) to inform historical tracing analyses.

This method is best suited to an attribution analysis of major events, such as adoption of new building codes or policies. It is not typically applicable to EE programs. However, various elements of this approach may be used in the analysis of very large custom projects that essentially require case study approaches.

While this method draws from multiple information sources, it is difficult or impossible to determine the magnitude of the effects, so the evaluator cannot assign statistical precision to the estimate (NMR et al. 2010). However, as part of making a persuasive case for attribution and providing evidence supporting a statistically derived net savings estimate, this method can be very important. Statistics alone often are not a complete attribution assessment. They often require context using supporting logic to enhance the validity of the statistical estimates.

EXAMPLE 1: Historical Tracing for a Residential New Construction Program
Keneipp et al., (2011) used historical tracing in conjunction with Delphi panels to develop energy savings for new homes. This study used historical tracing spanning 14 years of regulatory documents to create timelines of the residential new construction program presence and activities for Arizona Public Service Company. Using these data, the evaluators created an influence diagram of market influences on specific building practices. This information was then shared with two in-person Delphi panels of market experts, who estimated the percentage of homes built in 2010 using specific building practices. These Delphi panels also developed the counterfactual scenarios that were used to show the net impact of the residential program on percentage of homes that were built to standards that would not have been met in the absence of the program. The Delphi outputs were then used to develop inputs for an engineering simulation model to calculate energy savings per home.

4 Conclusions and Recommendations

A central theme in this chapter is that all decisions have an implicit counterfactual, i.e., what would have happened if the decision had not been made. In the context of EE program investments, net savings are the savings that are attributable to the program, i.e., they would not have occurred if the program had not been offered. This chapter presents a number of approaches for addressing attribution and the net impacts resulting from EE programs. The conclusions and recommendations discuss issues affecting the choice of a net savings approach within an evaluation context.

4.1 A Layered Evaluation Approach

It is important that the approach selected be appropriate for the intended audience and that it presents analyses supported by evidence. A great statistical analysis will not be persuasive to many decision makers and stakeholders on its own. All approaches should be supported by a

narrative discussing why a specific approach was taken, the appropriate interpretation of the findings, and the larger picture of identifying net savings. The narrative and analysis should also recognize and indicate the uncertainty in net savings determination. Developing an appropriate narrative often leads to the application of layered methods of analyses.

Studies examining net savings from EE programs may contain both sophisticated quantitative analyses, as well as intuitive analyses that show that savings attributable to the program do in fact exist. A compelling part of the narrative can be a simple case study of one or two market participants. A case study can show with a very high degree of internal validity that net savings were obtained, and/or provide examples of NTG factors including FR, SO, and ME. An intuitive case study often is a useful first step in an analysis framework for addressing estimates of net savings. A framework might include:

> Part 1 – Establish the existence of the effect, possibly using a case study approach. This could be the establishment of the existence of savings attributable to the program or, if the focus of the research is on estimating FR or SO, the first step might be to establish the existence of these effects.[82] Once existence of an effect is established, the next step involves determining the magnitude of the effect. This can be easier when the audience has been convinced that the effect exists (it is non-zero), and the logic behind the attribution of the effect is set out.
>
> Part 2 – This involves the extrapolation of the findings of the case studies to the more general participant population. Once the logic of the case studies has been established, it is often possible to define and apply a statistical model consistent with this logic, or to develop an alternative approach to extrapolate the effect. This approach might be any of the methods discussed in this chapter – survey methods, common practice baselines, market data analyses and comparisons, structured expert surveys, or historical tracing to examine the influence of a program over time.

The framework discussed above might utilize three steps:

1. An initial "high internal validity study" to prove existence.
2. Utilize methods to establish a range estimate, i.e., what might be a reasonable lower bound for the impacts and what might be the highest reasonable value. This provides information on the importance of the effect being studied whether it is net savings or an NTG factor such as FR, SO, or ME.

[82] If the study is attempting to estimate the amount of spillover resulting from a program, the first step might be to isolate one or two case studies that compellingly show that SO exists at participating commercial sites.

3. Once the range is established, analyses can be performed using methods presented in this chapter to develop the best estimate of impacts within that range.[83]

4.2 Selecting the Primary Estimation Method

The selection of appropriate net savings analyses methods will depend in part on the questions that are to be answered by a net savings study. The questions to be answered and goals of the analysis will provide insights regarding the appropriate net savings approach to be used. Research questions that have implications for the net savings approach include:

1. What are the components of net savings to be estimated by the program? Candidate approaches would include:
    a. Random control trials and quasi-experimental designs employing DiD and regression methods along with RDD and RED designs are discussed in section 3.1 of this chapter. These approaches will capture net savings that addresses FR and participant SO. Nonparticipant SO is not directly addressed, but can be addressed through surveys of nonparticipants and ME studies with trade allies.
    b. Survey methods that can be used to adjust engineering-based gross savings estimates for FR and participant SO are discussed in section 3.2. Like the methods in "a." above, non-participant SO can be addressed through surveys of nonparticipants and ME studies using trade allies.
    c. Broader-based methods such as market sales, structured judgment, and historical tracing analyses discussed in sections 3.4, 3.6, and 3.8 respectively all can be used to provide program specific net savings estimates, and can address SO and ME.
    d. Common practice baseline methods discussed in section 3.3 can produce estimates by developing baselines on a program basis. This approach may not fully address FR or participant SO as it does not account for self-selection bias, and it does not directly address nonparticipant spillover. However, as noted above, nonparticipant SO can be addressed through surveys of nonparticipants and ME studies with trade allies. Common practice baseline methods might be viewed as a compromise that balances out over- and under-estimated NTG factors in the net savings estimate.
    e. Deemed or stipulated discussed in section 3.7 methods can, of course, be set at the program level.
    f. Top-down analyses discussed in section 3.5 use aggregate data that represents the overall level of EE effort across all programs, but are not able to isolate the effects of a single program.

[83] In a survey setting, this approach can help the survey respondent consider the behavior that might result in lower and then higher impacts that might have been achieved if the program had not existed. The thought process developed by this three-step construct can help survey respondents produce better estimates of their most likely behavior.

2. How can estimates of net savings on a program basis be combined with information on program implementation effectiveness? Approaches that provide estimates of net savings but also include elements that involve gathering information directly from participants, nonparticipants, and trade allies can be useful for improving program performance. For example, some programs are designed to minimize FR to improve overall resource effectiveness, and some programs may focus on expanding the magnitude of SO and ME. For these programs, specific estimates of FR, SO, and ME – particularly if they are provided over time (e.g., every two years) – can be used to assess overall program effectiveness.
3. Can estimates of aggregate net savings from a portfolio of programs be produced? All of the approaches can produce program specific estimates that can be aggregated up to the portfolio level. However, top-down methods are designed to work with aggregate data particularly at the regional level. Top-down models are the one approach that conceptually addresses all the NTG factors – FR, SO and ME.

Other factors that influence the selection of appropriate methods will vary by program type, delivery, sector, and maturity. Key elements for consideration when choosing a method are described in a recent FR and SO methodology study for the Massachusetts Program Administrators (Tetra Tech, Inc., KEMA, NMR Group, Inc., 2011). This study addressed factors as summarized below:

- Availability of market sales data with meaningful comparison group. If market sales data are available on the total sales of both efficient and standard equipment over time and if these data are available for the program area and if there is an appropriate comparison area for the appropriate time period, then total program effects may be estimated based on these data.

  The ideal strategy is to compare the magnitude of the change in sales of energy-efficient equipment relative to the sales of standard equipment in the program area and the comparison area. However, the program itself tends to produce systematic differences between the program and control areas. Therefore, where a program has been operating for a long period of time, it is very difficult to find a suitable comparison area.
- Homogeneity of the measure and the consumers. Random control trials and quasi-experimental designs work best when there are a large number of similar consumer types and measures. Since large custom programs are likely to have fewer projects, it is possible that a few (or even one) very large project(s) could have a significant influence on the FR or SO rate. Therefore, the evaluator should use multiple approaches that allow for a greater focus on those consumers that drive the overall impacts to confirm the findings for that program. Methods based on market data or samples of consumers who are making similar purchase decisions may not apply to programs with custom measures.
- Likelihood of substantial upstream effects unknown to end-use participants. If there is a reasonable likelihood of substantial upstream effects that an end-use participant would

not know about, then conducting an evaluation by using participating end-user surveys alone will tend to understate the effect of the program (even if consumers answer accurately from their perspective). These situations require either information for the market as a whole (if the market sales-based approach is viable) or a combination of participant end-user and vendor surveys. For example, the participating customer would not know that the program influence has changed what options are available, lowered the price of the efficient options, and/or increased the sales staff's knowledge of and/or interest in promoting the efficient option.

- Cost/value trade-offs. Some methods that provide more credible results are more costly. This cost may be justified for those program components that are important to the portfolio, but not for all components. Importance to the portfolio is typically related to the level of spending and/or savings associated with a program component. However, a component's importance may also depend on future program plans or other "visibility" factors. The systematic assessment of the value of information gained by net savings estimation approaches as compared to the cost of the research is needed to better balance the requests for confidence and precision of the estimates. A target of 90 percent confidence and 10 percent precision simply may not be reasonable for all but the largest programs in a portfolio. This systematic approach can examine the impacts on rate payers from incorrectly attributing savings to a program. If it is a small program, the impacts on rate payers even when incentives are being paid will be small from going to 90 percent confidence and 15 or 20 percent precision using a one tailed test. This can substantively reduce evaluation costs with little impact on the overall equity tradeoffs between ratepayers and utilities.
- Data quality. Data quality is a critical factor for all methods. Typical examples of potential data quality limitations are: (1) insufficient information in program tracking databases, (2) lack of clear definitions of what is contained in tracking systems (e.g., a data dictionary), (3) limitations on the availability of nonparticipant data, including billing data, and (4) limitations on the number of years of available billing data for participants.

4.3 Methods Applicable for Different Conditions

Methods that are suitable for programs with particular features are summarized in Table 1.[84] For a program in a particular context, choosing the appropriate methods requires balancing the advantages and disadvantages of each. Thus, this table does not attempt to prescribe a best method for a given situation. Rather, it indicates which methods available in the practitioners' toolkit are applicable to programs with the indicated features. This table focuses on information available for individual consumers. It does not address top-down methods which work with aggregated data across a large number of geographic cross-sections.

[84] This is based on Tetra Tech et al. (2011) prepared for the Massachusetts program administrators.

Table 1. Summary of methods applicable to different conditions

Surveyed Group	Net Savings Data Collection Method	Applicability				Typical Cost or Complexity
		Custom Measures	Measures with few, diverse participants	Large numbers of similar participants	Measures with substantial upstream influence invisible to consumers	
None (with defined participant and control group or matched nonparticipant comparison group	Energy consumption data for participants and nonparticipants	Poor	Poor	Good	Poor	Low
None (with defined market segment)	Market sales data for program area and comparison area	Poor	Poor	Good	Good	Low if data are available, High or not possible if data must be developed
Participating End-users	Survey based counterfactual	Good	Good	Good	Poor unless combined with retailer or contractor surveys	Medium
Nonparticipating end-users	Survey based	Poor	Poor	Good	Poor unless combined with retailer or contractor surveys	Medium-High
Manufacturers & Regional buyers and distributors	Market sales data	Poor	Poor	Good	Good	Low
Retail store managers and contractors	Sales data	Poor	Poor	Good	Good	Medium
Retail store managers and contractors	Sales and survey based counterfactual	Poor	Poor	Good	Good	Low
Retail store managers and contractors	Survey based (Customer-specific influence)	Good	Good	Medium	Good	Medium
Participating consumers and comparison group consumers	Billing data analyses with control variables	Poor	Poor	Good if have valid comparison group	Good if have valid comparison group	Low
Experts	Survey based, Delphi panel	Depends on quality of input methods				Low
None (Aggregate data available on geographic cross-sections)	Top-down methods for regional application	Requires data on aggregate energy consumption and information on EE effort (expenditures or related program variable) for a large number of cross-sectional observations over a period of time.				Depends on the cost of compiling the initial data set

4.4 Planning Net Savings Evaluations – Issues to be Considered

Evaluation planners must consider a number of practical issues when planning a net savings evaluation. These include: the use of the information, maturity of the program, the timing of the study, the frequency of net savings estimation, and whether multiple approaches should be used. The summary below provides direction when considering these issues:

- Use of the information. It is important to consider how the results of the net savings evaluation will be used and the audience for which the evaluation is intended. This use might include shareholder incentives, resource plans, program design, and environmental targets (e.g., carbon emissions) among other policy goals.[85]

- Maturity of program. Most all programs are assumed to have some freeridership. The conventional wisdom is that as the program matures (all else equal), FR will increase, but so will spillover and market effects. As a result, it becomes important to test for the existence of spillover and market effects as a program matures.

- Timing – when to collect data. To estimate freeridership, the timing of the data collection should occur as soon as possible after program participation. This timely measurement minimizes recall bias (Baumgartner. 2013), provides timely feedback on program design, and reduces the possibility that the key decision maker and/or market actor are no longer available. However, if the objective is to estimate spillover, the ideal time to collect data would be at least one to two years after program participation, as this allows sufficient time for spillover to have occurred. Finally, if the objective is to estimate market effects, then regular data collection over a period of time is required.

- Frequency of net savings estimation. The frequency of net savings or NTG analyses depends on the use of the information. If it is a component of financial incentives for a program administrator, there may be a need to conduct these studies more frequently. Usually, there is no need to perform these detailed net savings studies more frequently than every other year. But, it also depends on the methods used. A statistical analysis of a residential behavioral program may be estimated every year since persistence is an important issue and the costs of the study are not high. Northeast Energy Efficiency Partnerships recommends that net savings estimates be made every two to five years (Titus et al., 2008) as there are a number of factors that can make estimates of net savings change over time.

- Triangulation of NTG approaches. Using data from multiple sources described in this document limits the effects of self-report bias and measurement error (Baumgartner,

[85] For example, the regional net savings research project (NEEP, 2012) showed that "compared to New England and New York, states in the Mid-Atlantic more commonly use evaluated gross savings for utility regulatory compliance and net savings for program planning and measurement of cost effectiveness. In contrast, New England and New York are more likely to use evaluated net savings; in doing so, they apply NTG values prospectively rather than retrospectively."

2013). Using an in-depth methodology with multiple sources also allows evaluators to weight the value of responses from different decision-makers (Megdal et al., 2009). Other data sources often used are: (1) interviews with key decision-makers at the site; (2) project file reviews or project analysis that looks at barriers to project installation, how the project addressed those barriers, and documentation on the participant's decision to go forward with the project; and (3) market data collection, which might include analyses of market sales, shipping data and surveys of market actors (GDS Associates et al. 2010) (SEE Action, 2012b).

4.5 Trends and Recommendations in Estimating Net Savings

As discussed above, the choice of approach for estimating net savings will vary depending on the questions that are being asked, the characteristics of the program(s) being evaluated, and the ultimate use of the data. However, there are trends in the application of methods:

1. The expanded use of informational and behavioral EE programs is leading to a greater use of random control trials and quasi-experimental designs that employ some form of randomization (RDD or RED) to help address self-selection.
2. The complexity of programs and the need for assessing market effects is leading to a greater use of informed expert panels and Delphi-types of analyses.
3. The need to examine trends in program performance over time and impacts on markets over time is resulting in longer-term planning for net savings and NTG factor analyses, e.g., regular studies conducted periodically with panel data.
4. Net savings studies are increasingly embedded in survey analyses that are also designed to gather information on program implementation effectiveness.
5. The value of information from net savings studies is being considered in a more structured manner to help manage evaluation costs. Achieving 90 percent confidence and 10 percent precision may be important for a very large EE program, but for a program that is one tenth of the size of the largest program, precision levels are being generated that represent only one percent of the large program. Also, one tailed tests as targets should be more commonly considered as the importance of attaining a threshold level of attribution is likely to be more important information than is the likelihood that a program has exceeded the target by a given amount. Note that a one-tailed targeted precision level still allows for the calculation of the upper end of the confidence interval (Navigant, 2012) and there is value to knowing if there was a high likelihood that the target was exceeded by a given amount. The appropriate level of confidence and precision targets are now often reviewed by both EE program administrators and regulators to provide fair attribution estimates that protect risks to both ratepayers and utilities that may be receiving incentives.[86]

[86] Navigant (2013) discusses a loss function approach for assessing the value of information from net savings studies; and Navigant (2012) presents information on sampling and the tradeoffs between confidence and

It has always been important to consider evaluation options prior to implementing an EE program or portfolio of programs. However, the importance of planning the types of net savings studies that are needed and the frequency of this measurement prior to program implementation are becoming critically important. Net savings studies embedded in experimental designs that are established prior to consumers becoming program participants allow for:

- The consideration of randomized designs,
- The development of the data platform for estimating consumption based models including top-down models, and
- The collection of information needed for well-run structured expert panel studies.

In conclusion, net savings methodologies continue to evolve and improve over time. No one methodology is appropriate for all programs or measures, and a single methodology is often not the best choice for estimating program or measure net savings. In the end, it is recommended that the evaluation plan to assess net savings be designed in conjunction with the key stakeholders keeping the following in mind:

- The appropriate schedule for the evaluation effort over time taking into account the expected value of the information produced versus the cost of the research effort.
- Program design and maturity
- The contribution of the program to overall portfolio savings (past, current, planned)
- The evaluation budget, objectives, and value
- Observations and leanings from other jurisdictions

Finally, adequately documenting the methods used and effectively communicating the results of any net savings study is important. A framework for persuasive communication was presented at the beginning of this chapter.

---

precision.

Appendix A: *Price Elasticity Studies as a Component of Upstream Lighting Net Savings Studies*

Studies of upstream changes in the price for residential lighting products have been receiving attention as a way to complement surveys with market actors, or even replace these surveys, with econometric models. To date, these studies have focused on residential lighting products and, within that category, mostly on CFL sales. Two methods have been demonstrated – a demand equation approach in (Cadmus 2012 and 2013), and KEMA (2010) tested several different methods for estimating the increase in CFL sales resulting from program-induced lower prices due to program activities – coupons and markdowns negotiated with retailers.

Cadmus (2012) examined Efficiency Maine's residential lighting program using a price elasticity approach -- "a variety of specifications to ascertain price impacts (the primary instrument used by the program) on CFL demand." The basic concept is to estimate expected bulb purchases (and associated savings) at prices offered under the program and then the sales that would have occurred at original retail prices. The difference between these two values is viewed as net savings in this study.

The approach used in Cadmus (2012 and 2013) was a single equation regression model where the quantity of CFLs purchased was a function of the price of CFLs and a select set of other independent variables. The data used to estimate this equation included package and bulb sales for each retailer, by model number and by week. The data set does not include information on the customers that purchased the CFLs, but does contain information on quantities of CFLs sold and prices by retailer. Customer variables that one might like to see in a demand equation would include income and education, but often these variables are not available in a retailer sales tracking system.

Cadmus (2012 and 2013) used similar approaches. A regression was estimated relating quantities of CLFs sold by retailer to the price of CFLs that week for each retailer. Other factors were considered in determining consumer purchases such as promotional events. Programmatic factors such as labeling and information dissemination are pervasive throughout the lighting programs and, while potentially important, could not be addressed due to lack of variation across consumer purchases.

These two studies showed an increase in the sales of CFL blubs as prices declined due to markdowns negotiated with retailers and discount coupons provided to consumers. The second step of the approach involved estimating what the sales would have been at the higher prices that would have prevailed without the program, i.e., the counterfactual.

Considerable effort is made in these price elasticity studies to control for other factors that might also affect CFL sales other than price, but it is difficult to show that any method is free of bias. In the case of the Efficiency Maine lighting program, there were three components to the program.

Two were linked to price (markdowns and coupons) and a third was linked to overall participation in the Appliance Rebate Program, "with Appliance Rebate Program participants electing to receive a free six-pack of CFL bulbs, via a check-off on the Appliance Rebate Program application form." This third part of the program would provide CFLs at essentially a zero price and it is not clear how this would be factored into the analysis.

Cadmus (2012) presents several general caveats to the demand equation approach used in the study. First, it was acknowledged that "this estimation method has rarely been used in upstream lighting program evaluations, as such data generally have been unavailable. As Efficiency Maine … tracked these data and shared them for this evaluation, Cadmus found such econometric demand estimation provided the best method for estimating the program's freeridership." Second, Cadmus (2013) indicates that it "will continue to look for alternative methods to calculate net-to-gross," and that the model used for the … 2012 evaluation does not account for spillover.

KEMA (2010) used price variables to estimate net savings in an upstream lighting study. This study had the benefit of a sizeable data collection effort. As part of the in-store consumer intercept research, brief interviews were conducted with shoppers who had just made a lighting purchase (revealed preference) as well as "stated preference" surveys with other consumers recruited randomly. 1,463 customer intercept surveys were completed across 378 stores.

KEMA (2010) used three primary types of methods for estimating net savings:

- Supplier and consumer self-report methods;
- Econometric models (e.g., pricing/conjoint elasticity models, revealed preference purchase models, stated preference purchaser elasticity models); and,
- Total sales (market-based) approach

Among the econometric modeling efforts, four different econometric models were used:

1. Pricing (price formation model)
2. Conjoint Elasticity
3. Revealed Preference Purchase
4. Stated Preference Purchaser Elasticity

The first two econometric methods – pricing formation and the conjoint elasticity model are both needed to produce a net savings estimate. Revealed preference and stated preference models can produce net savings directly. As a result, there were four econometric models, but only three different approaches for estimating net savings.

The price formation model estimates the percent reduction in CFL prices that resulted from program incentives. This is combined with the conjoint analysis which estimates the corresponding percent increase in market share/sales that result from a price decrease. This

allows for net savings to be calculated by combining the findings from the pricing study with the conjoint demand elasticity study, i.e., the program-induced reduction in price times the estimated increase in sales due to the lower price.

Revealed preference survey respondents are limited to customers that actually purchased CFLs. These customers were asked to indicate how many CFLs they would have bought compared to their actual purchases at double the price they actually paid. Response categories were the same amount, fewer, or none. While still based on hypothetical, self-reported responses, revealed preference respondents may be a more reliable sample because they just made an active purchase decision (as compared to stated preference respondents, i.e., a random sample of customers). Revealed preference respondents, however, may be somewhat unlikely to indicate they would have paid more for what they just purchased. KEMA (2010) states that the magnitude of the potential bias is unknown, "but it is likely that NTGR estimates from stated preference respondents are biased downward and from revealed preference respondents are biased upward."

The revealed preference model developed by KEMA utilized survey data collected through the in-store consumer intercepts to model CFL purchase rates with and without the effects of the program. This model was based on a logistic regression to model the probability of buying a CFL rather than an "equivalent" non-CFL as a function of price, displays, customer characteristics, and bulb characteristics, by channel. The fitted models were evaluated under "program" and "non-program" conditions. For each channel, the difference between the probability of purchasing CFLs under the program condition and that under the non-program condition is the CFL sales share attributable to the program.

In summary, the price elasticity studies completed to date have addressed residential lighting. Cadmus (2012 and 2013) developed a demand model specification based on an examination of alternative specifications. KEMA (2010) develops several different approaches for examining the change in CFLs sold as a function of program induced lower prices. KEMA (2010) concludes that from the econometric approaches, the revealed preference model was viewed as the preferred approach. It should be noted that these approaches focus on freeridership and do not address SO or longer term ME. However, work is continuing on the use of price elasticity to produce net savings estimates in residential lighting. An expanded literature will likely provide additional confidence in this method for addressing FR from upstream lighting programs, and possibly an expansion to other residential product programs.

References

Abadie, A. and G.W. Imbens (2011). *Bias-Corrected Matching Estimators for Average Treatment Effects.* Journal of Business and Economic Statistics Vol. 29, No. 1.

AEP Ohio, (2012) Appendix H, Evaluation of Home Energy Reports, prepared by Navigant. May. http://dis.puc.state.oh.us/TiffToPDf/A1001001A12E15B14941H23668.pdf

Agnew, Ken, Goldberg, Mimi (2013). Chapter 8: "Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol." *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measure*s. http://www1.eere.energy.gov/wip/pdfs/53827-8.pdf.

Angrist, Joshua D. & Jorn-Steffen Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ: Princeton University Press.

Arimura, Toshi H., Shanjun Li, Richard G. Newell, and Karen Palmer, (2011). *Cost Effectiveness of Electricity Energy Efficiency Programs.* Resources for the Future Discussion Paper 09-48-Rev.

Auffhammer, Maximilian, Carl Blumstein, and Meredith Fowlie (2008). *Demand-Side Management and Energy Efficiency Revisited.* The Energy Journal 29 (3), 91-103

Baumgartner, R. (2013). *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures,* Chapter 12: Survey Design and Implementation Cross-Cutting Protocols for Estimating Gross Savings. http://www.nrel.gov/docs/fy13osti/53827.pdf.

BC Hydro (2012). *Review of a Top Down Evaluation Study: Rivers & Jaccard*. Prepared for BC Hydro, by Navigant Consulting, Inc., April.

Bradlow, E. (1998). *Encouragement Designs: An Approach to Self-Selected Samples in an Experimental Design.* Marketing Letters, Vol. 9, Vol. 9, No. 4, November.

Cadmus Group, Inc. (2012). *CPUC Macro Consumption Metric Pilot Study: Final Report*. Prepared for the California Public Utilities Commission, October.

Cadmus Group, Inc. (2012). *Efficiency Maine Trust Residential Lighting Program Evaluation: Final Report.* Prepared for the Efficiency Maine Trust. http://www.efficiencymaine.com/docs/Efficiency-Maine-Residential-Lighting-Program-Final-Report_FINAL.pdf

Cadmus Group, Inc. (2013). *Focus on Energy Calendar Year 2012 Evaluation Report*,

Volume II and Appendices A though O. Prepared for the Public Service Commission of Wisconsin. http://www.focusonenergy.com/sites/default/files/FOC_XC_CY%2012%20Report%20Volume%20II%20Final_08-28-2013.pdf

Cadmus Group and Navigant (2012). *New York Energy $mart^{sm} Products Program Market Characterization and Assessment Evaluation: Final Report.* Prepared for The New York State Energy Research and Development Authority, Victoria Engel-Fowles Project Manager, Project Number 9875, February. http://www.nyserda.ny.gov/BusinessAreas/Energy-Data-and-Prices-Planning-and-Policy/Program-Evaluation/NYE$-Evaluation-Contractor-Reports/2012-Reports/Market-Analysis.aspx

Cadmus Group, Navigant, Opinion Dynamics Corporation (2012). *2012 Residential Heating, Water Heating, and Cooling Equipment Evaluation: Net-to-Gross, Market Effects, and Equipment Replacement Timing*. Prepared for the Electric and Gas Program Administrators of Massachusetts.

Castor, Sarah (2012). *Fast Feedback Results.* 2011 Final Report prepared for EnergyTrust of Oregon. http://energytrust.org/library/reports/Fast_Feedback_-_20110.pdf

Commonwealth Edison Company (2012). *Home Energy Reports Evaluation*. Prepared by Navigant. http://www.icc.illinois.gov/downloads/public/edocket/323839.pdf

Cook, T. et. Al. (2010). *Contemporary Thinking About Causation in Evaluation*, American Journal of Evaluation, 31:105. http://aje.sagepub.com/content/31/1/105

Demand Research, LLC, (2012). *Macro Consumption Metrics Pilot Study: Final Repor*t. Prepared for: California Public Utilities Commission Energy Division, November.

Diamond, A. and J. Haninmueller (2007). *The Encouragement Design for Program Evaluation*. Harvard University and International Finance Corporation. http://www.docstoc.com/docs/5419170/The-Encouragement-Design-for-Program-Evaluation-September-Alexis-Diamond

EPRI (1991). *Impact Evaluation of Demand-Side Management Programs: A Guide to Current Practice*. EPRI CU-7179, February.

Energy Valuation Organization, International Performance Measurement and Verification Protocols (IPMVP) (January 2012). *Concepts and Options for Determining Water and Energy Savings*, Vol. 1.

Eto, J., (1988). "*On Using Degree-days to Account for the Effects of Weather on Annual Energy Use in Office Buildings*," Energy and Buildings, vol. 12, p. 113 – 127

Eto, Joe, Ralph Prahl, and Jeff Schlegal (1996). *A Scoping Study on Energy-efficiency Market Transformation by California Utility DSM Programs*. Lawrence Berkeley National Laboratory. http://eetd.lbl.gov/ea/ems/reports/39058.pdf.

Fagan, Jennifer, Mike Messenger, Mike Rufo, Peter Lai (2009). *A Meta-Analysis of Net to Gross Estimates in California*. Paper presented at the 2009 AESP conference.
GDS Associates, Inc., Nexant, and Mondre Energy (2010). "Net Savings: An Overview.*"* RFP 2009-1 prepared for the Statewide Evaluator.

Feng. W. (2006). *A Method/Macro Based on Propensity Score and Mahalanobis Distance to Reduce Bias in Treatment Comparison in Observational Study*. Public Health Research, paper pr05. http://www.lexjansen.com/pharmasug/2006/publichealthresearch/pr05.pdf

Fowlie, M and C. Wolfram (2009). *Evaluating the Federal Weatherization Assistance Program using a Randomized Encouragement Design (RED).* presented to Environmental Energy Technologies Division (EETD), Lawrence Berkeley National Laboratory, September. http://eetd.lbl.gov/sites/all/files/lbl_09-11-09.pdf

Fowlie, M and C.Wolfram, (undated). *An Experimental Evaluation of the Federal Weatherization Assistance Program.* Presentation prepared for the Michigan Public Service Commission.
http://www.dleg.state.mi.us/mpsc/electric/workgroups/lowincome/fowlie_wolfram.pdf

Fowlie, M and C.Wolfram, (undated). *Randomized Encouragement Design Analysis of the DOE Weatherization Assistance Program*.
http://iber.berkeley.edu/research/researchabstracts/Wolfram_RandomizedEncouragement.pdf

GDS Associates, Inc., Nexant, and Mondre Energy. (2010). *Net Savings: An Overview.* RFP 2009- prepared for the Statewide Evaluator.

Greene, W., (2011). *Econometric Analysis*, 7th Ed., Prentice Hall.

Haeri, Hossein. (2013). *Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures*, Chapter 1: The Introduction.
http://www.nrel.gov/docs/fy13osti/53827.pdf.

Haeri, Hossein, and M. Sami Khawaja (2012). *The Trouble with Freeriders*. Public Utilities Fortnightly. http://www.cadmusgroup.com/wp-content/uploads/2012/11/Haeri-Khawaja-PUF-TroublewithFreeriders.pdf.

Hall, N. et al. (2013). *Setting Net Energy Impact Baselines: Building Reliable Evaluation Approaches*. Paper presented at the 2013 International Energy Program Evaluation Conference, Chicago, Illinois.

Ho, D. et al. (2007). *Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Policy Analysis* 15(3); pp. 199-236.

Imbens, G. and Lemieux, T. (2010). *Regression Discontinuity Designs: A Guide to Practice*. Journal of Economic Literature 48, 281-355.

Itron, Inc. (2010). 2006-2008 Evaluation Report for PG&E Fabrication, Process and Manufacturing Contract Group. Prepared for the California Public Utilities Commission, Energy Division. http://www.calmac.org/publications/PG%26E_Fab_06-08_Eval_Final_Report.pdf http://www.calmac.org/publications/PG%26E_Fab_06-08_Eval_Final_Report_Appendices.pdf

Keating, Ken (2009). *Freeridership Borscht: Don't Salt the Soup*. Paper presented at the 2009 International Energy Program Evaluation Conference.

KEMA (2010). *Final Evaluation Report: Upstream Lighting Program*, Volume 1. CALMAC Study ID: CPU0015.01; Prepared for: California Public Utilities Commission, Energy Division. Prepared by: KEMA, Inc., Prime Contractor: The Cadmus Group, Inc. http://www.calmac.org/publications/FinalUpstreamLightingEvaluationReport_Vol1_CALMAC_3.pdf

Keneipp, M. et al. (2011). *Getting MIF'ed: Accounting for Market Effects in Residential New Construction Programs*. Paper presented at the 2011 International Energy Program Evaluation Conference. Boston, Massachusetts.

Kennedy, P. (2008). *A Guide to Econometrics*, 6th Edition. Wiley-Blackwell, April.

KEMA, NMR Group Inc., Itron, Inc., The Cadmus Group, Inc. (2010). *Residential New Construction (Single Family Home) Market Effects Study*. Phase II Report prepared for California Public Utilities Commission Energy Division. Study ID CPUC0051.01. http://calmac.org/publications/RNC_mkt_effects_Phase_2_report_final_120610-ID.pdf

Khawaja, M. Sami, Josh Rushton, Josh Keeling. (2013). *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measure*s, Chapter 11: "Sample Design Cross-Cutting Protocols. http://www.nrel.gov/docs/fy13osti/53827.pdf.

Mahalanobis, P. (1936). "On the generalised distance in statistics". *Proceedings of the National Institute of Sciences of India 2* (1): 49–55. http://www.new.dli.ernet.in/rawdataupload/upload/insa/INSA_1/20006193_49.pdf

McMenamin, J.S.,(2008). *Defining Normal Weather for Energy and Peak Normalization.* Itron, Inc.

https://www.itron.com/PublishedContent/Defining%20Normal%20Weather%20for%20Energy%20and%20Peak%20Normalization.pdf

Megdal, L. et al. (2009). *Feasting at the Ultimate Enhanced Freeridership Salad Bar.* Paper presented at the International Energy Program Evaluation Conference, Portland, Oregon. http://www.anevaluation.com/pubs/Salad%20Bar%202009%20IEPEC%20paper%205-12-09.pdf.

Messenger, Mike, Ranjet Bharvirkar, Bill Golemboski,Charles Goldman, and Steven Schiller. . (2010). *Review of Evaluation, Measurement and Verification Approaches Used to Estimate the Load Impacts and Effectiveness of Energy Efficiency Programs*. Lawrence Berkeley National Laboratory. http://emp.lbl.gov/sites/all/files/lbnl-3277e.pdf

Mort, Dan. (2013). *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures,* Chapter 9: Metering Cross-Cutting Protocols. http://www.nrel.gov/docs/fy13osti/53827.pdf

Navigant (2012). *A Sampling Methodology for Custom C&I Programs*. Prepared by D. Violette, D., and B. Rogers for the Ontario Natural Gas Technical Evaluation Committee, Ontario Energy Board, November.

Navigant Consulting (2013). *Custom Free Ridership and Participant Spillover Jurisdictional Review*. Prepared for the Sub-Committee of the Ontario Technical Evaluation Committee. http://www.ontarioenergyboard.ca/documents/TEC/Evaluation%20Studies%20and%20Other%20Reports/Ontario%20NTG%20Jurisdictional%20Review%20-%20Final%20Report.pdf.

New York Department of Public Service (July 31, 2013). *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approach*. Appendix H.

New York Department of Public Service (July 31, 2013). *Guidelines for Calculating the Relative Precision of Program Net Savings Estimates*. Appendix I.

New York Department of Public Service (November 2012). *Evaluation Plan Guidance for EEPS Program Administrators,* Update #3, Appendix F. Albany, New York.

Navigant Consulting (May, 2013). *Custom Free Ridership and Participant Spillover Jurisdictional Review*. Prepared for the Sub-Committee of the Ontario Technical Evaluation Committee.

NEEP (2012). *Regional Net Savings Research, Phase 2: Definitions and Treatment of Net and Gross Savings in Energy and Environmental Policy*, submitted to the Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum, by NMR Group and

Research Into Action, December. https://www.neep.org/Assets/uploads/files/emv/NEEP%20-%20Regional%20Net%20Savings%20Report%2012-05-12.pdf

NMR Group, Inc., KEMA, Cadmus Group, Inc., Tetra Tech (2011). *Massachusetts ENERGY STAR Lighting Program: 2010 Annual Report.* Final Report prepared for Energy Efficiency Advisory Council Consultants, Cape Light Compact, NSTAR, National Grid, Unitil, and Western Massachusetts Electric. https://www.efis.psc.mo.gov/mpsc/commoncomponents/viewdocument.asp?DocId=935690223.

NMR Group, Inc. and Research Into Action, Inc. (2010). *Net Savings Scoping Paper.* Revised Draft prepared for Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum. http://www.neep.org/Assets/uploads/files/emv/emv-products/FINAL_Net_Savings_Scoping_Paper_11-13-10.pdf.

NMR Group, Inc. and Tetra Tech (2011). *Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches*. Final report prepared for the Massachusetts Program Administrators. http://www.ma-eeac.org/Docs/8.1_EMV%20Page/2011/2011%20Residential%20Studies/Residential%20MA%20NTG%20Methods%20Final%20072011.pdf

Oak Ridge National Laboratories (1991). *Handbook to DSM Program Evaluation.* Eric Hirst and John Reed, eds., NTIS Pubs., Washington, DC, # ORNL/CON -336, December.

Parfomak, P. and L. Lave (1996). *How Many Kilowatts Are in a Negawatt? Verifying the Ex-Post Estimates of Utility Conservation Impacts at a Regional Level.* Energy Journal 17 (4).

Peters, Jane, Marjorie McRae (2008). *Freeridership Measurement Is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation*. In Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings, Washington, DC. http://www.aceee.org/files/proceedings/2008/data/papers/5_491.pdf.

Prahl, R. et al. (2013). *The Estimation of Spillover: EM&V's Orphan Gets a Home*. Proceedings of the 2013 International Energy Program Evaluation Conference, August.

Provencher, B. and B. Glinsmann (2013). *Evaluation Report: Home Energy Reports – Plan Year 4.* Prepared for Commonwealth Edison Company. February.

Provencher et al. (2013). *Some Insights on Matching Methods in Estimating Energy Savings for an Opt-In, Behavioral-Based Energy Efficiency Program.* 2013 International Energy Program Evaluation Conference, Chicago.

Regional Technical Forum. (2012). *Guidelines for the Development and Maintenance of RTF Savings Estimation Methods.* NWCouncil, Released December 4. Web page:

http://rtf.nwcouncil.org/subcommittees/deemed/ and document link: *http://www.google.co.uk/url?sa=t&rct=j&q=&esrc=s&frm=1&source=web&cd=3&ved=0CDwQFjAC&url=http%3A%2F%2Frtf.nwcouncil.org%2Fsubcommittees%2Fdeemed%2FDraftForReview_Guidelines%2520for%2520RTF%2520Savings%2520Estimation%2520Methods%2520(12-04-2012).docx&ei=epYGUs-SGOXy0gXo6YDAAg&usg=AFQjCNHDq5kHbCPklBvFfoobTohbVCoupA&sig2=Jh63EbYf2eFEmfzQ0nf9fw&bvm=bv.50500085,d.d2k*

Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall (2007). *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches*. Prepared for the California Public Utilities Commission.

Ridge, R. (1997). *Errors in Variables: A Close Encounter if the Third Kind*. Proceedings of the 1997 International Energy Program Evaluation Conference. August, Chicago.

Ridge, R. et al. (2009). *The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating Net-to-Gross Ratio*. Paper presented at the 2009 Energy Program Evaluation Conference, Portland, Oregon.

Ridge, R. et al. (2013). *Gross Is Gross and Net Is Net: Simple, Right*?. Paper presented at the 2013 International Energy Program Evaluation Conference, Chicago, Illinois

Rosenberg, M., and Hoefgen, L. (2009). *Market Effects and Market Transformation: Their Role in Energy Efficiency Program Design and Evaluation.* Prepared for the California Institute for Energy and the Environment and the California Public Utilities Commission Energy Division.

Rufo, Michael (2009). *Evaluation and Performance Incentives: Seeking Paths to (Relatively) Peaceful Coexistence,* Proceedings, the International Energy Program Evaluation Conference, Portland, Oregon, August.

Sacramento Municipal Utilities District (SMUD) (2011). *Evaluation Report: OPOWER SMUD Pilot Year2*. Prepared by Navigant. February. http://www.opower.com/uploads/library/file/6/opower_smud_yr2_eval_report_-_final-1.pdf

Sebold, Frederick D, Alan Fields, Lisa Skumatz, Shel Feldman, Miriam Goldberg, Kenneth Keating and Jane Peters (2001). *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*. http://www.calmac.org/events/20010301PGE0023ME.pdf ,

SEE Action (2012a). *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations.* Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkley National Laboratory. http://behavioranalytics.lbl.gov./reports/behavior-based-emv.pdf. (2012 a).

SEE Action (2012b). *Energy Efficiency Program Impact Evaluation Guide*. Prepared by Steven R. Schiller, Schiller Consulting, Inc. http://www1.eere.energy.gov/seeaction/pdfs/emv_ee_program_impact_guide.pdf. (2012 b).

SMUD (2013). *Load Impact Results from SMUD's Smart Pricing Options Pilot.* Prepared by Freeman Sullivan & Co. for Sacramento Municipal Utility District – SMUD contact Ms. Lupe Jimenez.

Southern California Edison (2012). Edison SmartConnect® Demand Response and Energy Conservation Annual Report, prepared by David Hanna et al., Itron, Inc. for Eric Bell, SCE project manager. https://www.pge.com/regulation/DemandResponseOIR/Pleadings/SCE/2012/DemandResponseOIR_Plea_SCE_20120430_237124.pdf

Stuart, Elizabeth A. (2010). *Matching Methods for Causal Inference: A Review and a Look Forward,* 25(1); pp. 1-21. Statistical Science.

TecMarket Works et al. (2012). *Indiana Evaluation Framework*. Prepared for the Indiana Demand Side Management Coordination Committee.

Tetra Tech, Inc., KEMA, NMR Group, Inc. (2011). *Cross-Cutting (C&I) Free Ridership and Spillover Methodology Study Final Report.* Massachusetts Program Administrators. http://www.ma-eeac.org/Docs/8.1_EMV%20Page/2011/2011%20Commercial%20&%20Industrial%20Studies/MA%20FR_SO%20CI%20%20Study%20w%20Exec%20Summary%205-26-2011%20v11.pdf

The Nonresidential Net-To-Gross Ratio Working Group (October, 2012). *Methodological Framework for Using the Self-Report Approach to Estimating Net-to-Gross Ratios for Nonresidential Customers.* Prepared for the Energy Division, California Public Utilities Commission.

Titus, E. and Michals, J. (2008). *Debating Net Versus Gross Impacts in the Northeast: Policy and Program Perspectives*. ACEEE Summer Study on Energy Efficiency in Buildings, (5);pp. 312-323. https://www.aceee.org/files/proceedings/2008/data/papers/5_429.pdf.

U.S. Department of Energy (2010). *Guidance Document #7: Topic: Design and Implementation of Program Evaluations that utilize Randomized Experimental Approaches".* Smart Grid Investment Grant Technical Advisory Group, November. http://www.smartgrid.gov/sites/default/files/pdfs/cbs_guidance_doc_7_randomized_experimental_approaches.pdf

Violette et al. (1991). *Impact Evaluation of Demand-Side Management Programs — Volume 1: A Guide to Current Practice*. Electric Power Research Institute Pubs., Palo Alto, CA, #EPRI CU-7179, February.

Violette, D. et al. (1993). *Statistically-Adjusted Engineering Estimates: What Can The Evaluation Analyst Do About The Engineering Side Of The Analysis?*. Published in the *Proceedings of the 1993 Energy Program Evaluation Conference*.

Violette, D. et al. (2005). *Commercial/Industrial Performance Program (CIPP) Market Characterization, Market Assessment And Causality Evaluation.* Prepared for NYSERDA, Jennifer Ellefsen, Project Number 7721, March.

Violette, D and B. Provencher (2012). *Review of a Top Down Evaluation Study: Rivers & Jaccard (2011).* Prepared for BC Hydro, Navigant Consulting, Inc., April.

Violette, D. M., Provencher, B., and Sulyma, I. (2012). *Assessing Bottom-Up and Top-Down Approaches for Assessing DSM Programs and Efforts*. International Energy Program Evaluation Conference Proceedings, Rome, June.

Violette, D. (2012). *Uniform Methods Project for Determining Energy Efficiency Program Savings for Specific Measures,* Chapter 13: "Persistence and Other Evaluation Issues Cross-Cutting Protocols", **Subcontract Report,** NREL/SR-7A30-53827, April. See: http://www1.eere.energy.gov/wip/pdfs/53827-13.pdf

West, S, (2008). *Alternatives to the Randomized Controlled Trial.* American Journal of Public Health. Vol. 98, No. 8. http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2007.124446
Wooldridge, J. M. (2002). Econometric Analysis of Cross Section and Panel Data. Cambridge, MA: MIT Press.

Wooldridge, J.M. (2010). *Econometric Analysis of Cross Section and Panel Data.* MIT Press Ltd..