TETRA TECH

complex world | CLEAR SOLUTIONS™

# Massachusetts Program Administrators

Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches (Final)

July 20, 2011

Prepared by NMR Group, Inc. with contributions by Tetra Tech and KEMA

NMR Group, Inc.

KEMA

# Massachusetts Program Administrators

Cross-Cutting Net to Gross Methodology Study for Residential Programs – Suggested Approaches (Final)

July 20, 2011

Prepared for: Massachusetts Program Administrators

**Tetra Tech**
6410 Enterprise Lane, Suite 300 | Madison, WI 53719
Tel 608.316.3700 | Fax 608.661.5181
www.tetratech.com

## TABLE OF CONTENTS

# 1. EXECUTIVE SUMMARY

The primary objective of the current methodology study was to develop suggested approaches for consideration by the PAs for estimating net program impacts for the Massachusetts Program Administrators' (PAs') residential programs. The study team (NMR) started with the revised methodology report for C&I programs (2010)[1] and adapted the decision framework, the literature review, and methodology guidelines to programs targeted to residential customers. Some of that study is repeated here so that those interested only in residential programs will be able to refer to a stand-alone document.

## 1.1 BACKGROUND

Prior to 2003, the Massachusetts Program Administrators (PAs) independently quantified the impacts of free-ridership and spillover (net-to-gross factors) for their energy efficiency programs in Massachusetts. These independent evaluation approaches resulted in reported impacts that were measured using a variety of survey instruments, analysis techniques, and assumptions. In 2003, a consortium of Massachusetts PAs funded a study to develop standardized methods that each of the PAs could use to determine free-ridership and spillover factors for C&I programs.[2] In 2010, a subsequent study revisited the 2003 standardized methodology for C&I programs by reviewing other methodologies now being used across the nation. This review explored the pros and cons of alternative methods for estimating what would have happened absent the program in different contexts. In addition to the review, the report provided a decision framework with guidelines for when the standardized self-report methodology is appropriate and when other methods should be used.

Thus, while methods for estimating net program impacts for the MA C&I programs have been standardized since 2003, the methods for estimating these impacts for the residential programs have never been standardized. Rather, the PAs continue to quantify net effects for residential programs independently; for some residential programs, net effects have never been quantified at all.

## 1.2 METHODOLOGY

In order to develop the decision framework and methodological guidelines for residential programs, we first conducted a review of the PAs' current residential programs, focusing on program elements most relevant to methodological decisions regarding the estimation of net effects. As part of the program review, the study team reviewed the three-year plans and information collected from the PAs by the NMR team and interviewed PA staff about their residential programs. The product of this review is a taxonomy of programs organized by incentive structure, role of trade allies in customer decisions, and other program features. (Appendix A of the report provides this taxonomy).

As the C&I report provided a comprehensive literature review of methods for measuring net-to-gross methods that is largely relevant to residential programs, we did not conduct an extensive additional literature review. Rather, we added to the existing review a recent scoping paper[3] on estimating net savings that we used in part to inform our methodological suggestions for the residential programs.

Based on the program information garnered from the program review, the Net Savings Scoping Paper, and the decision matrix from the C&I report (adapted to the context of the residential programs), we

---

[1] "Cross-Cutting Free-Ridership and Spillover Methodology Study for Commercial and Industrial Programs". Tetra Tech, November 18, 2010.
[2] "Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised)", National Grid, NSTAR Electric, Northeast Utilities, Unitil, Cape Light Compact, June 16, 2003.
[3] "Net Savings Scoping Paper," NMR & Research Into Action, November 13, 2010.

developed suggested approaches for consideration by the PAs for estimating net-to-gross effects for each residential program.

## 1.3 METHODS FOR STUDYING NET-TO-GROSS EFFECTS

Net-to-gross (NTG) effects can be estimated using a variety of different approaches. In deciding which approach or combination of approaches to use in a given situation, several factors need to be considered. Here we provide an overview of the major approaches, including general guidelines on when it is appropriate to use each method.

### 1.3.1 Market sales data analysis

Approaches based on market sales data analysis develop an estimate of the total net effect of the program, including both free-ridership and participant and non-participant "like" spillover. Program effects can be estimated through analysis of market sales data using a few different experimental designs. Typically, a cross-sectional, comparison area approach is used. Post-program data can be compared with data from a non-program comparison area (or multiple comparison areas) at the same point in time, or the change in the program area from the pre-program period to the post-program period can be compared with the change in the non-program area over the same period of time. Factors to consider in deciding whether this approach is appropriate and feasible to use for a particular program include the following:

- *Existence of appropriate comparison area*. There must exist a comparison area (or, ideally, multiple areas) sufficiently comparable to the area of study such that a credible baseline for the area of interest can be constructed from it, possibly with a set of systematic adjustments to control for differences in total size of the areas, or demographic differences between the areas. However, if there are unique characteristics in the program area that could affect the market, such as special climates that are not found in other areas, a comparison area approach cannot be used.

- *Available market data.* Market data analysis requires obtaining sufficiently comprehensive market data for both the area of interest and an appropriate comparison area(s). Comprehensive sales/shipment tracking systems have never been available for most markets. Absent such comprehensive sales data, a general picture of market coverage can be obtained by interviewing 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. Manufacturers and regional buyers/distributors can sometimes provide market sales and/or shipment data or self-reported sales/shipments. In some cases, participating end-users' self-reported purchases can provide market data if the sample is sufficiently large and representative of the market. If there are gaps in the data that are substantial enough to make the results misleading, this approach cannot be used.

- *Features of program.* In general, market data analysis is appropriate for programs that promote large numbers of homogenous measures and that have substantial influence "upstream" from the end-user.

### 1.3.2 Self-reported counterfactual surveys

Self-Reported counterfactual surveys attempt to determine what would have happened absent the program by asking people what actions they would have taken if the program were not available (for free ridership) and what subsequent actions the program influenced them to take (for spillover). Self-reports of the counterfactual by participating customers (or key decision-makers) is the approach historically used to estimate free-ridership and spillover in Massachusetts. This information can be combined with self-reported counterfactual surveys of retail store managers, contractors,

manufacturers, and distributors. For high-volume measures, the vendor, contractor, or other market actor is asked about general influences of the program on the vendor's sales, stocking patterns or installations. The end-user survey indicates the influence of the incentive or other assistance the customer received, as well as the influence of the contractor or other supply-side agent in the customer's decision to adopt a measure, and the supplier survey indicates the role of the program in changing the supplier's promotional efforts. Self-reported counterfactual surveys can be used with almost any type of downstream program, but should not be used for upstream programs.

The end-user counterfactual approach has often been criticized. Concerns about the method include the following:

- potential biases related to respondents' giving "socially desirable" answers and the tendency to rationalize past decisions

- people's inability to know what they would have done in a hypothetical alternative situation, especially in current program designs that use multiple methods to influence behavior

- potential arbitrariness of scoring methods that translate responses into free-rider estimates

While these concerns are legitimate, good survey techniques can mitigate many of these potential biases.

### 1.3.3   Pricing and elasticity analysis

Several methods can be used to estimate price elasticity—that is, the effect on purchases of lowering the price through upstream or downstream incentives. These methods are used for programs whose influence on the market is affected chiefly through reducing prices for efficient equipment.

*Stated Preferences.* Stated preference experiments systematically ask potential customers what they would choose from a set of options with different features and prices, or how a change in price affects purchases. The "choice sets" offered each customer are designed so that the effect of price and features can be estimated from the data set of all the customers' responses. The key challenge for stated preference methods is the potential difference between what customers say they will buy in a hypothetical situation, and what they do buy when they actually have to spend money.

*Revealed Preferences.* Revealed preference studies observe the actual choices customers make from true choices available to them when making purchases. The key challenges for revealed preference studies are as follows:

- It is necessary to observe choices in contexts that are similar to those of the study area, except for the presence of the program.

- It is usually necessary to observe the items as they are being purchased, as customers usually cannot reliably report the efficiency levels of recently purchased equipment. Direct observation can be accomplished via store intercepts or via onsite visits. Another challenge for this method is the potential non-response bias—that is, potential differences between customers willing to have their purchases observed and those who decline.

*Shelf and stocking observations.* To obtain a NTG estimate from the survey-based elasticity estimates, it is necessary to estimate the effect of the program on prices in the region. The price effects can be estimated using in-store shelf and stocking surveys. Alternatively, price effects can be asked about in supplier surveys at various levels.

## 1.3.4   Billing analysis

Billing analysis develops an estimate of program savings by analyzing consumption data. The most common form of such analysis compares usage after program participation with usage prior to program participation, with some form of weather normalization. Since billing analysis typically requires up to 12 months of post-implementation consumption data, this approach may not be feasible where more timely feedback on net savings is required. In general, billing analysis is appropriate to use only when participant whole-house savings are substantial relative to total and when there are large numbers of fairly homogenous participants.

When a comparison group is used in the analysis, the resulting estimate of savings is considered to be net of free-ridership; the analysis does not isolate NTG effects from adjustments to gross savings. The comparison group change in usage represents how the participants would have changed absent the program. However, if there is non-participant spillover, this approach will understate the savings attributable to the program. Essentially, the non-participant spillover will be incorrectly counted as part of naturally occurring savings, and subtracted from participant change, instead of adding to it. Thus, not only is the program not credited with the non-participant spillover, but also the true participant savings attributable to the program are underestimated.

This approach depends heavily on the "comparability" of the comparison group. In most programs, customers who choose to participate are different from those who do not participate, in ways that can affect their year-to-year consumption changes. This self-selection effect can be controlled for, to some extent, by including in the analysis customer characteristics that might be associated with both the propensity to participate and the consumption changes absent the program. A second situation where a valid comparison is available occurs when customers are randomly assigned to receive the participant "treatment" or not. With random assignment, there is no systematic difference between the untreated or control group customers and the participating customers, other than the treatment itself. Therefore, the control group provides an unbiased estimate of what participants would have done absent the program treatment.

## 1.3.5   Structured expert judging

Structured expert judgment studies assemble panels of individuals with close working knowledge of the technology, infrastructure systems, markets, and political environments addressed by a given energy efficiency measure, to estimate baseline market share and, in some cases, forecast market share with and without the program in place. The *Delphi process* is the most widely known method of this family of methods. Properly executed, a Delphi process includes at least one iterative step in which each expert is provided with the judgments and rationales of all the other experts, and offered an opportunity to re-assess or defend the original position.

In the context of energy efficiency program evaluation, structured expert judging has as its foundation the experts' experience with other programs and the NTG findings for other programs, typically based on other methods. Structured expert judging allows experience from other contexts to be applied to situations in which all feasible methods may have substantial threats to validity. Expert judging also allows adjustments to be made, albeit subjectively, for some of these threats.

A particularly useful role for structured expert judging is to develop a "consensus" estimate to consolidate results from multiple estimation methods. For example, recognizing weaknesses in every feasible method, the evaluators of the Massachusetts Residential Lighting Program are using a Delphi approach to assess net-to-gross estimates derived from five different methods and to develop a recommended consensus estimate. There are some markets and programs that are not amenable to

other methods for estimating net effects, in which case expert judgment, aided by a summary of the history and current state of the market and the program, may be the best available approach.[4]

### 1.3.6 Historical tracing: case study method

This method involves the careful reconstruction of events leading to the outcome of interest—for example, the launch of a product or the passage of legislation, to develop 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 argument. These include:

- Compiling, comparing, and weighing the merits of narratives of the same set of events provided by individuals with different points of view and interests in the outcome.

- Positing a number of alternative causal hypotheses and examining their consistency with the narrative fact pattern.

- Assessing the consistency of the observed fact pattern with linkages predicted by the program logic model.

Researchers use information from a wide range of sources to inform historical tracing analyses. These include public and private documents, personal interviews, and surveys.

This method is best suited to attribution analysis of major events such as adoption of new building codes or policies. It is not typically applicable to energy efficiency programs, and is not suggested as a primary method for the Massachusetts programs.

### 1.3.7 Program delivery staff surveys

Some practitioners have used reports from program staff on their influence on customers as input to NTG estimates, particularly for custom projects. An argument for this approach is that customers may be inclined to want to take credit for a good idea, and would not necessarily remember how the program staff helped provide information and develop the project. A critical counter-argument is that program staff have an explicit vested interest in obtaining high attribution credit, and may not realize when the suggestion they made was already something the customer was considering.

Thus, we do not suggest using delivery staff reports of their influence on customers as a valid basis for estimating NTG. We do however suggest obtaining information from delivery staff as background for understanding and clarifying projects and decision points.

### 1.3.8 Deemed or stipulated NTG estimates

In deemed or stipulated approaches to estimating the net effects of programs, a specific NTG ratio is assumed and applied to the program. Generally, program administrators and regulators draw on available evidence of program impacts, including previous NTG studies, to help decide on the percentage of gross savings that can be claimed by a program. These approaches are particularly useful for forward-looking NTG estimates, possibly building on current and previous estimates, and also relying on program theory and examples from the histories of other programs.

---

[4] See, for example, the 2010 CPUC Residential New Construction Market effects Study by KEMA, NMR, Itron, and Cadmus.

The credibility of the negotiated net savings figure would depend on the type, amount, and quality of the information informing it; such information can include program tracking data, adjusted gross savings estimates, net savings estimates derived from periodic research, sales/shipment data, etc. Although there is always the potential for a deemed estimate to be incorrect, as it is not based on information that is specific to the program and program period, the approach is simple and inexpensive, and avoids depending entirely on controversial measurements of the counterfactual of net savings.

## 1.4     SUGGESTED APPROACHES FOR MASSACHUSETTS RESIDENTIAL PROGRAMS

Below is a summary of the suggested approaches for each of the PAs residential programs based on the key characteristics of programs, including the types and number of measures promoted, the likelihood of substantial market effects, and the availability of sales data and an appropriate comparison area.

The *Energy Star Lighting Program* promotes large numbers of similar prescriptive measures. As an upstream program, there is a high likelihood of important influences unknown to customers and a high likelihood of substantial market effects. While these program features preclude the end-user self-report counterfactual approach and point to a market-level approach based on sales data, at this time comprehensive sales data are not available for lighting products. Further, it is difficult to find a valid comparison area to use as a baseline because so many states now have similar lighting programs. Our suggestion is to use multiple methods to derive several NTG estimates. These estimates would be presented to a Delphi panel of experts, who would come to a consensus on the overall NTG estimate for the program. Estimation methods could include: 1) market data analysis on purchase/sales data collected from customer self-reports of purchases and interviews with vendors and suppliers, 2) in-store revealed preferences observations, and 3) shelf and stocking surveys of retail stores, paired with price elasticity analysis.

*The Energy Star Appliances Program* also promotes large numbers of similar prescriptive measures, with a moderate degree of influences unknown to the customer and high potential for market effects. Again, in principle a comprehensive sales-based, cross-sectional approach is appropriate, as was used in NMR's 2004 evaluation of the ES Appliances Program.[5] However, comprehensive sales data, which had been tracked by the Department of Energy since 1998, are no longer available. In their absence, we suggest an approach similar to that suggested for Energy Star Lighting (above), with a Delphi panel of experts estimating the NTG for the program based on a number of estimates derived from multiple methods. However, the market data would come from interviews with supply-side market actors instead of from customers' self-reported purchases. In addition, customer self-reported counterfactual surveys would be included in the estimation methods for the Appliance program, as participants in the program know they are participants and are more likely to be able to report the effect of aspects of the program (e.g., incentives) on their purchases.  However, the costs of using multiple methods may be prohibitive considering the relatively modest size of this program. Therefore, it may be more appropriate to use a modified approach involving interviews with supply-side market actors, or customer self-reported counterfactual surveys, or a combination of the two. For the Appliance Retirement Program, estimates could be based on participant self-reported counterfactual surveys, with research on the appliance secondary market helping to inform the counterfactual—what would have happened to the appliances if they had not been picked up by the program.

*MassSave* is primarily an "umbrella program," through which customers receive whole-house efficiency audits and recommendations for measures to improve the efficiency of their home. While some measures (e.g., light bulbs and low-flow showerheads) are installed directly through the program, many of the recommended measures (e.g., weatherization measures such as insulation, air sealing, and heating equipment) are promoted by other residential programs within the portfolio. Therefore, although

---

[5] NMR Group. 2007. Massachusetts Energy Star Appliance Program: Market Share Tracking and Analysis.

many of the program's measures are themselves prescriptive, a sales-based approach would not work, as it would not be able to distinguish the effects of the program from those of other programs that primarily promote the same measures. We suggest using customer self-reported counterfactual surveys, supplemented by input by the auditors and contractors, to gauge the effects of the audit and the incentives on customers' purchase decisions.

The *Weatherization Program* is largely integrated with MassSave. In order to receive incentives for weatherization measures, customers must have a MassSave auditor visit their homes and recommend particular measures, such as duct sealing and air sealing. Free ridership and spillover could be estimated through self-reported counterfactual surveys of MassSave participants who installed weatherization measures, supplemented by input by the auditors and contractors.

*Energy Star Homes,* or *Residential New Construction* promotes efficiency on a whole-building level, not at the level of individual measures. As explained previously, this house-as-a-system approach makes sales-based approaches non-viable. At the same time, a cross-sectional approach will not work because of unique conditions in the local building context, such as building codes and their level of enforcement. Self-report counterfactual surveys with participant builders can be used to estimate free ridership, but this approach will not capture the potentially substantial market effects of the program. To estimate market effects, expert judging can be used. For example, a recent evaluation of the California investor owned utilities' (IOUs') residential new construction (RNC) programs, covering the 2006-2008 program years, examined how the RNC programs could affect the efficiency of California homes built outside those programs. The market effects study examined net impacts achieved in two ways: 1) net impacts achieved through the IOU programs' influence on above-code practices in homes built outside the IOU programs; and 2) net impacts achieved through the IOU programs' influence on increased code compliance in homes built outside the IOU programs. The evaluation team provided two Delphi panels—a panel of Title 24 Consultants[6] and one of industry experts—with gross savings calculated by comparing the efficiency between above-code and just-code homes, and between just-code and below-code homes. Using these gross savings estimates, the experts assigned attribution scores to the RNC programs and other factors to derive net savings estimates.[7]

The *Residential Heating and Cooling* and the *Heating and Hot Water Equipment* programs have similar features—large numbers of prescriptive measures, likelihood of substantial market effects, etc.—and can be treated similarly. Again, a sales-based approach would be ideal for these programs, but such comprehensive data is not currently available. Therefore, self-report surveys of both customers and contractors can be used to estimate free ridership, while market effects can be estimated through interviews with contractors and suppliers in comparison areas. The interviews would gather information on sales levels and market share of efficient and standard equipment.

Market effects of the *natural gas training programs* could be estimated through the self-report surveys of participants in the training programs. Surveys could collect data on installation practices and other behavioral changes caused by the training programs with energy savings estimated from the behavioral changes.[8] In addition, should budgets permit, it would also be important to employ a simple

---

[6] Title 24 Consultants advise builders and provide certificates of compliance with the energy efficiency portion of the building code for newly constructed homes, as required by California state law.

[7] KEMA, NMR Group, Itron, Cadmus Group. 2010. Phase II Report Residential New Construction (Single-Family Home) Market Effects Study.

[8] A recent evaluation of education and training programs in California used this method to estimate net and gross energy savings of the programs. Opinion Dynamics, Wirtshafter Associates, Jai J Mitchell Analytics, Summit Blue Consulting. 2010. *Indirect Impact Evaluation of the Statewide Energy Efficiency Education and Training Program.*
http://www.calmac.org/publications/06-08_Statewide_Education_and_Training_Impact_Eval_Vol_I_FINAL.pdf

test/comparison approach with field studies to determine whether training is in fact associated with better practices.[9]

The *Multi-family Retrofit* program has similar features to the C&I programs: a custom, whole-building approach to energy efficiency, with relatively few participants. Therefore, the methods suggested for the C&I programs are appropriate for this program as well. Specifically, surveys with decision-makers and market actors (e.g., retail store managers, contractors, etc.) can be used to gauge the influence of the program on the customer's purchase decisions.

The *O Power* behavioral pilot is unique among the residential programs in at least two ways that affect how program influence can be estimated. First, it neither promotes particular measures nor provides incentives. Rather, the program aims to change participating customers' energy usage by promoting energy-saving behaviors. Second, the pilot program employs a true experimental design, with a randomly selected control group of non-participants in the same geographical area. This design makes the program ideal for a billing analysis approach comparing the overall energy usage of participants (i.e., those who were randomly selected to receive the Home Energy Reports) with that of the non-participants (a randomly selected group of customers who did not receive the reports) over the same time period. This approach would involve statistical analyses of differences in pre- and post-treatment electric and natural gas consumption by treatment group members (compared to control group households). These analyses yield an estimate of program influence on energy savings net of free ridership. However, non-participant spillover would be counted negatively, as it would decrease the difference in energy use between the treatment group and the control group. Adjustments to non-participant spillover might be possible through surveys with non-participants, gauging their awareness of the program as well as any changes in energy-related behavior as a result of this awareness, might be used to assess the degree to which any non-participant spillover is occurring.

---

[9] Northeast Energy Efficiency Partnerships, Inc., May 2006, Strategies to Increase Residential HVAC Efficiency in the Northeast.

## 2. INTRODUCTION

The focus of this report is on the general methods for estimating net program effects for residential programs in Massachusetts.

To create this report, the study team started with a recent report providing recommendations for estimating net program effects for non-residential programs in Massachusetts[10] and adapted it to the context of MA's residential programs. The purpose of 2010 Methodology Study for Commercial and Industrial (C&I) programs was to revise an earlier methodology report for C&I programs.[11] The 2010 effort involved a number of steps. In addition to the methodological recommendations for the MA C&I programs, the C&I report included a literature review, an overview of different methods for estimating net impacts, an analysis examining appropriate methods for different types of programs, and a discussion of best practices for self-report methods. In contrast, the scope of the current effort was more circumscribed. Our primary purpose was to provide methodological suggestions for measuring net program impacts for the Massachusetts Program Administrators' (PAs') residential programs. The current report retains those portions of the C&I report that are relevant to residential programs, with minor changes, deletions and additions to reflect the residential context.

As defined in the C&I report, the net program effect is the observed effect, less the estimate of what would have happened absent the program. Several methods to estimate what would have happened absent a program are discussed in Chapter 2. The most common method is to rely on participant self-reports to estimate free-ridership and participant spillover, and non-participant self-reports to estimate non-participant spillover.

At this point it is useful to define free-ridership and spillover. Program attribution refers to energy impacts that can be attributed with some level of confidence to program efforts. A program's *free-ridership rate* is the percentage of program savings attributed to free-riders. A *free-rider* refers to a program participant who received an incentive or other assistance through an energy efficiency program who would have adopted the same high-efficiency measure[12] on their own at that same time if the program had not been offered. For free-riders, the program is assumed to have had no influence or only a slight influence on their decision to install or implement the energy efficiency measure. Consequently, none or only some of the energy (and demand) savings from the energy efficiency measures taken by this group of customers should be credited to the energy efficiency program.

In addition to simply identifying free riders, it is important to measure the *extent* of free-ridership for each customer. Pure free-riders (100%) would have adopted exactly the same energy efficiency measures at that time absent the program. Partial free-riders (1–99%) are those customers who would have adopted some measures at that same time on their own, but of a lesser efficiency or a lesser quantity, or they would not have adopted the efficient measures until a later time. Thus, the program had some impact on their decision. Non-free-riders (0%) are those who would not have installed or implemented any program energy efficiency measure (within a specified period of time) absent the program services.

---

[10] "Cross-Cutting Free-Ridership and Spillover Methodology Study for Commercial and Industrial Programs". Tetra Tech, November 18, 2010.

[11] "Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised)", National Grid, NSTAR Electric, Northeast Utilities, Unitil, Cape Light Compact, June 16, 2003; Rathbun, Pam, Carol Sabo, and Bryan Zent. PA Consulting Group. Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised). Prepared for the Massachusetts Utilities, June 13, 2003.

[12] For purposes of this discussion, an "energy efficiency measure" includes high efficiency equipment or appliances, an efficiency measure such as weatherization, or an energy efficient practice such as turning off a computer when not in use.

*Spillover* refers to additional energy efficiency measures adopted by a customer due to program influences, but without any financial or technical assistance from the program. *Participant "like" spillover* refers to the situation in which a customer installed energy efficiency measures through the program, and then installed additional measures of the same type due to program influences. *Participant "unlike" spillover* is in which the customer installs other types of energy efficient measures than those offered through the program, but is influenced by the program to do so.

*Free-drivers,* or non-participant spillover, refers to any energy efficient measures adopted by program non-participants due to the program's influence. The program can have an influence on contractors and vendors as well as an influence on product availability or practices, product or practice acceptance, customer expectations, and other market effects. All of these may induce non-participants to install energy efficient measures. *Non-participant "like" spillover* refers to additional measures of the same type as the programs that are adopted due to the program's influence.

Some approaches to measuring net effects, based on market-level data instead of participant self-reports, do not isolate free ridership and spillover; rather, both are embedded in an estimate of the net-to-gross ratio for the program. This ratio, instead of separate free ridership and spillover rates, is applied to adjusted gross savings. These approaches are most appropriate for programs with the potential for significant non-participant spillover or market effects.

Over the past decade, there has been extensive debate across the country regarding the need to estimate what would have happened absent a program and how to do so. Isolating the effects of these program factors from other influences in the decision to adopt energy efficiency measures is often referred to as "attribution." The increasing importance of program attribution given aggressive savings targets and DSM incentive mechanisms has intensified this debate. Some argue that free-ridership and spillover cancel each other out and should not be measured, that they are too difficult to estimate reliably, or that funds are better spent on program implementation. However, other evaluators and regulators note the advantages of consistent measurement of free-ridership and spillover. Fagan, Messenger, Rufo and Lai (2009)[13] list the following reasons why understanding a program's net savings is important:

- Understand program and portfolio cost-effectiveness

- Improve portfolio design and resource allocation

- Refine program design

- Understand market transformation

- Align program administrators' financial interests with societal interests

- Understand how energy efficiency programs affect baseline load forecasts and short-term power procurement decisions

## 2.1   BACKGROUND

Prior to 2003, the Massachusetts Program Administrators (PAs) independently quantified the impacts of free-ridership and spillover (net-to-gross factors) for their energy efficiency programs in Massachusetts. These independent evaluation approaches resulted in reported impacts that were measured using a variety of survey instruments, analysis techniques, and assumptions. In 2003, a consortium of Massachusetts PAs funded a study to develop standardized methods that each of the

---

[13]   "A Meta-Analysis of Net to Gross Estimates in California". Jennifer Fagan, Mike Messenger, Mike Rufo, Peter Lai, paper presented at the 2009 AESP conference.

PAs could use to determine free-ridership and spillover factors for C&I programs.[14] In 2010, a subsequent study revisited the 2003 standardized methodology for C&I programs by reviewing other methodologies now being used across the nation. This review explored the pros and cons of alternative methods for estimating what would have happened absent the program in different contexts. In addition to the review, the report provided a standardized methodology for situations in which end-users are able to report on program impacts via self-report methods and a decision framework with guidelines for when the standardized self-report methodology is appropriate and when other methods should be used.

Thus, while methods for estimating net program impacts for the MA C&I programs have been standardized since 2003, the methods for estimating these impacts for the residential programs have never been standardized. Rather, the PAs continue to quantify net effects independently for individual residential programs; for some residential programs, net effects have never been quantified at all.

## 2.2    STUDY OBJECTIVES

The primary objective of the current methodology study was to develop suggested approaches for consideration by the PAs for estimating net program impacts for the Massachusetts PAs' residential programs by reviewing the revised methodology report for C&I programs (2010) and adapting the decision framework and methodology guidelines to programs targeted to residential customers. The study team particularly sought to identify residential programs for which market-level approaches to measuring net-to-gross effects, rather than standard self-report methods, might be appropriate and feasible.

## 2.3    SCOPE OF THE RESEARCH

In order to develop the decision framework and methodological guidelines for residential programs, the study team first conducted a review of the PAs' current residential programs, focusing on program elements most relevant to methodological decisions regarding the estimation of net effects. As part of the program review, the study team reviewed the three-year plans and information collected from the PAs by the NMR team. Where additional information was required to develop a complete understanding of the programs (or specific aspects of programs), and to adapt the C&I decision framework and methodology to residential programs, the cross-cutting study team interviewed PA staff about their residential programs. The product of this review is a taxonomy of programs organized by incentive structure, role of trade allies in customer decisions, and other program features. (Appendix A provides this taxonomy).

As the C&I report provided a comprehensive literature review of methods for measuring net-to-gross methods that is largely relevant to residential programs, we did not conduct an extensive additional literature review. Rather, we added to the existing review a recent scoping paper[15] on estimating net savings that we used in part to inform our methodological suggestions for the residential programs.

Based on the program information garnered from the program review, the Net Savings Scoping Paper, and the decision matrix from the C&I report (adapted to the context of the residential programs), we developed suggested approaches for consideration by the PAs for estimating net-to-gross effects for each residential program. The study team also provided suggestions for acquiring the data required for some of the suggested methods.

---

[14]  "Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised)", National Grid, NSTAR Electric, Northeast Utilities, Unitil, Cape Light Compact, June 16, 2003.

[15]  "Net Savings Scoping Paper," NMR & Research Into Action, November 13, 2010.

## 2.4　OVERVIEW OF REPORT

Chapter 2 discusses the methodologies available for studying what would have happened absent the program. It includes a decision framework for selecting appropriate methods as well as suggestions for measuring net program impacts for residential programs based on applying the decision framework to the Massachusetts residential programs. Chapter 3 presents the findings from the literature review on self-report methodologies from the 2010 C&I report, which we modified to reflect the context of residential programs. The taxonomy of PA residential programs is contained in Appendix A. Appendix B summarizes the sources reviewed as part of the best practice review.

# 3. METHODOLOGIES FOR STUDYING FREE-RIDERSHIP AND SPILLOVER

## 3.1 TYPES OF METHODOLOGIES AND THEIR ADVANTAGES/DISADVANTAGES

### 3.1.1 Overview

The best methods for estimating what would have happened absent a program can vary by program type. In some cases, method selection depends on the details of a program and may be a matter of judgment or degree. Some methods are fundamentally not applicable for certain program categories. The most common methodology is self-reporting of the counterfactual by participating customers. Other methods, discussed further below in Section 2.1.2, include market sales data analysis, self-reported counterfactual surveys of non-participating customers and market actors, pricing and elasticity analysis, billing analysis, structured expert judging, historical tracing (case study method), program delivery staff surveys, and deemed savings estimates.

Section 2 provides a typology of methods that identifies the sources of data, the types of data collected, and how the data are analyzed. Then, for each method described in these terms, key factors affecting the method's validity are identified, as well as key data collection issues.

Methodological characterizations, in terms of risks to validity, reliability, applicability, and costs, are never black and white. Accordingly, the choice of the "best" method for a particular situation is often not clear-cut. Section 2.2 includes a decision framework showing how well the different methods apply to programs with particular features. Finally, we apply this decision framework to the residential programs, providing suggestions for methods to estimate net effects for each program.

### 3.1.2 Discussion of general approaches

#### a. Market sales data analysis

Approaches based on market sales data analysis in many ways approach the ideal when the necessary data can be obtained. These methods can largely capture the total net effect of the program, including both free-ridership and participant and non-participant "like" spillover. However, this ideal is often not possible to implement because of the difficulty in obtaining sales data, the lack of an appropriate comparison area (i.e., baseline) with which to compare sales from the program area, and other factors.

*Experimental design.* Program effects can be estimated through analysis of market sales data using a few different types of experimental designs. First, pre-program data can be compared to with-program data from the program area. The major problem with this method is that markets change over time naturally, due to many different factors that are unrelated to the program; therefore, market changes due to the program cannot be isolated from changes due to other factors. Second, post-program data can be compared with data from a non-program comparison area at the same point in time. This design can better account for market changes in the program area due to non-program factors. However, the comparison area approach introduces problems with potential non-comparability of the two areas, as well as difficulty in attributing differences between the two areas to the program rather than other factors in a particular time period (i.e., the period of program activity). Finally, the change in the program area from the pre-program period to the post-program period can be compared with the change in the non-program area (or multiple comparison areas) over the same period of time. Even with a pre/post-test-comparison design, with before and after measurements in both the program area and the comparison area, long-standing programs tend to render the program and comparison areas systematically non-comparable in ways that are aggravated by this particular experimental design.

It is not necessary for the comparison area to be exactly comparable to the area of study interest. Rather, it is necessary for a credible baseline for the area of interest to be constructed based on the comparison-area data, possibly with a set of systematic adjustments. For example, sales data may be expressed in terms of sales shares to control for differences in total size of the two areas; if available, shares may be calculated separately by segment to allow adjustment for a different segment mix between the areas. The comparison between the program and non-program areas can also be made more reliable by controlling for certain variables, such as demographic differences in the two areas, in a regression model. This is the approach currently being used to estimate the Net-to-gross ratio for lighting programs in Massachusetts and other states.[16] An evaluation of the cumulative market effects of the MA Energy Star Appliance Program in 2003 and 2004 used a similar modeling approach.[17,18]

*Sources of data.* The main challenge of this method is in obtaining comprehensive market data for both the area of interest and an appropriate comparison area (or ideally, multiple comparison areas). Comprehensive sales/shipment tracking systems have never been available for most markets. Although this problem can be ameliorated to some extent by requiring participating market actors, such as suppliers and retailers, to provide sales data for research purposes, at this time they are rarely required to do so. Absent such comprehensive sales data, a general picture of market coverage (i.e., sales/purchases of efficient and standard equipment) can be obtained in other ways.

Vendors and contractors can be asked about sales volumes and efficient equipment sales shares for conditions with and without the program, or for in-territory and comparison area sales. This approach can be analyzed similarly to market-level sales data. The difference is that the market sales data approach usually refers to comprehensive or nearly comprehensive reporting of sales. By contrast, vendor surveys may collect "best guess" estimates of sales volumes and shares from a sample, then use sampling weights and other measures of size (such as employment) to expand the survey responses to the full market. In addition, manufacturers and regional buyers and distributors can sometimes provide either market sales and/or shipment data or self-reported sales/shipments. In some cases, participating end-users' self-reported purchases can provide market data if the participant sample is sufficiently large and representative of the market. Self-reported purchase estimates can be obtained through telephone surveys or through on-site data collection. For example, studies evaluating lighting programs have used bulb purchase and socket saturation data from on-site visits to customers' homes because they have been found to be more reliable than self-reported estimates from telephone surveys.

## b.  *Self-reported counterfactual surveys*

Self-reported counterfactual surveys attempt to determine what would have happened absent the program by asking people to report on a hypothetical situation—a situation in which a particular existing program did not exist (or the participant had not heard of it, was not eligible, etc.).

---

[16] "Results of the Multistate CFL Modeling Effort," NMR Group, February 4, 2010.
[17] "Massachusetts Appliances Market Progress and Evaluation Report: Regression Analysis," Nexus Market Research, April 12, 2004.
[18] "Massachusetts Appliances Market Progress and Evaluation Report: Statistical Analyses of Market Penetration of ENERGY STAR-Compliant Appliances," Nexus Market Research, February 25, 2005..

**Cross-Cutting Net to Gross Methodology Study for Residential Programs**

*Participating and non-participating end-users/decision makers.* Self-reports of the counterfactual by participating customers (or key decision-makers) is the approach historically used to estimate free-ridership and spillover in Massachusetts. It can be used with almost any type of downstream program. The self-report approach to the counterfactual, which involves asking participants about the effect of the program on their decision to adopt specific measures, has often been criticized. For example, Peters and MacRae[19] identify the following primary concerns:

- potential biases related to respondents' giving "socially desirable" answers

- people's inability to know what they would have done in a hypothetical alternative situation, especially in current program designs and a larger milieu that use multiple methods to influence behavior

- the tendency of respondents to rationalize past decisions

- potential arbitrariness of scoring methods that translate responses into free-rider estimates

- lack of customer recognition of the influence the program may have on other parties influencing their decisions (e.g., program may have influenced contractor practices, which in turn may indirectly impact the participant's decision).

Another concern about customer self-reports of the counterfactual is that they can underestimate spillover. Ideally, data on free ridership and spillover are collected at two different times. Free ridership is ideally assessed as close as possible to the time of the decision to install the program-supported measure, while spillover is best assessed months after program participation. Assessing spillover at the same time as free-ridership may be too soon to learn about spillover effects on later purchases.

While these concerns are legitimate, they have been recognized for as long as free-ridership has been estimated. Good survey techniques can mitigate many of these potential biases. (See, for example, Ridge et al.[20]) Moreover, the majority of methods available for estimating program effects on customer measure adoption are based on some type of surveys. Even market sales data often come from survey responses which may be incomplete or subject to various biases. *The key difference is that self-reported free ridership relies on the respondent's judgment about the counterfactual, or what would have happened absent the program—something that in fact did not happen—whereas self-reported sales data, however imperfect, relate to something that actually did happen.*

In addition to certain biases working toward over-reporting of free-ridership, there are other biases working toward under-reporting. Some of the recent literature paints a bleak picture of our ability to learn anything by talking to people. However, there are well established methods that can mitigate many of these problems. In particular, biases in both directions can be minimized by using well designed surveys with good set-up questions. Furthermore, scoring systems can be validated and calibrated by methods like what was done for the residential/small customer net-to-gross protocol for

---

[19] Free-ridership Measurement Is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation?, Jane S. Peters and Marjorie McRae, Research Into Action, Inc. In Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings. Washington, DC., 2008.

[20] The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio, Ridge, Richard, Ridge & Associates, Philippus Willems, PWP Inc, and Jennifer Fagan, Itron, Inc., and Randazzo, Katherine, KVD Research Consulting. Presented at Proceedings of the 2009 International Energy Program Evaluation Conference, Portland, Oregon.

the California Public Utilities Commission 2006-08 evaluations and the Massachusetts C&I Net to Gross methods study in 2010.[21]

*Market actors.* Self-reported counterfactual surveys of retail store managers, contractors, manufacturers, and distributors can also be used, usually to supplement end-user/decision-maker self-reported counterfactual surveys. For high-volume measures, the vendor, contractor, or other market actor is asked about general influences of the program on the vendor's sales, stocking patterns or installations. For custom or relatively rare measures, the vendor, energy-efficiency auditor, or contractor can be asked about the decision process and influence for individual customers. This information can be combined with the customer's report of influences. If either the customer or the vendor reports that the program influenced the purchase or sale, some amount of credit is given to the program, depending upon the degree of influence.

*Combining end-user and market actor counterfactual surveys.* Self-reported counterfactual surveys of customers can be combined with those of upstream actors to capture the effect of the program on the upstream actors, who in turn influence the end-user. Studies combining participating customer and participating vendor self-reports in this way have found the following:

- The end-user survey indicates the influence of the incentive or other assistance the customer received, as well as the influence of the vendor, contractor, or other supply-side agent, in the customer's decision to adopt a measure.

- The supplier survey indicates the role of the program in changing the supplier's promotional efforts.

- Together these two sources indicate the likely role of the program in the customer's decision to adopt the measure.

### c. Pricing and elasticity analysis

Several methods can be used to estimate price elasticity—that is, the effect on purchases of lowering the price through upstream or downstream incentives.

*Stated preferences.* Stated preference experiments systematically ask potential customers what they would choose from a set of options with different features and prices. The "choice sets" offered each customer are designed so that the effect of price and features can be estimated from the data set of all the customers' responses. Conjoint and double-bounded techniques are two approaches structured for this type of estimation.[22] The key challenge for stated preference methods is the potential difference between (1) what customers say they will buy in a hypothetical situation, and (2) what they do buy when they actually have to spend money.

Simpler stated preference surveys can be used when features are not being investigated and the only question is how a change in price affects purchases.

---

[21] Response to Overarching Comments Regarding the Use of Self-Reported Net-to-Gross (NTG) and the Residential and Small Commercial Self-Report Approach NTG Method, January 28, 2010, Residential/Small Commercial Joint Simple Net-to-Gross (Self-Report) Committee.

[22] Conjoint experiments provide customers with a series of hypothetical choices among a set of products with varying combinations of attributes and prices. Double-bounded questions ask a series of questions to determine upper and lower bounds on the price at which a customer would buy a particular product.

*Revealed preferences.* Revealed preference studies observe the actual choices customers make from true choices available to them when making purchases. The key challenges for revealed preference studies are as follows:

- It is necessary to observe choices in contexts that are similar to those of the study area, except for the presence of the program. This is a similar challenge to that of obtaining a valid comparison area for market sales data. The difference is that if the valid comparison area can be defined, it may be more practical to collect revealed preference data there than to obtain comprehensive sales data.

- To obtain accurate revealed preference information, it is usually necessary to observe the items as they are being purchased. Customers usually 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 onsite can be quite costly, as interviewers wait for customers to purchase the key items. The remaining challenge for this method is the potential non-response bias—that is, potential differences between customers willing to have their purchases observed and those who decline.

*Shelf and stocking observations.* To obtain a NTG estimate from the survey-based elasticity estimates, it is necessary to estimate the effect of the program on prices in the region. The price effects can be estimated using in-store shelf and stocking surveys. Alternatively, price effects can be asked about in supplier surveys at various levels.

*Spillover.* Whether revealed preference or stated preference surveys capture spillover depends on the structure of the program and on the study design. For an upstream program, these methods will capture the total change in purchase rate due to the price reduction caused by the program, or due to other factors. For rebate programs, while it is possible for revealed preference surveys combined with discrete choice analysis to tease out the effect of the rebate both on those who take the rebate and those who do not, stated preference surveys for rebate programs are not typically able to make this separation accurately because respondents to this type of survey are not identified as participants or non-participants.[23]

### d. Billing analysis

Billing analysis develops an estimate of program savings by analyzing consumption data. The most common form of such analysis compares usage after program participation with usage prior to program participation, with some form of weather normalization. Analysis may be done separately for each customer and then aggregated, or may be done in a pooled time series cross-sectional model. Since billing analysis typically requires up to 12 months of post-implementation consumption data, this approach may not be feasible where more timely feedback on net savings is required. It works better when the savings are substantial relative to total and when customers are fairly homogenous.

When a comparison group is used in the analysis, the resulting estimate of savings is sometimes interpreted as savings net of free-ridership. The rationale is that the comparison group change in usage represents how the participants would have changed absent the program. However, as discussed further below, bill analysis counts as free ridership any non-participant spillover.

This approach depends heavily on the "comparability" of the comparison group. In most programs, customers who choose to participate are different from those who do not participate, in ways that can

---

[23] Discrete choice analysis simulates the decision to purchase various types of equipment and then uses the model to determine the probability of purchasing high-efficiency equipment absent the program.

affect their year-to-year consumption changes. This self-selection effect can be controlled for, to some extent, by including in the analysis customer characteristics that might be associated with both the propensity to participate and the consumption changes absent the program.

The most complete treatment along these lines is the Inverse Mills ratio self-selection correction developed by J. J. Heckman. Ken Train and Miriam Goldberg expanded this approach to the context of Statistically Adjusted Engineering analysis for program evaluation. However, this approach still assumes that there is no correlation between the change in usage not explained by the model and the remaining unobserved factors that determine program participation. Because there are no data for the unobserved factors that determine program participation, it is not clear whether the assumption of no correlation between change in usage and unobserved factors affects efforts to correct for self-selection.

The assumption that the comparison group change is a good representation of how participants would have changed absent the program can be justified in some cases. A common example is low-income programs, where past and future participants may serve as a comparison for current participants
In cases where there is a credible comparison group and model or analysis structure, the billing analysis provides net savings. The analysis does not isolate a net-to-gross effect from adjustments to gross savings.

A second situation where a valid comparison is available occurs when customers are randomly assigned to receive the participant "treatment" or not. With random assignment, there is no systematic difference between the untreated or control group customers and the participating customers, other than the treatment itself. Therefore, the control group provides an unbiased estimate of what participants would have done absent the program treatment.

This approach is rarely possible in full-scale programs, but is sometimes possible in pilots. The OPower program that has been instituted in several territories in recent years relies on this type of random assignment and comparison of consumption data to determine savings from an informational treatment. Other behavioral programs using this type of approach are under development or consideration.

The random assignment informational programs use many thousands of customers in the treated and control groups, far more than in typical pilots. On the other hand, the ability to estimate savings from these programs requires the use of random assignment, unlike a typical full-scale program.

For any type of program, if the comparison group is otherwise valid, but there is non-participant spillover, net savings estimation based on comparison of participants with non-participants will understate the savings attributable to the program. Essentially, the non-participant spillover will be incorrectly counted as part of naturally occurring savings, and subtracted from participant change, instead of adding to it. Thus, not only is the program not credited with the non-participant spillover, but also the true participant savings attributable to the program are underestimated.

Another effect included in the net savings yielded by billing analysis is "takeback." Takeback is an increase, in usage as a behavioral response to the lower cost of using the equipment after an energy efficiency improvement. Takeback effects arguably should be considered as adjustments to gross savings, not as a net-to-gross adjustment. The net-to-gross factor indicates how much of the gross savings occurred due to the program and would not otherwise have occurred. If there is takeback associated with a measure, that takeback in most cases is an effect of the measure and not program influence on measure implementation. For example, if the effect of adding better insulation is that occupants set their thermostats warmer; this is not an effect of the program but an effect of the measure, whether the measure was program-induced or not.

Like non-participant spillover, takeback is difficult to isolate from other factors affecting participant and non-participant changes. Agnew and Goldberg[24] present analysis exploring the relationship between takeback estimates and engineering assumptions in the context of HVAC measures. Takeback can be measured to some extent via surveys asking what people are doing differently since getting the measure, and to what extent that is because of the measure. This approach means relying again on self-reports.

### e. Structured expert judging

Structured expert judgment studies assemble panels of individuals with close working knowledge of the technology, infrastructure systems, markets, and political environments addressed by a given energy efficiency measure, to estimate baseline market share and, in some cases, forecast market share with and without the program in place. Structured expert judgment processes employ a variety of specific techniques to ensure that the participating experts specify and take into account key known facts about the program, the technologies supported, and the development of other influence factors over time.

The *Delphi process* is the most widely known method of this family of methods. Properly executed, a Delphi process includes at least one iterative step in which each expert is provided with the judgments and rationales of all the other experts, and offered an opportunity to re-assess or defend the original position.

In the context of energy efficiency program evaluation, structured expert judging has as its foundation the experts' experience with other programs and the NTG findings for other programs, typically based on other methods. Structured expert judging allows experience from other contexts to be applied to situations in which all feasible methods may have substantial threats to validity. Expert judging also allows adjustments to be made, albeit subjectively, for some of these threats.

A particularly useful role for structured expert judging is to develop a "consensus" estimate to consolidate results from multiple estimation methods. For example, recognizing weaknesses in every feasible method, the evaluators of the Massachusetts Residential Lighting Program are using a Delphi approach to assess net-to-gross estimates derived from five different methods and to develop a recommended consensus estimate. There are some markets and programs that are not amenable to other methods for estimating net effects, in which case expert judgment, aided by a summary of the history and current state of the market and the program, may be the best available approach.[25]

### f. Historical tracing: case study method

This method involves the careful reconstruction of events leading to the outcome of interest—for example, the launch of a product or the passage of legislation, to develop 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 argument. These include:

- Compiling, comparing, and weighing the merits of narratives of the same set of events provided by individuals with 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.

---

[24] Getting to the Right Delta:  Adjustment and Decomposition of Billing Analysis Results, Agnew, Ken and Goldberg, Miriam, *Proceedings of the 2009 International Energy Program Evaluation Conference.*
[25] See, for example, the 2010 CPUC Residential New Construction Market effects Study by KEMA, NMR, Itron, and Cadmus.

- Positing a number of alternative causal hypotheses and examining their consistency with the narrative fact pattern.

- Assessing the consistency of the observed fact pattern with linkages predicted by the program logic model.

Researchers use information from a wide range of sources to inform historical tracing analyses. These include public and private documents, personal interviews, and surveys.

This method is best suited to attribution analysis of major events such as adoption of new building codes or policies. It is not typically applicable to energy efficiency programs, and is not suggested as a primary method for the Massachusetts programs. However, elements of this approach may be used in analysis of very large custom projects requiring essentially case-study approaches.

### g. Program delivery staff surveys

Some practitioners have used reports from program staff on their influence on customers as input to NTG estimates, particularly for custom projects. An argument for this approach is that customers may be inclined to want to take credit for a good idea, and would not necessarily remember how the program staff helped provide information and develop the project. A critical counter-argument is that program staff have an explicit vested interest in obtaining high attribution credit, which the customers do not. Moreover, program staff do not necessarily realize when the suggestion they made was already something the customer was considering.

Thus, we do not suggest using delivery staff reports of their influence on customers as a valid basis for estimating NTG. We do however suggest obtaining information from delivery staff as background for understanding and clarifying projects and decision points.

### h. Deemed or stipulated NTG estimates

In deemed or stipulated approaches to estimating the net effects of programs, a specific NTG ratio is assumed and applied to the program. Generally, program administrators and regulators draw on available evidence of program impacts, including previous NTG studies, to help decide on the percentage of gross savings that can be claimed by a program. These approaches are particularly useful for forward-looking NTG estimates, possibly building on current and previous estimates, and also relying on program theory and examples from the histories of other programs.

The credibility of the negotiated net savings figure would depend on the type, amount, and quality of the information informing it; such information can include program tracking data, adjusted gross savings estimates, net savings estimates derived from periodic research, sales/shipment data, and more. Although there is always the potential for a deemed estimate to be incorrect, as it is not based on information that is specific to the program and program period, the approach is simple and inexpensive, and avoids depending entirely on controversial measurements of the counterfactual of net savings.

## 3.1.3 Summary of methods

NTG methods can be described in terms of the following:

- Type of approach: Market sales data analysis, self-report counterfactual, pricing and elasticity analysis, billing data analysis, or structured expert judging.

- Who the respondent is or from whom the data are obtained:  participating or non-participating customers, participating or non-participating suppliers at various points in the supply chain, industry experts, etc.

- What types of information are obtained from the respondent.

- How the data are analyzed.

Table 3-1 summarizes the common NTG methods available, based on these elements. For each of these methods, Table 3-2 indicates key research design factors affecting method validity, as well as issues that need to be addressed in the data collection. Whereas the NTG Methods table in the C&I report focused on self-report methodologies, we re-organized the table by type of approach, as some of the methods in the table are not survey-based. This organization better reflects the context of the residential programs, for which evaluators are increasingly turning to methods that do not rely on counterfactual self-report surveys.

**Table 3-1. Common NTG Methods**

| Method | Data Source | Types of data collected | Data Description | Analysis | Includes Like Spillover | Includes Unlike Spillover | Includes Non participant Like Spillover |
|---|---|---|---|---|---|---|---|
| Market sales data analysis | Sales/shipment data provided by industry groups or, ideally, mandated by the federal government | Comprehensive market sales data for program area and comparison area | Sales of efficient and standard equipment in program and non-program areas over time | Weighted/averaged area-to-area comparison, or statistically derived baseline | Yes | No | Yes |
| | Manufacturers & Regional buyers and distributors | Market sales/shipment data | Sales of efficient and standard equipment in program and non-program areas over time | Weighting and/or averaging | Yes | No | Yes |
| | Retail store managers and contractors | Sales data | Sales of efficient and standard equipment when program is and isn't present | Weighting and/or averaging | Yes | No | Yes |
| | Retail store managers and contractors, manufacturers, distributors | Self-reported sales or shipments (not sales/shipment data) | Sales of efficient and standard equipment when program is and isn't present | Weighting and/or averaging | Yes | No | Yes |
| | End-users/decision makers | Self-reported purchases | Self-reported purchases in a specific time period, along with other household behaviors, attitudes, and characteristics, in program and non-program areas | Weighted/averaged area-to-area comparison, or statistically derived baseline | Yes | No | Yes |
| Self-reported counterfactual | Participating and non-participating end-users/decision makers | Post Hoc self-reported counterfactual | Self-reported likelihood of buying absent program assistance, and influence of program on purchases outside of program | Scoring and averaging | Optionally | Optionally | No |
| | Retail store managers and contractors, manufacturers, distributors | Post Hoc self-reported counterfactual | Promotional activity and sales with and without program | Weighting and/or averaging | Yes | No | Yes |
| | | Customer-specific influence—self-reported counterfactual | Promotional activity and sales with and without program, among specific customer groups | Weighting and/or averaging | Typically not | No | No |
| Pricing and Elasticity Analysis | Non-participating end-users | Stated preferences-likelihood | Likelihood of purchase at varying conditions | Scoring and averaging | No | No | No |
| | Participating and non-participating end-users | Revealed preferences | Actual purchases, prices, and customer characteristics | Discrete choice analysis or simple average adoption rates | Optionally | No | Optionally |
| | Retail store | Shelf and stocking observations | Observed shelf volumes and prices | Modeling or averaging | Depends on accompanying elasticity | No | Depends on accompanying elasticity |

| Method | Data Source | Types of data collected | Data Description | Analysis | Includes Like Spillover | Includes Unlike Spillover | Includes Non participant Like Spillover |
|---|---|---|---|---|---|---|---|
| | | | | | analysis | | analysis |
| Billing Data Analysis | Participating customers and comparison group customers | Billing data | Consumption data for roughly one year pre- and post-participation | Weather normalization and change analysis | Yes | Yes | Counted negatively |
| Structured Expert Judging | Experts | Various | NTG estimates from multiple methods, or judging by weight of evidence | Delphi process | Depends on input methods | Depends on input methods | Depends on input methods |

### 3.1.4    Design and data collection considerations

For any research method, details of the situation and of the study design affect the method's validity. The table below indicates design factors and data collection issues that must be considered in implementing each of the NTG methods listed in the previous table. Some of these factors are within the control of the researcher, whereas others cannot be controlled.

**Table 3-2. Design Factors and Data Collection Issues to be Addressed for Common NTG Methods**

| Method | Data Source | Types of data collected | Validity: Depends on— | Data collection issues |
|---|---|---|---|---|
| Market sales data analysis | Sales/shipment data provided by industry groups or, ideally, mandated by the federal government | Comprehensive market sales data for program area and comparison area | Degree of comprehensiveness. Truly comprehensive sales/shipment tracking systems have never been available; if they were, validity would likely be greater than with any other method | With voluntary efforts, some parties often don't cooperate, leaving major holes in data. No mandatory comprehensive tracking system has ever existed. |
| | Manufacturers & Regional buyers and distributors | Market sales/shipment data | ability to construct non-program baseline from pre-program and/or comparison area market coverage | Often some key suppliers don't cooperate, "holes" need to be plugged |
| | Retail store managers and contractors | Sales data | Market coverage and weighting | Often some key suppliers don't cooperate, "holes" need to be plugged |
| | Retail store managers and contractors, manufacturers, distributors | Self-reported sales or shipments (not sales/shipment data) | Market coverage and weighting; Market actor gaming and recall of sales/shipments | Often some key suppliers don't cooperate, "holes" need to be plugged |
| | End-users/decision makers | Self-reported purchases | Verification of measures by on-site auditors and accuracy of respondent recall; Ability to construct non-program baseline from comparison area market coverage | Requires considerable effort to assure consistency of data collection protocols across on-site auditors |
| Self-reported counterfactual | Participating and non-participating end-users/decision makers | Post Hoc self-reported counterfactual | Scoring; accuracy of self-reported hypothetical action absent the program | Requires well designed surveys to minimize response bias. As with all sample surveys, nonresponse bias can also be an issue.. |
| | Retail store managers and contractors, manufacturers, distributors | Post Hoc self-reported counterfactual | Market actor gaming and recall of sales/shipments market coverage and weighting; accuracy of supplier-reported actual and hypothetical activities | Accuracy of supplier's report on influencing factors for customers as a whole; Ensuring a representative sample of suppliers |
| | | Influence of program on purchase decision—self-reported counterfactual from market actors | Market actor gaming and recall of sales/shipments; accuracy of supplier's report on influencing factors for individual customers | Accuracy of supplier's report on influencing factors for individual customers |
| Pricing and Elasticity Analysis | Non-participating end-users | Stated preferences-likelihood | Scoring;  accuracy of self-reported hypothetical actions if purchasing with and without program | Requires companion study to determine change in prices and stocking absent the program |
| | Participating and non- | Revealed preferences | Comparable conditions between purchases in presence | Typically requires in-store intercept or onsite observation, which may |

| Method | Data Source | Types of data collected | Validity: Depends on— | Data collection issues |
|---|---|---|---|---|
| | participating end-users | | of program and those in absence of program; accuracy of purchase and pricing data obtained from customers | require permission from stores |
| | Retail store | Shelf and stocking observations | Stores that represent a meaningful comparison set to those with active programs | Need access to stores; some companies deny access |
| Billing Data Analysis | Participating customers and comparison group customers | Billing data | Valid comparison group with minimal self-selection effects. | Requires obtaining participant and non-participant billing data from utilities |
| Structured Expert Judging | Experts | Various | Well documented methods, Effective iteration process, Experts' expertise; Quality and comprehensiveness of information with a weight-of-evidence approach | Cooperation from a knowledgeable panel |

## 3.2 DECISION FRAMEWORK FOR SELECTING APPROPRIATE METHODOLOGIES

### 3.2.1 General

Key dimensions to be considered in choosing a NTG method are the following:

### a. Availability of market sales data

In general, market-level approaches to estimating net effects require the ability to obtain comprehensive sales data for the measure(s) promoted by a program indicating the total sales of efficient and standard equipment.[26] If sales data are unavailable, or if there are important gaps that would cause estimates based on the available data to be misleading, market-based approaches cannot be used.

### b. Existence of appropriate comparison area

Cross-sectional approaches, including those based on market-level sales data, require the existence of a comparison area that is sufficiently similar to the program area. Many factors can limit the extent to which two areas can be compared, including the existence of similar programs in the comparison state, unique market characteristics in one or both areas, climate differences, significant demographic differences, and others.

While some of these differences can be controlled for in statistical models, for others it is impossible to do so. For example, one of the approaches being used in a study estimating the NTG ratio for lighting programs in MA and other states involves a cross-sectional approach using a regression model. The model controls for several variables, including demographic factors like race and income, as well as area-level factors such as number of Wal-Marts per person and county metropolitan status.[27] By contrast, a cross-sectional approach was considered to be infeasible for a recent market effects study of the California investor owned utilities' residential new construction programs because California has unique building codes, a localized building industry, and various unique climates that do not exist in other areas—factors that cannot be controlled for.[28] Thus, for some programs, market-level approaches cannot be used because there is no valid comparison area, even if such approaches would otherwise be feasible and appropriate.

### c. Features of measures promoted through program

*Homogeneity of measures and customers.* Most of the available methods require large numbers of similar measures and similar customer types. For custom measures, measures with few participants, or measures with applications in widely disparate conditions, methods based on market data or on samples of customers making similar purchase decisions do not easily apply. The only methods that work well for custom or case-specific measures are end-user post-hoc counterfactual surveys, and vendor surveys asking about specific customers.

Some programs are primarily designed to serve as "umbrella programs," or gateways to other programs within a portfolio. Such programs provide customers with recommendations for energy

---

[26] Ideally, sales data of efficient equipment will include the efficiency level of the equipment, such as tier designation under the Consortium for Energy Efficiency (CEE) Super-Efficient Home Appliance Initiative (SEHA). http://www.cee1.org/resid/seha/seha-main.php3

[27] "Results of the Multistate CFL Modeling Effort," NMR Group, February 4, 2010.

[28] "CPUC Residential New Construction Market effects Study," KEMA, NMR, Itron, and Cadmus, 2010.

efficient measures through whole-house audits. Along with the recommendations, customers receive information about incentives available through the program and other programs that promote those measures. Market-level approaches do not work well for these "umbrella programs," because they do not promote large numbers of specific measures; rather, they promote a wide range of measures and largely serve outreach and informational purposes, by facilitating the installation of measures primarily promoted through other programs.

*System-level versus measure-level programs*: For some programs, energy efficiency is promoted not through individual measures, but on a whole-house level in which the building is treated as a system. Multiple energy-efficient measures and practices may act synergistically in a building, leading to efficiency of the house as a whole. The degree of energy efficiency of the system, not the installation of particular measures, is the outcome of interest. Therefore, for such whole-house programs, market sales data cannot be used to estimate net effects.

### d. Likelihood of substantial upstream effects unknown to end-use participant

When there is a reasonable likelihood of substantial upstream effects that an end-use participant would not know about, participating end-user counterfactual surveys alone will understate the effect of the program, even if the customer answers accurately from the customer's own perspective. 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, or increased the sales staff's knowledge of and interest in promoting the efficient option. A related complication of many upstream programs is that participants and non-participants can be difficult to identify for evaluation purposes. Since products are often incentivized upstream (e.g., as markdowns or buydowns to retailers), customers can simply purchase the discounted products without filling out a rebate form or otherwise identifying themselves. These situations either require information for the market as a whole, if the market sales-based approach is viable, or else require input from upstream market actors.

### e. Cost/Value tradeoffs

Some methods can provide more credible results, but are also more costly. This cost may be justified for program components that are important to the portfolio, but not for other components. Importance to the portfolio is typically related to the level of spending and/or savings associated with a program component, but may also depend on future program plans, or other "visibility" factors. In the tables below, we indicate rough levels of cost or difficulty typical for each method, but these can vary with individual circumstances.

### 3.2.2    Methods applicable for different conditions

The table below summarizes which methods are suitable for programs with particular features. The table does not attempt to prescribe a best method for a given situation. Rather, it indicates what options will do better or worse for programs with the indicated features. For a particular program in a particular context, the choice of methods can be made by balancing the advantages and disadvantages of each.

**Table 3-3. Summary of methods applicable to different conditions**

| Method | Data Source | Types of data collected | Applicability | | | | | Typical Cost or Complexity |
| | | | Custom Measures | Whole Building Measures | Programs with few, diverse participants | Prescriptive Measures with large numbers of similar participants | Measures with substantial upstream influence invisible to customers | |
|---|---|---|---|---|---|---|---|---|
| Market sales data analysis | Sales/shipment data provided by industry groups or, ideally, mandated by the federal government | Comprehensive market sales data for program area and comparison area | Poor | Poor | Poor | Good | Good | Low if data are available, High or not possible if data need to be developed |
| | Manufacturers & Regional buyers and distributors | Market sales/shipment data | Poor | Poor | Poor | Good | Good | Low |
| | Retail store managers and contractors | Sales data | Poor | Poor | Poor | Good | Good | Medium |
| | Retail store managers and contractors, manufacturers, distributors | Self-reported sales or shipments (not sales/shipment data) | Poor | Poor | Poor | Medium | Medium | Medium |
| | End-users/decision makers | Self-reported purchases | Poor | Poor | Poor | Good | Good | High |
| Self-reported counterfactual | Participating and non-participating end-users/decision makers | Post Hoc self-reported counterfactual | Good | Good | Good | Good | Poor unless combined with retailer or contractor surveys | Medium |
| | Retail store managers and contractors, manufacturers, distributors | Post Hoc self-reported counterfactual | Poor | Poor | Poor | Good | Good | Low |
| | | Customer-specific influence—self-reported counterfactual | Poor | Poor | Poor | Good | Medium | Low |

| Method | Data Source | Types of data collected | Applicability | | | | | Typical Cost or Complexity |
|---|---|---|---|---|---|---|---|---|
| | | | Custom Measures | Whole Building Measures | Programs with few, diverse participants | Prescriptive Measures with large numbers of similar participants | Measures with substantial upstream influence invisible to customers | |
| Pricing and Elasticity Analysis | Non-participating end-users | Stated preferences-likelihood | Poor | Poor | Poor | Good | Good if combined with pricing study | High |
| | Participating and non-participating end-users | Revealed preferences | Poor | Poor | Poor | Good | Good if combined with pricing study | High |
| | Retail store | Shelf and stocking observations | Poor | Poor | Poor | Good if combined with stated or revealed preference study | Good if combined with stated or revealed preference study | High |
| Billing Data Analysis | Participating customers and comparison group customers | Billing data | Poor | Poor | Poor | Good if have valid comparison group | Good if have valid comparison group | Low |
| Structured Expert Judging | Experts | Various | Depends on quality of input methods | | | | | Low |

### 3.2.3   Application of decision framework to Massachusetts residential programs

The application of immediate interest for the NTG decision framework outlined above is the Massachusetts Residential programs. These programs differ from the Massachusetts C&I programs in several important ways. As noted in the 2010 C&I methodology report, most of the non-residential programs provide custom measures and have relatively few participants. Therefore, market sales data are unlikely to be available for these measures, and econometric methods involving revealed or stated preference data are not well-suited to these programs. Therefore, the primary suggestion for estimating net-to-gross factors for the C&I programs was to rely on participating end-users, plus design team member interviews and supplier surveys. In contrast, many of the residential programs promote prescriptive measures and have large numbers of participants. Sales data are likelier to be available for these measures, and econometric approaches are more viable. Also, end-user self-report approaches are inappropriate or insufficient for upstream programs, because the program's effects on purchases are relatively "invisible" to customers. Therefore, we consider a broader range of methods for estimating net effects for the residential programs, including market-based approaches in addition to surveys of decision-makers and market actors who influence purchase decisions.

Several of the residential programs meet the requirements for market-based approaches, including measure homogeneity, large numbers of participants, and possible valid comparison areas. Such approaches rely on sales data—evidence of what actually happened, as opposed to respondents' ideas about what *might* have happened in hypothetical circumstances. However, the use of these approaches is severely limited by the lack of available market-level sales data. Increasing the availability of comprehensive sales data by requiring participant market actors—or ideally, through a federal mandate, nonparticipant as well as participant market actors—to provide sales data would be a first step toward making such approaches more viable.

An important issue involved in using market-based approaches is that, while these methods estimate the savings that are *realized* in the study period, typically they do not distinguish between savings *caused* by the program in the study period and those caused in prior program periods. This is not a problem in itself because if the same market-based method is used year after year, annual program effects can be tracked over time. However, switching from market-based to other methods, or vice versa, can result in systematic over- or under-reporting of savings. In this case it would be appropriate to use the same methods over time, or, if switching methods, to make sure that the potential overlap or gap is accounted for, which could involve using both sets of methods in at least one year.

Also, when market-based methods are used for the first time to estimate program effects of a mature program over a particular time period, the question arises whether the program should be given the credit for savings that were realized during the program period but may have been caused in prior program years. For example, the Massachusetts residential gas programs are currently moving from an unregulated to a regulated framework and are now required to report NTG estimates. This calls for a policy decision. A resolution of the Energy Efficiency Advisory Council resolution issued on September 8, 2009 establishes January 1, 2010 as the starting point for fulfilling the requirements of the Green Communities Act of 2008. Hence it appears that this is the critical date in the current regulatory framework, and that only savings caused by program activity after that date should be credited to the current programs.

Absent sufficiently comprehensive sales data, a triangulation approach can be used in which estimates from multiple methods are either combined in some way or are presented to an expert panel to come up with an overall estimate. Multiple methods are appropriate in circumstances in which program costs and savings are high, the market is changing rapidly, and/or there is likely to be significant non-participant spillover.

For programs with little likely effect on the broader market, well-designed participant and contractor surveys to elicit the self-reported counterfactual may be adequate. For programs expected to influence

the market non-trivially, beyond the effects on individual participants, the end-user self-reported counterfactual surveys should be combined with supplier surveys (when comprehensive sales data are not available). The supplier survey would collect data on sales volumes and shares with and without the program. This can be done by asking suppliers about current and prior practice, and/or about current sales and hypothetical sales if the program did not exist; alternatively suppliers in a comparison area can be surveyed.

Table 3-4 indicates suggested approaches for each residential program based on the characteristics of programs described in the decision framework—the types and number of measures promoted, the likelihood of substantial market effects, and the availability of sales data and an appropriate comparison area.

As shown in Table 3-4, the *Energy Star Lighting Program* promotes large numbers of similar prescriptive measures. As an upstream program, there is a high likelihood of important influences unknown to customers and a high likelihood of substantial market effects. While these program features preclude the end-user self-report counterfactual approach and point to a market-level approach based on sales data, at this time comprehensive sales data are not available for lighting products. Further, it is difficult to find a valid comparison area to use as a baseline because so many states now have similar lighting programs. Our suggestion is to use multiple methods to derive several NTG estimates. These estimates would be presented to a Delphi panel of experts, who would come to a consensus on the overall NTG estimate for the program. Estimation methods could include: 1) market data analysis on purchase/sales data collected from customer self-reports of purchases and interviews with vendors and suppliers, 2) in-store revealed preferences observations, and 3) shelf and stocking surveys of retail stores, paired with price elasticity analysis.

*The Energy Star Appliances Program* also promotes large numbers of similar prescriptive measures, with a moderate degree of influences unknown to the customer and high potential for market effects. Again, in principle a comprehensive sales-based, cross-sectional approach is appropriate, as was used in NMR's 2004 evaluation of the ES Appliances Program.[29] However, comprehensive sales data, which had been tracked by the Department of Energy since 1998, are no longer available. In their absence, we suggest an approach similar to that suggested for Energy Star Lighting (above), with a Delphi panel of experts estimating the NTG for the program based on a number of estimates derived from multiple methods. However, the market data would come from interviews with supply-side market actors instead of from customers' self-reported purchases. In addition, customer self-reported counterfactual surveys would be included in the estimation methods for the Appliance program, as participants in the program know they are participants and are more likely to be able to report the effect of aspects of the program (e.g., incentives) on their purchases. However, the costs of using multiple methods may be prohibitive considering the relatively modest size of this program. Therefore, it may be more appropriate to use a modified approach involving interviews with supply-side market actors, or customer self-reported counterfactual surveys, or a combination of the two. For the Appliance Retirement Program, estimates could be based on participant self-reported counterfactual surveys, with research on the appliance secondary market helping to inform the counterfactual—what would have happened to the appliances if they had not been picked up by the program.

*MassSave* is primarily an "umbrella program," through which customers receive whole-house efficiency audits and recommendations for measures to improve the efficiency of their home. While some measures (e.g., light bulbs and low-flow showerheads) are installed directly through the program, many of the recommended measures (e.g., weatherization measures such as insulation, air sealing, and heating equipment) are promoted by other residential programs within the portfolio. Therefore, although many of the program's measures are themselves prescriptive, a sales-based approach would not work, as it would not be able to distinguish the effects of the program from those of other programs that primarily promote the same measures. We suggest using customer self-reported counterfactual

---

[29] NMR Group. 2007. Massachusetts Energy Star Appliance Program: Market Share Tracking and Analysis.

surveys, supplemented by input by the auditors and contractors, to gauge the effects of the audit and the incentives on customers' purchase decisions.

The *Weatherization Program* is largely integrated with MassSave. In order to receive incentives for weatherization measures, customers must have a MassSave auditor visit their homes and recommend particular measures, such as duct sealing and air sealing. Free ridership and spillover could be estimated through self-reported counterfactual surveys of MassSave participants who installed weatherization measures, supplemented by input by the auditors and contractors.

*Energy Star Homes,* or *Residential New Construction* promotes efficiency on a whole-building level, not at the level of individual measures. As explained previously, this house-as-a-system approach makes sales-based approaches non-viable. At the same time, a cross-sectional approach will not work because of unique conditions in the local building context, such as building codes and their level of enforcement. Self-report counterfactual surveys with participant builders can be used to estimate free ridership, but this approach will not capture the potentially substantial market effects of the program. To estimate market effects, expert judging can be used. For example, a recent evaluation of the California investor owned utilities' (IOUs') residential new construction (RNC) programs, covering the 2006-2008 program years, examined how the RNC programs could affect the efficiency of California homes built outside those programs. The market effects study examined net impacts achieved in two ways: 1) net impacts achieved through the IOU programs' influence on above-code practices in homes built outside the IOU programs; and 2) net impacts achieved through the IOU programs' influence on increased code compliance in homes built outside the IOU programs. The evaluation team provided two Delphi panels—a panel of Title 24 Consultants[30] and one of industry experts—with gross savings calculated by comparing the efficiency between above-code and just-code homes, and between just-code and below-code homes. Using these gross savings estimates, the experts assigned attribution scores to the RNC programs and other factors to derive net savings estimates.[31]

The *Residential Heating and Cooling* and the *Heating and Hot Water Equipment* programs have similar features—large numbers of prescriptive measures, likelihood of substantial market effects, etc.—and can be treated similarly. Again, a sales-based approach would be ideal for these programs, but such comprehensive data is not currently available. Therefore, self-report surveys of both customers and contractors can be used to estimate free ridership, while market effects can be estimated through interviews with contractors and suppliers in comparison areas. The interviews would gather information on sales levels and market share of efficient and standard equipment.

Market effects of the *Natural Gas Training Programs* could be estimated through the self-report surveys of participants in the training programs. Surveys could collect data on installation practices and other behavioral changes caused by the training programs with energy savings estimated from the behavioral changes.[32] In addition, should budgets permit, it would also be important to employ a simple test/comparison approach with field studies to determine whether training is in fact associated with better practices. In one field study meant to assess the benefits of contractor training, two groups of homes in New Jersey were examined: one with central air conditioning systems installed by North American Technical Excellence (NATE)-trained and certified contractors, and one with systems

---

[30] Title 24 Consultants advise builders and provide certificates of compliance with the energy efficiency portion of the building code for newly constructed homes, as required by California state law.

[31] KEMA, NMR Group, Itron, Cadmus Group. 2010. *Phase II Report Residential New Construction (Single-Family Home) Market Effects Study.*

[32] A recent evaluation of education and training programs in California used this method to estimate net and gross energy savings of the programs. Opinion Dynamics, Wirtshafter Associates, Jai J Mitchell Analytics, Summit Blue Consulting. 2010. *Indirect Impact Evaluation of the Statewide Energy Efficiency Education and Training Program.*
http://www.calmac.org/publications/06-08_Statewide_Education_and_Training_Impact_Eval_Vol_I_FINAL.pdf

installed by contractors who were not NATE certified.[33] The study found no statistically significant difference between the two groups in the quality of refrigerant charge, airflow, and equipment sizing, although there was a significant difference in duct sealing quality (better quality with the certified group).

The *Multi-family Retrofit* program has similar features to the C&I programs—a custom, whole-building approach to energy efficiency, with relatively few participants. Therefore, the methods suggested for the C&I programs are appropriate for this program as well. Specifically, surveys with decision-makers and market actors (e.g., retail store managers, contractors, etc.) can be used to gauge the influence of the program on the customer's purchase decisions.

The *O Power* behavioral pilot is unique among the residential programs in at least two ways that affect how program influence can be estimated. First, it neither promotes particular measures nor provides incentives. Rather, the program aims to change participating customers' energy usage by promoting energy-saving behaviors. Second, the pilot program employs a true experimental design, with a randomly selected control group of non-participants in the same geographical area. This design makes the program ideal for a billing analysis approach comparing the overall energy usage of participants (i.e., those who were randomly selected to receive the Home Energy Reports) with that of the non-participants (a randomly selected group of customers who did not receive the reports) over the same time period. This approach would involve statistical analyses of differences in pre- and post-treatment electric and natural gas consumption by treatment group members (compared to control group households). These analyses yield an estimate of program influence on energy savings net of free ridership. However, non-participant spillover would be counted negatively, as it would decrease the difference in energy use between the treatment group and the control group. Adjustments to non-participant spillover might be possible through surveys with non-participants, gauging their awareness of the program as well as any changes in energy-related behavior as a result of this awareness, might be used to assess the degree to which any non-participant spillover is occurring.

---

[33] Northeast Energy Efficiency Partnerships, Inc., May 2006, Strategies to Increase Residential HVAC Efficiency in the Northeast.

Table 3-4. Suggested Approaches for Massachusetts Residential Programs

| Program | Importance to Portfolio (savings goals) | Measure Homogeneity | Customer Homogeneity | Likelihood of important influences unknown to customer | Likelihood of large NP effects/market effects | Potential for Market data availability | Availability of meaningful comparison area or non-program market baseline | Suggested Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ES Lighting | 37% tot MWh; 16% total BTUs for all res progs | Prescriptive, large numbers | High | High | High | Medium | Possible, but decreasing | Market sales data analysis—sales/ shipment data (if available) | Market sales data analysis-- customer/ decision-maker self-reported purchases | Self-reported counterfactual-- supply-side market actors | Pricing and elasticity analysis-- revealed preferences | Pricing and elasticity-- shelf and stocking survey (w elasticity analysis) | Structured expert judging (if multiple methods used) | |
| ES Appliances | 8% MWh; 3% BTUs | Prescriptive, large numbers | High | Medium (manufacturer and retailer involvement) | High | Medium | Possible | Market sales data analysis sales/ shipment data (if available) | Market sales data analysis-- supply-side market actor self-reported sales/ shipments | Self-reported counterfactual-- supply-side market actors | Pricing and elasticity analysis-- revealed preferences | Pricing and elasticity-- shelf and stocking survey (w elasticity analysis) | Self-reported counterfactual-- participating end-users/decision makers | Structured expert judging (if multiple methods used) |
| MassSave/ Weatherization | 18% MWh; 17% Therms; 18% BTUs | Prescriptive, large measures; Whole-house measures | High | Low | Low | Low | No | Self-reported counterfactual— participating/ non-participating end-users/decision makers | Self-reported counterfactual-- supply-side market actors | | | | | |
| ES Homes | 2% MWh 8% Therms; 5% BTUs | Whole-house measures | High | High | High | Low | No | Structured expert judging (weight of evidence) | | | | | | |
| Residential Heating and Cooling Equipment | 1% MWh; <1% BTUs | Prescriptive, large numbers | High | High | High | High | Possible | Market sales data analysis—sales/ shipment data (if available) | Self-reported counterfactual-- participating/ non-participating end-users/decision makers | Self-reported counterfactual-- supply-side market actors | | | | |

| Program | Importance to Portfolio (savings goals) | Measure Homogeneity | Customer Homogeneity | Likelihood of important influences unknown to customer | Likelihood of large NP effects/market effects | Potential for Market data availability | Availability of meaningful comparison area or non-program market baseline | Suggested Methods | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Heating and Hot Water Heating | 49% Therms; 28% BTUs | Prescriptive, large numbers | High | High | High | High | Possible | Market sales data analysis—sales/ shipment data (if available) | Self-reported counterfactual--participating end-users/decision makers | Self-reported counterfactual--supply-side market actors | | | | |
| Natural gas training programs | N/A | Prescriptive, large numbers | High | High | High | N/A | Possible | Self-reported counterfactual--participating contractors | Field verification of installation practices | | | | | |
| Multi-family Retrofit | 9% MWh 6% Therms; 7% BTUs | Prescriptive, large numbers; Custom measures; Whole-house measures | Medium / Low | Low | Low | Low | No | Self-reported counterfactual — participating/ non-participating end-users/decision makers | Self-reported counterfactual--supply-side market actors | | | | | |
| O-Power | 25% MWh; 20% Therms; 22% BTUs | Behavioral changes | High | Low | Low | N/A | High | Billing data analysis | | | | | | |

# 4. LITERATURE REVIEW ON SELF-REPORT APPROACH (SRA) METHODOLOGIES

Since the objective of the 2010 C&I study was to develop a standardized methodology for situations where end-users are able to report on program impacts via self-report methods, much of the best practice review in that report focused on those methodologies.[34] Since small commercial methodologies are often appropriate for residential customers, much of this methodology could be applied to residential programs as well. To adapt the guidelines to residential programs, we retained those that are equally applicable to residential programs, with minor changes or additions to reflect the residential context, and removed those that are only applicable to C&I sector.

## 4.1 BEST PRACTICES IN SELF-REPORT APPROACH (SRA) METHODOLOGIES

Free-rider and spillover estimates inherently rely on counterfactuals: what would or would not have happened absent a program. Every method of estimating the counterfactual relies on certain assumptions, as well as on data collection.

The literature review found that many self-report attribution techniques are based on sound methodologies and are consistent with analytical methods used in the social sciences[35, 36] Ridge, Willems, Fagan, and Randazzo (2009) point out that it does not make sense to treat all self-report approaches equally, as some conform to best practice, while others do not.[37] Keating (2009) adds that many of the criticisms of the SRA can be alleviated through more careful research design, sampling, survey timing, and question wording.[38]

### 4.1.1 Necessary data elements for implementation of SRA free-ridership and spillover measurement

In order to employ best practice methods, it is critical that the PAs (or program implementers) keep complete and accurate electronic records of data needed to implement the SRA approach. This includes, but is not limited to, contact information on customer decision makers; contact information for contractors and vendors involved with the project; detailed information on services provided (e.g., recommendations from an audit), measure or services incentivized through the program; rebates paid (at the measure level); gross energy and demand savings (at the measure level); date of installation; and project cost. In addition, there should be a unique identifier for each project and a way to link individual measures to those projects. For cases in which participating customers are not tracked (i.e., upstream programs such as lighting), of course, much of this information is not available.

---

[34]   Although some residential programs are also amenable to non-self report approaches, particularly market-level methods using sales data, it is beyond the scope of the current study to provide a literature review on these non-self report methods.

[35]   "Self-report Methods for Estimating Net-to-gross Ratios in California: Honest!", Richard Ridge, Philippus Willems, Jennifer Fagan, paper presented at AESP national conference, San Diego, CA, January 27-29, 2009.

[36]   "Response to Overarching Comments Regarding the Use of Self-Reported Net-to-Gross (NTG) and the Residential and Small Commercial Self-Report Approach NTG Method", paper presented to the California Public Utilities Commission by Members from Residential/Small Commercial Joint Simple NTG (Self-Report) Committee, the Large Nonresidential NTG Committee, and the evaluation contractors, January 28, 2010.

[37]   "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating Net-to-Gross Ratio, Richard Ridge, Phillipus Willems, Jennifer Fagan, Katherine Randazzo", paper presented at the 2009 Energy Program Evaluation Conference, Portland.

[38]   "Free-ridership Borscht: Don't Salt the Soup". Ken Keating, paper presented at the 2009 IEPEC conference.

## 4.1.2   Elements of good design for self-report free-ridership and spillover measurement

The reviewed literature presented a number of best practice elements for survey design, data collection, and analytic methods. The literature also points out the importance of making the whole process transparent. The question sequence, scoring algorithms, and handling of inconsistent and/or missing data should be included in the report so that stakeholders can understand how each question and its responses impacts the final estimate.

Generally, the methodologies reviewed focused on free-ridership SRA measurement, with much less attention given to SRA spillover techniques.

### a.   Survey design elements

Important survey design elements prevalent in the literature for estimating free-ridership (and spillover) include:

- Identifying the key decision-maker(s) for the specific project. This may include the end-user, the builder (for new construction programs), and other market actors (e.g., contractors, vendors, trade allies) who were influential in the decision-making process.

- Use of set-up or warm-up questions to help the decision maker(s) recall past events, the sequence of these events, and how they events affected their decision to adopt the measure.

- Use of multiple questions to limit the potential for misunderstanding or the influence of individual anomalous responses.

- Use of questions to rule out rival hypotheses for installing the efficient equipment.

- Testing questions for validity and reliability.

- Setting up consistency checks within the survey so that inconsistent responses can be clarified immediately.

- Making questions measure-specific in order to improve the respondent's ability to provide concrete answers and to recognize that different measures installed by a customer may have involved different motivations.

- To estimate partial free-ridership and spillover, including questions that capture partial efficiency improvement (accounting for savings above baseline but less than program eligible) where applicable for a measure.

- To estimate partial free-ridership and spillover, including questions that capture program effects on the quantity purchased (accounting for situations where the customer would have installed some of the efficient equipment without the program, but not as much) where applicable for a measure.

- To estimate partial free-ridership and spillover, explicitly asking how the program affected the timing of the measure adoption, (accounting for situations where the measure would have been adopted without the program, but not as soon) where applicable for a measure.

Many of these design considerations also apply to the other kinds of survey efforts discussed in Chapter 2 that can be used to estimate free-ridership (or spillover) surveys.

### b. Data collection elements

Even if stellar survey design elements are present, best practice data collection is key to collecting reliable and valid estimates. Key data collection elements include:

- Pre-testing the survey instrument to ensure questions are understandable, skip patterns are correct, and interview flows smoothly.

- Using techniques to minimize non-response bias, following professional standards for conducting surveys, and training and monitoring telephone interviewers.

- Timing of the data collection should occur as soon as possible after installation, to minimize recall bias and provide timely feedback on program design. Recognize, however, that timely data collection to estimate free-ridership will underestimate participant spillover and will increase data collection costs.

- Sampling a census (or oversample) of the largest savers and measures with few installations, to ensure these are sufficiently represented in the survey sample.

### c. Analytic elements

In addition to survey design elements, much of the literature discussed best practice analytic elements.

- Treatment of acceleration[39] to produce lifetime net savings rather than first-year net savings.

- Incorporating the influence of previous participation in the program. This recognizes that past program participation may have had a positive impact on customers' behaviors and decisions to install additional equipment through the program.

- Establishing a priori rules for treatment of missing/don't knows in the scoring algorithm.

- Weighting the estimates by annual savings to account for the size of the savings impacts for each customer.

- Calculating and reporting the precision of the estimate at the measure level.

- Algorithm characteristics and sensitivity testing of the scoring algorithm. Some of the methodologies reviewed relied on using different sets of questions to develop multiple scores of free-ridership, which are then averaged (or added together).

- Defining what the spillover measurement is and is not attempting to estimate, and why that approach was used.

---

[39] Acceleration is discussed in more detail in Section 3.1.4 below.

# APPENDIX A: TAXONOMY OF PA PROGRAMS BY SECTOR, TYPE OF ASSISTANCE, ELIGIBILITY, INCENTIVES, AND DELIVERY

In order to make suggestions on free-ridership/spillover methodologies that are best tailored to Massachusetts residential program designs and needs, the evaluators attempted to interview residential program managers/staff from several PAs to better understand program delivery. In-depth telephone interviews were completed with residential program managers/staff for each residential non-low income program (except for O Power). Because these program managers/staff generally said they were not well enough informed to answer questions about past free-ridership methodologies and interpretation of results, we sought responses to these questions from program evaluation staff. The following matrix summarizes the responses from these program manager/staff interviews, as well as some additional information from the evaluation staff. Also included in the matrix is information from the MA Joint Statewide Three-year Gas and Electric Energy Efficiency Plans.

Microsoft Office Excel
Worksheet

The attached file will appear as Appendix A in the final report.

# APPENDIX B:  BEST PRACTICE REVIEW SOURCES OF INFORMATION

Cook, Gay, Summit Blue Canada Inc. *Attribution Methodology Wars: Self-Report Methods versus Statistical Number Crunching—Which Should Win?* Paper delivered at ACEEE. 2008.

| Year(s) implemented | 2008 |
| --- | --- |
| Sponsoring agency/utility | ACEEE |
| Key Issues Addressed | Describes various approaches of assessing impact of free riders and spillover on program savings estimates, provides pros/cons of the methods, along with suggestions for determining which method is appropriate for certain types of evaluations. |
| Background | Three methods are commonly used to assess the NTG ratio:<br><br>1. Self-report methods:  Survey of participants and non-participants what they would have done in absence of program support. *Enhanced self-report* methods involve calibrating information from other sources (interviews with vendors, trade allies, review of program documentation, analysis of market-based sales data, etc) with the survey results. The enhanced methods can also determine what additional efficiency improvements participants have made outside the program, how participating vendor sales practices would have been different without the program and how non-participating vendor and customer practices have changed since the program was implemented.<br><br>&bull; Pros: Simplest and lowest cost method is a telephone survey. Can triangulate different perspectives to measure correct construct with increased accuracy; directly addresses behaviors program attempts to affect; and flexible enough to take into account the program-participant interactions.<br><br>&bull; Cons: Can provide biased results; difficult to systematically convert opinions of participants into quantifiable free ridership values; limited participant recall; tends to underestimate spillover; virtually impossible to define a precision target and statistically valid sample size.<br><br>2. Econometric methods:  Application of statistical tools and techniques to economic issues and data to develop models to compare participants' and non-participants' energy usage and demand patterns. Overall pros and cons:<br><br>&bull; Pros: Provides quantitative estimates of magnitude of net impacts; can provide more accuracy because tests for bias and precision can be included.<br><br>&bull; Cons: participants and non-participants included in a model; sample not randomly selected since participants are self-selected; no trade ally effects included.<br><br>&bull; Billing Analysis: Used to calculate annual demand and energy savings<br><br>  a. Pros: Can be used with complex retrofits and controls projects<br><br>  b. Cons: Large customers can have a significant influence; usable sample is reduced to customers with sufficient billing history; does not estimate spillover<br><br>&bull; Econometric models: Used to analyze co-relational relationships, usually with the hope of determining causation<br><br>  a. Pros: Can avoid concern over potential bias and cognitive dissonance issues with survey research; tests for bias and precision can be included; can predict free ridership and spillover. |

|  |  |
|---|---|
|  | b. <u>Cons</u>: Inability to estimate spillover upstream in the distribution channel; robust study requires large budget for evaluation<br><br>• Discrete choice analysis models: Simulates the decision to purchase various types of commercial equipment, and then uses the model to determine the probability of purchasing high-efficiency equipment absent the program.<br><br>3. Market share methods: Market shares approach uses aggregated sales volumes compared to baseline estimates of volume that would have been sold in absence of program. Saturation data analysis uses observations of the share of high efficiency equipment at two points in time.<br><br>    a. <u>Pros</u>: Assesses trends for the entire market; can estimate net energy impacts for programs where participation is not well defined.<br><br>    b. <u>Cons</u>: Collecting reliable saturation data requires a large budget and usually not repeated frequently; difficulty in collecting sales data (vendor concerns of releasing competitive data) and matching available data to unit of analysis (region, utility territory, etc). |
| **Free Ridership Methodologies** | Selection of method depends on:<br><br>• Objectives of program being evaluated<br><br>• Evaluation budget and resources<br><br>• Specific aspects of measure and program participants<br><br>For some programs, methodology selection is straightforward, as in the example below, with self-reported methods preferable.<br><br>Example Comparison of Methods for C&I Custom Programs:<br><br>_(table below)_<br><br>Methods selection may be less straightforward, for example if the screening criteria point towards a market-based approach, but the market data is not available. |
| **Conclusions** | Market share approach is preferable when either there is not good data on participants and/or the goal is to assess market transformation. Surveys with participants and non-participants should be done as part of self-report methods or statistical methods.<br><br>Best approach includes following the guidelines for SR surveys in combination with interviews with other market actors, market share data, etc., to use the survey results in statistical modeling where appropriate and to select the approach that best meets the evaluation goal within the available resources.<br><br>A combination of methods may be used, with certain methods being used on an annual basis and another method being used at certain intervals. |

| Program Characteristic | Self-Report Methods | Statistical Models |
|---|---|---|
| Targets large customers. | In-person or telephone surveys can be used with large customers. | Large customers can overly bias results. |
| Non-participants difficult to identify. | Does not require non-participant data for free riders or inside spillover. | Requires both participants and non-participants in analysis. |
| May not detect savings at whole building/facility level. | Targets measure level information. | Energy use data generally only available at building/facility level. |
| External factors likely to be significant. | Survey accounts for relevant external factors. | Need to collect appropriate data to adjust for external factors. |
| Focused on process changes rather than equipment. | Survey accounts for changes to processes as well as equipment. | Discrete choice and other models focus on equipment choices. |

Cooney, Kevin, Beth Baker, Timea Zentai, and Adam Knickelbein, Summit Blue Consulting, LLC, *Gas Furnace Market Transformation Model Development and Market Research.* Submitted to Energy Trust of Oregon. August 5, 2009. Presented by Fred Gordon at AESP Brownbag. 2010.

| | |
|---|---|
| **Year(s) implemented** | 2009 |
| **Sponsoring agency/utility** | Energy Trust of Oregon |
| **Sector** | Residential: Single family, multifamily and manufactured homes |
| **Key Issues Addressed** | 1. Develop a baseline estimate for the gas furnace market absent the Energy Trust program<br>2. Adapt a current market transformation model for gas furnaces for use with other programs and increase the usability and functionality of the current model |
| **Background** | Several methods were used to develop the inputs to the model:<br>• Secondary data review to facilitate more robust baseline and penetration values<br>• Vendor interviews to help inform the baseline estimate. Trade ally contractors were selected that represented major portion of sales as well as ones who had not been as active. Distributors were selected based on territory. Additional distributor survey collected data on unit sales for the county.<br>• Researching savings associated with federal code changes to assess relationship of utility program to savings achieved through the new standards in their territory.<br>Due to the uncertainty of some of the inputs of the model, two baselines were developed:<br>• *Low case baseline*: Compilation of interviews within county plus national market data<br>• *High case baseline*: National market data alone<br>Two factors change for each scenario: the baseline and number of gas conversions in the service area. Because the high case uses a lower baseline and higher number of conversions, the results show higher attributable energy savings.<br>The model considers only the retrofit market and the units incented under the program, but the baselines penetration values do not differentiate between the retrofit market and new construction market.<br>The market transformation model was originally created in Excel and was updated with user interface enhancements, functionality and updated assumptions and baseline values.<br>There is neither precise data nor perfect sample and it is unlikely there will ever be. Indicators of market change include:<br>• Multiple sources agree that market share is high<br>• Non-participant vendors show high market share<br>• Surrounding territories show high market share<br>• Data over several years shows high market share<br>• A small fraction on sales use program incentive.<br>Sources of information include distributor interviews, contractor interviews, customer free rider questions, studies of nearby areas, parties to federal standards agreement.<br>Keys to agreement of market change include:<br>• Sustained effects |

| | |
|---|---|
| | • Complimentary sources of data pointing in same direction<br>• Both market tracking and causal evidence<br>• Open process and ongoing discussions, acknowledge uncertainty<br>• Input of trade allies and evaluation experts and other key stakeholders<br>• Pick a middle-lower number out of a range of possible savings numbers<br>Only claim to accelerate for a few years. |
| **Conclusions** | NEAA has been doing for a while—occasional market studies, and then annual vendor surveys. |

Fagan, Jennifer, Mike Messenger, and Mike Rufo, Itron, Inc. and Peter Lai, CPUC Energy Division.
*A Meta-Analysis of Net-to-Gross Estimates in California.* AESP. 2009.

| Year(s) | 2009 |
|---|---|
| Sponsoring agency/utility | AESP |
| Key Issues Addressed | This paper provides an overview of estimates of the proportion of free riders in CA, reviews the pros and cons of different net estimation methods for free ridership only and methods used to estimate net market effects that include participant and non-participant effects. |
| Background | The authors claim that understanding net is important for: <br> 1. understanding program and portfolio cost-effectiveness <br> 2. improving portfolio design and resource allocation <br> 3. refining program design and tactics <br> 4. understanding market transformation <br> 5. aligning program administrators' financial interests with societal interests <br> 6. understanding how energy efficiency programs affect baseline load forecasts and short-term power procurement decisions. |
| Methodology | Seven different methods used to estimate NTGR, and their challenges: <br> 1. Customer self-reports. Use for more traditional downstream programs. Challenge for upstream programs is that product buy-downs or instant discounts make the program invisible by design to many customers. <br> 2. Supplier self-reports. Manufacturer/retailers' predictions of product sales with and without the program rebates often been used (especially for upstream rebate programs). Supplier responses may be biased because they realize that giving the right answers can effect continuation of program incentives. <br> 3. Sales based method—per-capita sales comparisons with a comparable state that does not have a program (representing baseline sales). Comparison states must be very similar to program area or normalized statistically for differences in customer and market characteristics. Numerous limitations to this method--available adoption data is not always reliable, nor are data on the necessary normalizing variables always available. <br> 4. Sales based method—paired comparison approach (used in Wisconsin). Comparing energy efficiency product sales data for a leading big box retailer in a state with rebates vs. a similar nearby state without rebates. Challenge is to find a representative retailer and controlling for differing demographics. <br> 5. Econometric—discrete choice analysis. Estimates efficient product purchases as a function of factors that influence energy efficiency demand such as product awareness, prices, and other factors. Can be difficult in rapidly transforming markets where prices and product content are dynamic. Hasn't been used extensively to estimate net impacts due to complexity and expense. It relies on a large body of non-participant survey data (usually 3,000+). <br> 6. Econometric—estimating a demand model. This model predicts the relationship between changes in energy efficient product price, different levels of customer awareness, and energy efficient product sales in different regions of the country (e.g., CA CFL study). <br> 7. Econometric--net billing analysis. Can be used for measures that account for a minimum of 5-10% of end use consumption. Not useful where measures and savings in question are very site specific (e.g., |

| | |
|---|---|
| | industrial customers). |
| | The NTG estimation method to use depends on the specific circumstances and goals of the evaluation. Answering the following questions can help guide the choice:<br><br>• What are the policy goals of the program? For short term resource acquisition, quantifying non-participant spillover and broader market effects may not be of interest. Where market effects are a strong objective, methods 3, 4, and 6 above are usually used.<br><br>• How mature is the program? If program is in infancy and measures are less well-known, a sales based approach is not useful. Mature programs that have been successful and run for at least three years are likely to have some market effects and warrant use of sales based approach. Exception is for products promoted in neighboring jurisdictions for years that may be new to your program (e.g., CFLs).<br><br>• What is the program design?  Upstream or downstream. If upstream, method 2 may not work due to sample bias. If program promotes customized measures only, the econometric and sales based approaches are not feasible, and for industrial billing analysis can't be used.<br><br>• How much budget do I have?  Methods 7, 3, and 4 are less costly than Method 5.<br><br>• What data are available?  Sales based approaches (method 4) rely heavily on publicly available data sources for information. These data are often incomplete. If a good, complete and reliable data source is available, then this approach may be the best choice for assessing the full influence (FR+SO).<br><br>• Is a suitable comparison group available?  Method 4 relies entirely on finding a representative retailer operating in jurisdictions with and without rebates. Methods 3 and 6 also require a good deal of diversity in the market conditions and the availability of non program areas.<br><br>• What level of precision is desired?  Should be based on the needs of policy makers and the program. If need high level but budget is limited, may rule out use of multiple methods or more costly methods.<br><br>• Are there performance-based metrics to be met?<br><br>If there is sufficient budget and available data to support use of more than one method, it is best to use multiple approaches and triangulate the results. |
| **Analysis** | Their review found that despite widespread changes in markets and multitude of NTG methodologies, in general, portfolio-level NTG ratios have been relatively constant since 1980. |
| **Conclusions** | Authors note that challenges associated with measuring net effects of energy efficiency programs are no more daunting than those facing other professions (e.g., education, public health, pharmaceutical, other policy and medical interventions). A wide range of NTG estimates can be derived from the same baseline data depending on NTG definitions, analysis methods, and time frame (immediate past or forecast of near-term future).<br><br>Jurisdictions should cooperate in the collection of sales and market share data for efficiency products to expand available data and reduce evaluation costs.<br><br>Need for oversight agencies to plan for potential market effects of programs operated over long periods of time. |

Goldberg, Miriam L., J. Ryan Barry, Tammy Kuiken, Ben Jones, Paulo Tanimoto, Nicole Buccitelli, Colin Rickert, and Darcy DeAngelo-Woolsey; KEMA, Inc., *Business Programs: Acceleration Treatment and Life Cycle Net Savings.* Submitted to the Public Service Commission of Wisconsin. March 10, 2010.

| Year(s) implemented | 2010 |
| --- | --- |
| Sponsoring agency/utility | Public Service Commission (PSC) of Wisconsin |
| Goal | 1. To review methods utilized by other jurisdictions and investigate the effect of acceleration on Focus attribution results by employing other methods<br>2. To investigate the effects of using life cycle net savings (LCNS) assumptions on the Focus attribution results |
| Key Issues Addressed | The intent of the effects of acceleration treatment analysis was to clarify how much of the difference between Focus and other programs' NTG ratios may be due to differences n the treatment of acceleration when determining program attribution.. |
| Effects of Acceleration Treatment | The study reviewed the attribution methodologies of well-established, large-scale, nonresidential programs in California, Massachusetts, New York, Oregon and Vermont.<br><br>**Comparison of Focus Method to Other Jurisdictions**<br><br>(see table below) |

**Comparison of Focus Method to Other Jurisdictions**

| State | Primary Treatment of Acceleration | Primary Data Collection Technique |
| --- | --- | --- |
| Wisconsin Focus Y1NS | Acceleration less than 48 months receives partial credit towards attribution.<br><br>Acceleration 48 months or more receives full attribution. | Self report participant surveys |
| California[1] | Acceleration less than 6 months receives no acceleration credit.<br><br>Acceleration more than 6 months, but less than 48 months receives partial credit towards attribution.<br><br>Acceleration 48 months or more receives full attribution. | Self report participant surveys |
| Massachusetts[2] | Acceleration more than 12 months receives full attribution.<br><br>No partial acceleration credit given for less than 12 months. | Self report participant surveys |
| New York[3] | Acceleration less than 60 months receives partial credit toward attribution.<br><br>Acceleration 60 months or more receives full attribution. | Self report participant surveys |
| Oregon[4,5] | The evaluation uses the program's effect on timing (yes/no) in developing the scores used to determine attribution.<br><br>The length of the acceleration period is not considered. | Self report participant surveys |
| Vermont[6] | The most recent Efficiency Vermont Program C&I impact evaluation did not attempt to assess attribution. | N/A |

| | |
|---|---|
| | The various methods of acceleration were applied to the Focus evaluation. Results of this analysis indicated that the current Focus evaluation treatment of acceleration provides attribution results comparable to those in other jurisdictions. Final attribution scores are not highly dependent on the acceleration calculation methodology.<br><br>The effects of efficiency and acceleration on attribution are relatively equal: removing partial credit for either causes attribution to decline by roughly 10% versus the current Focus evaluation method.<br><br>The study examined why varying the acceleration treatment had such a limited effect upon attribution. The attribution scores were grouped into categories:<br>• None: an attribution score of zero<br>• Partial: an attribution score between zero and one<br>• Full: an attribution score of one<br>• Market-based: an attribution score determined by a market study<br>Varying the acceleration method only affected the net savings for measures that received partial attribution scores. The measures with no attribution or market-based attribution scores were not affected by changes to the acceleration treatment. The majority of measures with full attribution would not be affected by changing the acceleration method, except for ones with acceleration between 48 and 60 months. For these, the attribution scores would be reduced using the NY acceleration method, but would remain the same under any of the other acceleration methods. |
| **Life Cycle Net Savings (LCNS) Approach** | The LCNS method provides for a different treatment of accelerated projects and produces lifetime net savings instead of the first year net savings that the current Focus method employs. Savings in the LCNS method are based partly on length of time a measure remains operational, so measure life is a key input to this method. The LCNS method does not incorporate a discount rate such as what would be included in a full benefit/cost analysis.<br><br>Similar to the 1st-year method, LCNS calculates attribution as a ratio of net savings to a ratio of verified gross savings, but has two differences:<br><br>1. LCNS looks at the total lifetime savings of the equipment<br>2. LCNS increases the annual verified gross savings in the acceleration period for custom measures where the existing equipment had lower than standard efficiency. In the post-acceleration period and for non-accelerated measures, the annual verified gross savings are the same as those used in the 1st-year method. The ratio of the two savings is referred to as the A/P ratio (Annual savings in acceleration period divided by savings in the post-acceleration period).<br><br>For some measures, the annual gross savings had to be estimated since the input data needed to calculate annual gross savings was not available.<br><br>The study used two different A/P ratios to investigate the uncertainty in the assumption of the A/P ratio and to confirm the robustness of the results. The table below shows the differences between the 1st-year savings method and the two methods of LCNS tested in this study. |

| Assumption | Y1NS | LCNS Method A | LCNS Method B |
|---|---|---|---|
| Type of savings | First year savings | Lifetime savings | |
| Annual acceleration period verified gross savings | The difference between the energy use of the rebated equipment and the energy use of its standard efficiency replacement. | The difference between the energy use of the rebated equipment and the energy use of the equipment replaced. | |
| Annual post-acceleration period verified gross savings | | The difference between the energy use of the rebated equipment and the energy use of its standard efficiency replacement. | |
| Acceleration period net savings | n/a | Acceleration period verified gross savings multiplied by the acceleration period. | |
| Post-acceleration period net savings | n/a | Post-acceleration period verified gross savings times the simple program attribution (SPA). | |
| A/P ratio assumed for custom CATI | n/a (implied 1) | 2 | Based on sector level A/P ratios observed in the engineering survey |
| Net savings calculation | Verified gross savings times [SPA + (Acc/48)(1-SPA)] | Acceleration period net savings plus post-acceleration period net savings | |

Both LCNS methods resulted in lower attribution factors than those calculated using the 1st-year method. The study found the difference was less about the acceleration treatment than the difference between weighting measure attribution by 1st-year versus lifetime savings. The 1st-year method results in a higher savings for shorter-lived measures than on measures with longer lifetimes.

The lower attribution from the LCNS method was apparent across all sectors, but the largest difference was in the Agriculture and Commercial sectors. These sectors had a large amount of savings from CFLs, which receive high market-based attribution scores. A shorter measure life, such as with CFLs, results in less lifetime savings than measures with similar savings with longer lifetimes.

The two methods of LCNS tested resulted in similar attribution factors, when the attributions were rounded to the nearest percent. Custom CATI measures accounted for only a small portion of savings, so the A/P ratio had a limited ability to affect the results.

The paper concludes with a recommendation to the PSCW to consider further development and refinement of the LCNS method. The two main differences between the approaches:

1. The first-year approach treats the reported acceleration period more as an indicator of the likelihood the measure would have been installed without the program rather than as a literal indicator of the time until the measure would have been installed.

2. The first-year approach determines aggregate attribution for a program, sector, or portfolio weighting measures only by first-year savings. The life cycle approach weights measures according to lifetime savings. The first-year approach gives more weight to shorter-lived measures

[1] Nonresidential Net-to-Gross Working Group. *Methodological Framework for Using the Self-report Approach to Estimating Net-to-Gross Ratios for Nonresidential Customers.* February 9, 2009.

[2] Sponsor utilities included National Grid, NSTAR Electric, Northeast Utilities, Unitil, and Cape Light Compact.

[3] NYSERDA. *Annual Report for 2008 – Program Evaluation and Status Report – Issued March 2009*, Section 2.3 Largest Savers Impact Evaluation. December 31, 2008. http://www.nyserda.org/publications/default.asp.

[4] Energy Trust of Oregon, Inc. *Evaluation Committee Report.* May 11, 2007. http://www.energytrust.org/meetings/board/2007/070808/04a_EvalMay.pdf.

[5] ADM Associates, Inc. *Impact Evaluation of New Building Efficiency Program for 2004 and 2005, Final Report.* February 2008.

[6] KEMA, Inc. and RLW Analytics. *Final Report: Phase 2 Evaluation of the Efficiency Vermont Business Program.* February 2006. http://publicservice.vermont.gov/pub/other/evaluationoftheefficiencyvtbusprogfinalreportphase2.pdf.

Goldberg, Miriam, KEMA Inc., Oscar Bloch, Wisconsin Department of Administration, Ralph Prahl, Prahl & Associates, David Sumi and Bryan Ward, PA Consulting Group, and Rick Winch and Tom Talerico, Glacier Consulting Group. *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs.* Prepared for Public Service Commission of Wisconsin. March 16, 2006

| | |
|---|---|
| **Year(s) implemented** | 2006 |
| **Sponsoring agency/utility** | Public Service Commission (PSC) of Wisconsin |
| **Sector** | Residential, Business Programs (Agriculture, Commercial, Industrial, Schools & Government), Renewable Energy Programs |
| **Key Issues Addressed** | Provide a method selection framework to make the rationale for the choice of methods more transparent to users of the evaluations and to provide a greater confidence in the results. |
| **Background** | In 2006, the Focus on Energy (FOE) Evaluation team and the PSC developed a decision tree to guide evaluators in deciding whether a self-reported program response method or market sales-based method would be more appropriate for the evaluations. |

Three steps are involved in choosing a method:

1. Definition of measure groups to be analyzed separately
2. Determining the best net-to-gross (NTG) method for each group
3. Determining the detailed data collection and analysis methods for the NTG method for each group.

The method choice is based on the following considerations:

1. Sales data availability:  Current and baseline market sales data
2. Accuracy of self-reports:  Ability of end-users and/or vendors to report accurately what would have occurred absent the program
3. Likelihood of large non-participant market effects:  likelihood of substantial non-participant market effects, indicating need for methods to capture such effects
4. Narrowness of technology definition: whether the technology to be addressed by a single analysis effort is a single technology or multiple categories of technologies.
5. Uniformity of unit savings:  Whether the savings per unit is sufficiently consistent across types of units and customers that the program effect can be adequately quantified in terms of the total number of units sold, rather than requiring information on unit characteristics by customer type.

**NTG Method Selection Screening Criteria**



| **Methodology** | Different methods can be chosen for different measures within a program area or program. |
|---|---|
| | Method selection is an iterative process based on the way that the groups to be |

| | |
|---|---|
| | analyzed are defined. The first step is defining groups to be analyzed separately, i.e. single measure in a broad market, entire program spanning a broad set of measures in a broad market, etc. Once a method is or methods are assigned for those groups, the groups may be combined to analyze together or split into smaller groups if it is anticipated that different subgroups would lead to different answers. |
| **Eligible respondents** | Market sales-based methods:  Relies on aggregate data on total sales of specific technology in WI<br><br>Self-reported Program Response:  End-users and/or vendors |
| **Types of measures** | Method can be applied to specific technologies or defined groups of technologies |

Itron, Inc. and KEMA. *2004/2005 Statewide Express Efficiency and Upstream HVAC Program Impact Evaluation*. December 31, 2008.

| | |
|---|---|
| **Year(s) implemented** | 2004-2005 |
| **Sponsoring agency/utility** | CPUC |
| **Sector** | Commercial retrofit |
| **Goal** | Encourage the installation of select high efficiency equipment. |
| **Timing of measurement** | Unknown |
| **Eligible respondents** | Participants and non-participants |
| **Types of measures** | Lighting, HVAC, refrigeration, and motors. |
| **Free ridership questions for customers (downstream self-report)** | Two approaches were used to estimate free-ridership, and the two resulting estimates were averaged. <br><br> **Three-criteria approach** consisted of three questions: <br><br> 1. LI42: If the rebate or cash incentive did not exist, which of the following best describes what you would have purchased… <br> • You would NOT have purchased new equipment <br> • You would have purchased fewer new equipment or less new equipment <br> • You would have purchased the same quantity of equipment as you did through the program <br> 2. LI43: If the rebate or cash incentive did not exist, which of the following best describes what you would have purchased… <br> • Standard efficiency equipment or the least expensive alternative available <br> • Less efficient than the equipment we just discussed <br> • The same high efficiency equipment as you purchased through the program <br> 3. LI44: If the rebate or cash incentive did not exist, would you have installed the rebated lighting equipment… <br> • More than 1 year later <br> • Within 1 year <br> • At the same time <br><br> **Program-influence approach** consisted of one question: <br> On a scale of 1-10, with 1 being *NOT AT ALL* Influential and 10 being *EXTREMELY* Influential, how influential was the Express Efficiency program rebate or cash incentive on your decision to install the rebated equipment? |
| **Free ridership algorithm (downstream self report)** | **Three-criteria**--If the respondents states that he or she would have purchased the same quantity and type of equipment, at the same time, and at the same level of efficiency, they are scored as a free rider. Likewise, if the respondent states they would not have purchased the equipment or would have purchased standard equipment, they are scored as a non-free ridership. <br><br> Partial free ridership is scored based on the frequency (0-3) of partial free ridership responses (e.g., would have purchased less new equipment). <br><br> Self-report and discrete choice methods are used to evaluate CFLs, T8s, and AC systems. For CFLs and AC system, the self-report method is used as the evaluators speculate that the discrete choice methodology is not accounting for upstream program effects. <br><br> **Program-influence** --Free ridership was calculated directly from this response, with a 1 indicating a customer was a free rider (FR = 1.00) and 10 indicating a customer was a net participant (FR = 0.00). All other values of free ridership |

| | |
|---|---|
| | were interpolated between these two points using the following equation:<br><br>Free Ridership = 1−(influence rating −1)9<br><br>The average of these two results was taken as the final free-ridership estimate. |
| **Free ridership approach (upstream program)** | An upstream approach was utilized the calculation of NTFR ratios for motors and central air conditioners (CAC). Using in-depth interviews and CATI surveys, participating motor and CAC distributors were asked:<br><br>• "What proportion of the rebated <SPECIFIC CAC/MOTOR MEASURE> you sold in 2004 and 2005 do you think you would have sold in California if you hadn't participated in the program?" This was followed by a confirmation question which read, "Okay, just to confirm – you are saying that <PROPORTION STATED> < SPECIFIC CAC/MOTOR MEASURE> would have been sold anyway in California if the program rebates were not available in 2004 and 2005. Is this correct?"<br><br>Evaluators asked motors distributors to provide free ridership estimates for four different motor size categories and the CAC distributors to provide estimates for five different CAC size/efficiency categories. |
| **Discrete choice modeling** | A discrete choice modeling methodology was used to estimate a net of free ridership model for the non-residential audit-only, rebate program-only and combined-program net of free-ridership ratio for lighting and HVAC equipment measures. |
| **Spillover questions for customers** | Not assessed. |
| **Spillover questions for vendors** | Not assessed. |
| **Spillover algorithm** | NA |

Keating, Kenneth M., PhD. *Free-Ridership Borscht: Don't Salt the Soup.* IEPEC. 2009.

| | |
|---|---|
| Year(s) implemented | 2009 |
| Sponsoring agency/utility | IEPEC |
| Key Issues Addressed | This paper documents the inadvertent bias that multiplicative algorithms can introduce into net-to-gross estimates. |
| Background | In most energy efficiency evaluations, free ridership receives the most attention as stakeholders want to avoid spending ratepayer or public funds on measures or behaviors that would have occurred without those funds. The self-report approach (asking the participants a set of related questions to try determine their motivation) is a popular methodology due to its direct approach and transparent nature. However, when converting the set of responses from participants into a probability of free ridership, evaluations can advertently bias results in one direction by multiplying individual measurements scores together into one summary score. |
| Free Ridership Methodology | Rather than using multiplication, employ averages of individual but similarly scaled measurements or averages of macro indices (such as a four question series of questions). |
| | Multiplication is acceptable if only conditional probabilities are factored and each probability is independent of the others (e.g., efficiency levels cannot be included in timing or quantity probabilities) or if applied to actual savings values (e.g., kWh saved). |
| | This type of self-report protocol is typically used in residential or small commercial evaluations. Large commercial or industrial evaluation requires more detailed inquires as efficiency levels can be assessed easily with a scaled response. |
| Conclusions | Many of the criticisms of the self-report approach can be mitigated through careful research design, sampling, survey timing, and question wording. |
| | Avoid the use multiplication except in carefully worded questions that assess conditional probabilities or when applied to a real world value such as gross savings. |

Keneipp, Marshall, Floyd Keneipp, and Jeff Erickson, Summit Blue Consulting, LLC and Bill Norton, Opinion Dynamics Corp. *APS Measurement, Evaluation, & Research (MER) Report, Consumer Products Program (CPP).* APS. September 30, 2008.

| | |
|---|---|
| **Year(s) implemented** | August 2005 through December 2007 |
| **Sponsoring agency/utility** | APS |
| **Sector** | Residential |
| **Goal** | Program promotes the purchase of high-efficiency ENERGY STAR-rated CFLs through discounted pricing at participating retail outlets |
| **Timing of Measurement** | Surveys conducted in December 2006 and September 2007 |
| **Eligible respondents** | APS customers who purchased CFLs |
| **Type of measures** | CFL bulbs |
| **Free ridership methodology** | Participants were asked the same questions for 2 types of bulbs in a phone survey, with sample data including type of CFL purchased, the amount paid , date of purchase, and store where purchased:<br><br>FR1  Would you have paid up to $[price paid + buydown amount + 50% of buydown] for [desc of product]? *(if yes, skip to 4th question)*<br><br>FR2  Would you have paid up to $[price paid + buydown amount]? *(if yes, skip to 4th question)*<br><br>FR3  Would you have paid up to $[price paid - 50% less than buydown] for [desc of product]?<br><br>FR4  if the [type] CFLs you purchased at [store] on [date] had cost $[price paid +buydown amount]  would you have purchased:<br><br>• More CFLs [+0%]<br>• Definitely the same number [+0%]<br>• Probably the same number [-10%]<br>• Probably fewer [-25%]<br>• Definitely fewer [-50%]<br>• Don't know/don't recall [-0%]<br><br>10.  Prior to purchasing these bulbs were you…<br>  1.  Not at all familiar with CFL bulbs (also called CFLs) [skip to 11)<br>  2.  slightly familiar with CFLs<br>  3.  Somewhat familiar with CFLs<br>  4.  Very familiar with CFLs<br>  5.  (don't know/refused)<br><br>10b. Prior to purchasing these bulbs, would you say that you used …<br>  1.  No CFLs (0%)<br>  2.  Some CFLs<br>  3.  About half CFLs<br>  4.  Mostly CFLs<br>  5.  all CFLs in the screw-in sockets in your home (100%)<br>  6.  (don't know/refused)<br><br>14. Prior to purchasing the bulbs at [store] on [date], had you purchased any CFL bulbs?<br>  1.  Yes, I have purchased CFLs before<br>  2.  No, this was my first CFL purchase |

| | |
|---|---|
| | 3.   (don't' know/don't recall) |
| **Free ridership questions for vendors** | None |
| **Free ridership algorithm** | A participant is initially defined as a 100% free rider if they would have bought the CFLs at an unsubsidized price (by answering yes to either question FR1 or FR2). 100% free riders where then asked question FR4 and their free ridership percentage was adjusted by the percent shown next to the response of question FR4 above. |
| | The NTG survey results were weighted according to the number of CFLs purchased by the respondents to give a savings weighted NTG estimate. Summing gross and net Watts across the surveyed population then dividing the net Watts by the gross Watts gives the final savings-weighted NTG ratio. One minus the NTG ratio is the free ridership percentage. |
| | The initial free ridership rate was discounted based on answers to questions 10, 10b and 14. the discounted free ridership rate was calculated assuming the following were NOT free riders: |
| | • Someone "not at all familiar" with CFLs |
| | • Someone that had used no CFLs before |
| | • Someone that had purchase no CFLs before |
| | The free ridership total was also examined if someone "slightly familiar" with CFLs was considered not a free rider, but that result is less defensible. |
| **Spillover questions for customers** | The participant survey included the following questions: |
| | SO1.   Have you purchased any additional CFLs since the purchase that we've been discussing? |
| | SO2.   [if yes to SO1] Did you receive a discount or did you buy these additional CFLs at a reduced price? |
| | SO3.   Would you have purchased these additional CFLs if you did not have the prior experience of using the CFLs that we've been discussing? |
| | SO4.   How many CFL bulbs have you purchased since [date]? |
| | SO5.   To the best of your knowledge, did the information you received from APS in any way influence your decision to purchase these CFLs? |
| **Spillover questions for vendors** | None |
| **Spillover algorithm** | Questions SO3 [No] and SO5 [Yes] define who purchased CFLs that ought to be counted as spillover for the number in SO4. To extrapolate to the population, the average CFL Watts for respondents were used as a proxy for spillover bulbs. If the number of bulbs was not known, the average of those who did know was used. Summing spillover Watts across the respondents and dividing by gross reported Watts for all respondents yielded spillover as a percent of total reported savings. |

Klos, Mary and Joan Huston, Summit Blue Consulting, LLC. *Impact Evaluation of 2007 CFL Buy-Down Pilot.* Prepared for Progress Energy—Carolinas. May 20, 2008

| | |
|---|---|
| **Year(s) implemented** | 2007-2008 |
| **Sponsoring agency/utility** | Progress Energy—Carolinas (PEC) |
| **Sector** | Residential |
| **Goal** | Pilot program to increase consumer awareness of benefits of ENERGY STAR CFLs by providing educational material and a discounted bulb price to consumers. |
| **Timing of Measurement** | 2 surveys – one shortly after purchase, one four months later |
| **Eligible respondents** | Purchasers of CFL multi-packs |
| **Type of measures** | CFL bulbs in multi-packs |
| **Free ridership questions for customers** | Three questions from the surveys were used to assess free-ridership: <br>• Thinking about the price of the CFL bulbs last fall at Home Depot, on a scale from 1 to 5, where 5 means "very important" and 1 means "not important at all," how important was the sales price in your decision to buy CFL bulbs at that time? (follow-up survey) <br>• Do you already have any CFL bulbs like these in your home? (initial survey) <br>• How many CFL bulbs are you already using in your home an in which rooms are you using them? (initial survey) |
| **Free ridership data from vendors** | Collected CFL sales levels at Home Depot 9 weeks before and during the pilot. |
| **Free ridership algorithm** | Looked at free-ridership in three ways: <br>1. Responses indicating the sales price was of little or no importance in their decision to purchase the bulbs were viewed as free-riders (as a percentage of respondents). <br>2. Respondents who previously had installed CFLs were considered free-riders (percentage of respondents). [Not quantifiable, but low based on large increase in number of CFL bulbs installed per home.] <br>3. Level of product sales (bulbs per week) recorded pre-program were considered the level of free-ridership during the program. <br>The midpoint between the first and the third measures was used as the best estimate for free-ridership. |
| **Spillover questions for customers** | Three questions addressed spillover: <br>• Have you purchased additional CFL bulbs for your home since the special sales price …? <br>• How likely are you to purchase additional CFL bulbs for your home in the future? (1=Very unlikely, 4=Very likely) <br>• Based on your experience with CFL bulbs, how likely would you be to recommend them to family or friends? (1=Very unlikely, 4=Very likely) |
| **Spillover data from vendors** | Collected CFL sales data for before, during and after buy-down pilot |
| **Spillover algorithm** | Four indicators were examined: <br>1. For people that said they purchased additional bulbs, deduced how many additional bulbs were purchased by combining rate of installation, total number of CFLs installed and number of people who said they purchased additional bulbs to calculate average purchase of bulbs per customer who bought additional bulbs. |

|  | Estimated spillover rate from this number. |
|  | 2. Percentage of customers very likely to purchase additional bulbs in the future |
|  | 3. Percentage of customers very likely to recommend CFLs to family or friends |
|  | 4. Comparison of bulb sales before, during and after buy-down pilot. The increase in sales after the event is considered the spillover effect. |
|  | Indicators 2 & 3 were not easily quantifiable to spillover estimate, but high numbers lend support to substantial spillover. |
|  | Indicators 1 & 4 were quantifiable, so a mean of the two was used as the spillover estimate for evaluation purposes. |

Megdal, Lori, Megdal & Associates, LLC, Yogesh Patil, Energy & Resource Solutions, Inc., Cherie Gregoire and Jennifer Meissner, New York State Energy Research and Development Authority, and Kathryn Parlin, West Hill Energy & Computing, Inc. *Feasting at the Ultimate Enhanced*
*Free-Ridership Salad Bar.* IEPEC. 2009

| | |
|---|---|
| **Year(s) implemented** | 2005-2007 |
| **Sponsoring agency/utility** | NYSERDA (2009 IEPEC conference) |
| **Key Issues Addressed** | While unsophisticated batteries of questions with arbitrary scoring will generate NTG ratios at little cost, complex projects require a "salad bar" approach using independent review, mixed modes, and multiple viewpoints. This methodology provides results that "demonstrate construct validity, consistency, and low variation." |
| **Background** | Because net-to-gross ratios were a major point of uncertainty for NYSERDA's large C&I evaluations, the researchers designed an evaluation of 25 of the largest savers in NYSERDA's programs. These programs included: <ul><li>CIPP (Commercial/Industrial Performance Program)</li><li>DG-CHP (Distributed Generation – Combined Heat and Power)</li><li>NCP (New Construction Program).</li><li>PLMP (Peak Load Management Program)</li><li>TA (Technical Assistance)</li></ul> These large savers represented 18 percent of the incremental savings for the entire portfolio. |
| **Free ridership and spillover methodology** | The researchers employed a "salad bar" approach where all respondents received a core set of questions, and select instruments were applied to specific projects as appropriate. This selection was often determined by the decision-making process at each project. An initial telephone interview was conducted to determine the decision-makers for each measure at each site and obtain contact information for those decision-makers. Then, follow-up instruments were administered to the decision-makers either via telephone or during in-person interviews. <br><br> These interviews yielded a wealth of quantitative and qualitative data. This data was then independently reviewed by three senior evaluation analysts. Each analyst determined a free ridership and spillover score for each measure and then teleconferenced with the other analysts to determine a consensus score. These scores included a range of values representing the uncertainty and potential measurement error of using both qualitative and quantitative data. |
| **Response to criticisms of SRA** | The researchers speculate that any social desirability bias found in energy efficiency self-report has limited effect on net-to-gross as research in more socially sensitive errors (drug abuse, sexuality) has shown small underreporting biases with little overall effect. <br><br> Using data from multiple decision-makers limits the effect of any self-report bias or measurement error from one respondent. An in-depth methodology with multiple sources also allows evaluators to weight the value of responses from different decision makers. <br><br> Using a large variety of free ridership questions allows for a detailed comparison of responses across questions. This comparison can enhance and test the consistency of the responses across questions. |

| | |
|---|---|
| **Conclusions** | When evaluating the decision making process at large projects (greater than 1.5 GWh in expected savings), using customized, site-specific methods leads to highly defensible and consistent results. However, this paper did not discuss the costs associated with this methodology – both financial and in terms of respondent burden. In addition, this methodology only presented an estimated range of free ridership; not a point value that is required in many regulatory environments. |
| **Evaluation(s) using method(s)** | NYSERDA Large C&I |

National Action Plan for Energy Efficiency (2007). *Model Energy Efficiency Program Impact Evaluation Guide.* Prepared by Diane Munns and Jim Rogers. <www.epa.gov/eeactionplan>

| | |
|---|---|
| **Year(s) implemented** | 2007 |
| **Sponsoring agency/utility** | EPA |
| **Key Issues Addressed** | Chapter 5 defines net savings, the four key factors that differentiate net and gross savings, provides descriptions of several approaches for determining net savings, and discusses criteria for selecting an appropriate net savings approach. |
| **Background** | There are three primary factors that differentiate gross and net savings: free ridership, spillover, and rebound. Free ridership is the most commonly evaluated net-to-gross factor (NTGR), then spillover, then rebound. |
| | Free riders are participants who would have taken the same action absent the program. The program can also influence the timing, the level of efficiency, and the number of units installed. These different levels of free ridership are referred to as partial or deferred free riders. A non free rider would not have installed the baseline measure without the program. Free ridership can vary from one measure to the next and over time. Free ridership is a source of energy and demand savings uncertainty. |
| | Spillover occurs when there are demand or consumption reductions because of the program, but the program doesn't directly influence the behavior. This may occur because of additional actions participants take outside the program as a result of their participation, changes in the mix of equipment that manufacturers, dealers and contractors offer all customers as a result of program availability, changes in specification practices of architects and engineers, and direct or indirect changes in energy use of non-participants as a result of the program (e.g., advertising, stocking practices, changes in buying habits). |
| | Estimating spillover and free ridership is complicated by market noise, making it difficult to estimate a program's influence. |
| | Rebound occurs when participants increase their use of the equipment as a result of it's improved efficiency. Could argue that there is a non-energy benefit associated with increased comfort, health, and safety. |
| **Free Ridership methodology** | Chapter 5 discusses four approaches for determining the NTGR. All four approaches can be used with any type of program (assuming a large number of participants for econometric). |
| | 1. Self-reporting surveys, which use survey-based stated intentions from participants and non-participants. The best use of self-reporting surveys involves asking a series of questions. Responses to questions are combined (additively or multiplicatively) into an individual free rider estimate. While this is the lowest cost approach, it has disadvantages such as potential bias and overall accuracy. Using techniques like adding consistency check questions can improve survey quality. |
| | 2. Enhanced self-reporting surveys, which combine interviews and other data sources. For example, interviews with multiple decision makers (e.g., managers, engineers, facilities staff, contractors, design engineers, manufacturers, distributors, retailers). Another data source is a project analysis which looks at how the project addresses barriers and/or documentation the participant may have of the decision to proceed (e.g., memos, feasibility studies). Other data sources include market sales data, review of similar programs, market potential or effects studies. |
| | 3. Econometric models to compare participant and non-participant energy and demand. Can only be used with programs having a large number of participants and a comparable non-participant group. Also, the program must be large enough to justify the cost |

|  | of this type of analysis. |
|---|---|
|  | 4. Stipulated net-to-gross ratios, which are multiplied by gross savings to obtain net savings. Typically stipulated by regulatory body when the expense and uncertainty of the results are significant barriers.<br><br>Due to the cost of estimating NTGR, it is acceptable to perform NTGR analyses less frequently (e.g., every few years) than gross savings impact evaluation as long as no major changes in market or behaviors. |

NMR Group, Inc., Research Into Action, Inc. *Net Savings Scoping Paper.* Paper submitted to Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum. 2010.

| Year(s) implemented | 2010 |
|---|---|
| **Sponsoring agency/utility** | Northeast Energy Efficiency Partnerships: Evaluation, Measurement, and Verification Forum. 2010. |
| **Key Issues Addressed** | Outlines various definitions of and measurement approaches to net savings used across the Northeast region, and documents the differing viewpoints of key stakeholders on issues related to net savings. Concludes with several recommendations to the Forum for next steps to resolve the issues discussed in the paper. |
| **Background** | NEEP commissioned the Scoping Paper to explore: 1) the possibility of consistency in defining and measuring net savings, 2) the extent to which current definitions and measurement approaches meet the current and future needs of the various audiences for net savings, and 3) the increasing challenges of attributing impacts to a particular program in the face of multiple programs and other influences promoting the same actions. |
| **Findings** | a. In addition to the established audiences for net savings estimates (energy regulators, program administrators, etc.), an important emerging audience is air regulators. In the near future, approaches to net savings might be required to meet the needs of air regulators in order to translate energy savings to reduction in emissions.<br><br>b. The challenge of attributing impacts to particular programs in the face of multiple other factors influencing the same actions is of great concern to the energy efficiency and air regulation communities. Some experts warned that it is important not to focus too narrowly on which program gets the "credit" for energy savings, while losing sight of the overall goal of reducing energy use. Also, there might be synergistic effects among the different influences.<br><br>c. Most experts who were interviewed supported the idea of having a consistent definition of net savings in the Northeast, but opinions diverged on whether consistent measurement approaches should be prescribed. Some experts thought that measurement approaches need to keep improving before any particular methods are prescribed. |
| **Recommendations and Conclusions** | The authors made several recommendations related to pursuing consistent approaches to defining and measuring net savings throughout the Northeast<br><br>1. Lead the process of developing consistent definitions of adjusted gross savings and net savings in the Northeast Region.<br><br>2. Consider taking action to improve the quality of data used to estimate net savings. This may involve advocating for legal requirements for manufacturers, retailers, and distributors to provide national sales and shipment data for key equipment and products, reported by size and efficiency at the county or state level. It may also involve encouraging program administrators to keep records of program activity by year, including in any possible comparison areas.<br><br>3. Clarify the degree to which programs must differentiate the impact of their activities from the impacts of other programs and efforts designed to bring about the same or |

similar actions or outcomes.

4. Encourage the energy efficiency community to expand its assessment of program success from a focus on net savings to the inclusion of additional factors that may more accurately capture the full range of program impacts, including non-energy impacts such as jobs, improved health, and increased productivity.

5. Forum should decide if it supports the development of consistent methodological approaches to estimating net savings for the Northeast, and, if so, take the actions necessary to develop regional guidelines for consistent methods.

6. Facilitate the development of a working group comprising members of the energy efficiency community, the system planning community, and the air regulation community with the ultimate goal of developing approaches to measuring energy savings and resultant reductions in greenhouse gas emissions in a manner that is mutually acceptable to and feasible for all three communities.

Nonresidential Net-To-Gross Ratio Working Group. *Methodological Framework for Using the Self-Report Approach to Estimating Net-to-Gross Ratios for Nonresidential Customers.* May 8, 2009.

| | |
|---|---|
| **Year(s) implemented** | For the evaluation of 2006-2008 programs |
| **Sponsoring agency/utility** | CA Public Utilities Commission (CPUC) |
| **Sector(s)** | Large nonresidential |
| **Goal** | Provide a standard framework using the Self-Report Approach (SRA) to estimate project and program-level Net-to-Gross Ratios (NTGR). The framework includes decision rules for systematically and consistently integrating findings from both qualitative and quantitative information. |
| **Key Issues Addressed** | <ul><li>The method uses a 0 to 10 scoring system for key questions used to estimate the NTGR, rather than using fixed categories that were assigned weights (as was done previously).</li><li>The method asks respondents to jointly consider and rate the importance of the many likely events or factors that may have influenced their energy efficiency decision making, rather than focusing narrowly on only their rating of the program's importance. This question structure more accurately reflects the complex nature of the real-world decision making and should help to ensure that all non-program influences are reflected in the NTGR assessment in addition to program influences.</li></ul> |
| **Background** | A working group was formed as part of the evaluation of the 2006-2008 programs, tasked with developing a standard methodological framework for estimating net-to-gross ratios. This approach was designed to fully comply with the *California Energy Efficiency Evaluation: Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals* (Protocols) and the *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches* (Guidelines). <br><br>The method is a general framework that can be customized for individual programs. The approach has three levels of free-ridership analysis based on project type. The evaluators determine which projects are assigned to each category. The categories are described as: <br><br>1. **Standard – Very Large Project** NTGR:  Most detailed level of analysis applied to the largest and most complex projects with the greatest expected levels of gross savings. <br>2. **Standard** NTGR:  Somewhat less detailed level of analysis applied to projects with moderately high levels of gross savings. <br>3. **Basic** NTGR: Applied to all remaining projects. <br><br>Each level of analysis relies on up to five sources of information: <br><br>1. Program Files:  Includes documentation such as completed application form(s), correspondence between customer and utility representatives, notes on project details, copies of rebate checks, etc. <br>2. Decision-Maker surveys: A survey is conducted with project decision makers to obtain highly structured responses concerning the probability that the customer would have implemented the same measure absent the program. <br>3. Vendor Surveys:  Completed for all Standard and Standard-Very Large NTGR projects that use vendors, as well as for Basic NTGR projects where customers indicated a high level of influence from vendors on their decision. Vendors include contractors, design engineers, distributors and installers. <br>4. Utility and Program Staff Interviews: Conducted for Standard and |

|  | Basic NTGR | Standard NTGR | Standard -Very Large NTGR |
|---|---|---|---|
| **Program File** | √ | √ | √ |
| **Decision Maker Survey** | Core | Core, Supplemental | Core, Supplemental |
| **Vendor Survey** | √[1] | √[1] | √ |
| **Utility & Program Staff Interviews** | √[2] | √ | √ |
| **Other Research Findings** |  |  | √ |

Standard-Very Large projects to obtain more insight into the customer's decision to implement the project and the extent of the utility's and program's role in the decision, along with vendor contact information.

5. Other Information:   Secondary research is performed for Standard-Very Large projects to obtain information from other sources.

[1]Only performed for sites that indicate a vendor influence score (N3d) greater than maximum of the other program element scores (N3b, N3c, N3g, N3h, N3l).

[2]Only performed for sites that have a utility account representative

| **Timing of measurement** | As close to the project completion as possible |
|---|---|
| **Free ridership methodology** | First, participants are asked about the timing of their program awareness relative to their decision to purchase or implement the energy efficiency measure.<br><br>Next, they are asked to rate the importance of the program versus non-program influences in their decision making.<br><br>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. These include:<br><br>• the age or condition of the equipment,<br>• information from a feasibility study or facility audit<br>• the availability of an incentive or endorsement through the program<br>• a recommendation from an equipment supplier, auditor or consulting engineer<br>• their previous experience with the program or measure,<br>• information from a program-sponsored training course or marketing materials provided by the program<br>• the measure being included as part of a major remodeling project<br>• a recommendation from program staff, a program vendor, or a utility representative<br>• a standard business practice<br>• an internal business procedure or policy<br>• stated concerns about global warming or the environment<br>• a stated desire to achieve energy independence.<br><br>The battery also obtains a description of what the customer would have done absent the program. If the implementation was not an early replacement action, information about the alternative measures is obtained, which is used to adjust the gross engineering savings estimate for partial free ridership.<br><br>The survey is based on a core set of questions for the Basic NTGR, with additional questions for both Standard and Standard-Very Large NTGR projects. The additional questions probe for more detailed information regarding how much of an influence the program had on the decision relative to a customer's internal policies such as financial criteria, corporate policy, and standard practice for implementing such projects. This information is used to |

| | |
|---|---|
| | check for consistency in responses for a project. |
| | Standard-Very Large projects may be subjected to additional questions that arise from review of other information sources. An internally consistent "story" is then created from all of the available data to support the NTGR calculated for Standard-Very Large projects. |
| | Vendors are asked about the program's significance in their decision to recommend the energy efficient measures, and on their likelihood to have recommended the same measure absent the program. |
| **Free ridership algorithm** | The NTG Ratio is calculated as an average of three scores, representing either the highest response or the average of several responses to the relevant question(s) about the decision to install the measure(s). The three scores are:<br><br>1. Timing and Selection score: Reflects the influence from the most important factor in the customer's decision to implement the project at this time. The vendor survey results enter directly into this score for projects where a high level of vendor influence was reported.<br><br>2. Program Influence score: Reflects the relative importance of the influence of the program to non-program factors on the customer's decision to implement the project. The program influence score is divided by 2 if the decision maker reported learning about the program after the decision was made on the specific measures to implement.<br><br>3. No-Program score: Represents the likelihood of actions the customer may have taken or would take in the future absent the project. This score also accounts for deferred free ridership by incorporating the likelihood that the customer would have installed program-qualifying measures at a later date if the program had not been available.<br><br>Each score is based on the maximum score, representing the most important factor in the customer's decision making. High scores that are inconsistent with other responses trigger consistency checks and can lead to follow-up questions to resolve the discrepancy. |
| **Spillover methodology** | None |
| **Spillover algorithm** | N/A |

Peters, Jane S. and Marjorie McRae, Research Into Action, Inc. *Free-Ridership Measurement Is Out of Sync with Program Logic… or, We've Got the Structure Built, but What's Its Foundation?* ACEEE. 2008.

| | |
|---|---|
| **Year(s) implemented** | 2008 |
| **Sponsoring agency/utility** | ACEEE |
| **Key Issues Addressed** | Paper discusses problems associated with free-ridership (FR) measurement in directing energy efficiency programs and policies as well as the measurement of free-riders through participant self-report. Suggest that program policy, design, and implementation decisions can be better informed through analysis of market changes than through FR measurement. |
| **Background** | Currently, most estimation methods rely on participant self-reports obtained during site visits or telephone surveys. Reasons why the authors feel there are problems with current approach: <br><br> 1. Customer decision making model is not appropriate in current program designs that use multiple methods to influence behavior. <br> 2. Psychological theories of how people explain their behaviors to themselves contraindicate the use of current FR methods (attribution theory, cognitive dissonance theory, and social desirability bias in survey responses suggest we are overestimating FR). <br> 3. Intentions are modest predictors of behaviors going forward, and certainly should not be used as predictor of behaviors retrospectively. <br> 4. Increased awareness of global warming will likely result in increases in self-report FR. <br><br> In 1994, Windel and Peters suggested a decision framework would be useful in helping to investigate how participants decided to invest in a measure through a program. This approach has become somewhat standard practice (e.g., 2003 Massachusetts guidelines and guidelines developed they the California Public Utility Commission Energy Division and the Master Evaluation Contractor Team in 2007). <br><br> Friedmann (2007) offers three examples of program designs that would likely accelerate the deployment of energy efficiency technologies and behaviors, but are risky for the utility that does not get cost recovery for savings attributed to FR: <br><br> • Upstream/midstream market programs with incentives to manufacturers and distributors and retailers <br> • Establishing long-term relationships at various levels of both the utility and the large office building manager/owners to enhance energy efficiency uptake. <br> • Addressing data centers with a variety of measures in a holistic manner. |
| **Conclusions** | The more effective EE programs are in changing typical market behavior, the less accurate self-report FR estimation methods are. They propose FR be proxied by the market saturation rate for the efficiency action. Market saturation potentially underestimates FR as those who already want to take the action may seek out the program. Yet, the actions from this group are partially offset by spillover actions. <br><br> The authors suggest policy makers set market targets (% of market share), and do market studies that track the progress toward increased market-share. This leaves open the possibility that incentives may be even more important for later adopters than for early ones. |

Peters, Jane S., Ph.D. and Ryan E. Bliss, Research Into Action. *Fast Feedback Pilot: Existing Buildings and Production Efficiency Programs.* Prepared for Energy Trust of Oregon. March 10, 2010.

| | |
|---|---|
| **Year(s) implemented** | July 2009, January 2010 |
| **Sponsoring agency/utility** | EnergyTrust of Oregon |
| **Sector(s)** | Commercial and Industrial |
| **Goal** | Primary research questions were whether and how the various methods affect completion rates and responses to survey questions. |
| **Key Issues Addressed** | Timing—survey conducted close to completion of project on a rolling basis throughout the year<br>Survey method—Paper, phone, web |
| **Background** | Previous participant surveys had asked respondent to recall details of program-supported projects that had been completed up to two years before. Pilot tested new approach to collecting rapid feedback from program participants and evaluate different survey methods (paper, telephone, and web) from program participants in Energy Trust's Existing Buildings (commercial) and Production Efficiency programs (industrial). |
| **Free-ridership methodology** | Each month, completed projects or projects near completion were assigned to one of three survey methods: paper/phone/web. PE projects requiring on-site verification were assigned to the paper survey; all unverified projects and EB site-verified projects were randomly assigned to phone or web survey method. Paper and phone resulted in higher completion rates than web.<br><br>Believe Fast Track approach results in more accurate free ridership figures than estimates gathered from participants a year or more after project completion. *(NOTE however, that the authors comment that savings weighted free ridership was comparable to the last program evaluation for the Existing Buildings program and somewhat lower for Production Efficiency.)*<br><br>Recommendations:<br>• Continue Fast Feedback Approach with phone method as this provides more immediate feedback to program staff and simplifies data collection/management<br>• Explore and test modifications to the current approach of how projects would have changed without program support (e.g., don't assume that continuing to use existing equipment implies no equipment upgrade).<br>• Expand Fast Feedback approach to all major programs. |
| **Eligible respondents** | Commercial and Industrial program participants. |
| **Types of measures** | HVAC systems, compressed Air, VSDs, motors, pumps, lighting, refrigeration, insulation, renewable, and commercial clothes washers. |
| **Free ridership questions for customers** | The free ridership assessment was based on the methodology developed for the evaluation of the 2006-07 PE program and adapted for the evaluation of the 2006-07 EB program. The assessment consists of 3 elements: 1) how the project would have changed without program assistance; 2) the availability of funds to do the project without program assistance; and 3) the program's influence on the project.<br><br>PROJECT CHANGE QUESTIONS--Respondents were asked how their project would have changed if they had not participated in the program. Responses were coded into one or more of the following categories:<br>1. cancelled the project altogether<br>2. postponed the project more than one year<br>3. repaired existing equipment<br>4. kept using existing equipment<br>5. purchased less expensive equipment |

| | |
|---|---|
| | 6. installed less energy-efficient equipment (slightly, somewhat or significantly less efficient)<br><br>7. reduced the project size or scope<br><br>8. not changed the project at all<br><br>9. don't know<br><br>AVAILABILITY OF PROJECT FUNDS QUESTION--Would firm have made available the funds needed to cover the entire project cost in hadn't received incentive (yes, no, don't know)<br><br>PROGRAM INFLUENCE—Asked to rate the influence on how the program was done for several program elements—the incentive, the installation vendor or contractor, the program representative, and a technical study (if applicable). Five-point scale, with 1 being not at all influential and 5 being extremely influential. |
| **Free ridership questions for vendors** | None |
| **Free ridership algorithm** | Using the above questions, they calculated 2 scores:  the *Project Change Score* (based on the project change questions and the availability of project funds question) and the *Program Influence Score*. Both scores ranged from 0 (no free ridership) to 50 (indicating high free ridership).<br><br>*Project Change Score*—Score of 0=project would postpone more than one year, repair, or continue using existing equipment (without specifying other changes, such as reducing the project scope or using less expensive or less efficient equipment), or use significantly less efficient equipment. Score of 25=respondent would reduce the scope of the project or use less expensive or somewhat less efficient equipment, or indicated some change but did not indicate what would have been done. Score of 50=respondents would do the project exactly the same or would use slightly less efficient equipment.<br><br>*Program Influence Score*—Score based on the highest rated influence from among the program incentive, the program representative, and the technical study if one was performed. Score of 0=high program influence; a score of 25=moderate program influence; and a score of 50=low program influence.<br><br>These scores were then summed with a resulting sum score ranging in value from 0 to 100. The scores were interpreted as a percentage, indicating a range from no to total free-ridership.<br><br>In cases where there was insufficient data to calculate one or both scores, they calculated 2 free ridership scores:  1) a low-scenario score, which assuming that the missing score was 0, and 2) the high scenario score, which assumed that the missing score was 50. To calculate a mean free ridership across all respondents, they also calculated a third free ridership score, which was the mid-point of the low scenario and high scenario scores.<br><br>Free-ridership scores across all respondents were reported as means of the low-scenario, mid-point scenario or high-scenario scores. Respondents without missing data had the same scores included in the mean calculations for all three scenarios. Thus, the magnitude of the range between the low-scenario score and the high-scenario score was based on the number of respondents with missing data.<br><br>Influence ratings were largely unrelated to survey method.<br><br>Requires good coordination between program staff and evaluator to receive monthly project information.<br><br>Increased expense as method requires monthly cleaning of project information for purposes of sampling, as well as increased costs for monthly survey management. |
| **Spillover questions for customers** | None |
| **Spillover questions for vendors** | None |

| **Spillover algorithm** | NA |
|---|---|

Prahl, Ralph, Prahl & Associates, Goldberg, Miriam and Bobbi Tannenbaum, KEMA Inc, David Sumi and Bryan Ward, PA Consulting Group, and Tom Talerico and Rick Winch, Glacier Consulting Group. *Integrating Supply-Side Results with End-User Net-to-Gross Self Reports.* Memorandum prepared for the Public Service Commission of Wisconsin. July 2, 2008.

| | |
|---|---|
| **Year(s) implemented** | 2008 |
| **Sponsoring agency/utility** | Public Service Commission (PSC) of Wisconsin |
| **Key Issues Addressed** | Establish a framework for the performance of supply-side (SS) research in order to supplement end-user self-reports (SR) done as the primary approach to NTG analysis |
| | Clarifies the approach when the Selection Framework has resulted in the use of end-user self-reports yet there is thought to be the potential for supply-side effects to call the veracity of the results into question |
| **Background** | The framework is intended to capture all potential situations on a continuum ranging from changes in the behavior of vendors directly participating in the program to changes in the behavior of all vendors in the market or markets targeted by the program (participating and non-participating). |
| | Criteria to decide whether to perform supplemental supply-side research includes: |
| | 1. The existence of a *plausible, credible, and specific* program theory predicting supply-side program effects, or some other sound logical or empirical basis for believing they are likely to exist. |
| | 2. Likelihood that the predicted effects can be meaningfully assessed through empirical research. |
| | 3. Likelihood that the needed research can be performed at reasonable cost, relative to the available budget and the likely impact. The burden of proof in point #1 above becomes stronger the more expensive the issue would be to research. |
| | Key provisos; |
| | 1. Not all SS effects imply immediate energy savings |
| | 2. It may not always be feasible to quantitatively integrate supply- and demand-side results for the following reasons: |
| | • Unavailability of sales data |
| | • Vendors and end-users may have perspectives that are difficult to reconcile quantitatively |
| | • May not be able to afford enough interviews with multiple categories of market actors to get a reliable picture of what is going on |
| | • Difficult to avoid double-counting when both supply- and demand-side savings are attributed to program. |
| | 3. When not possible to quantitatively integrate SS results with end-user SRs, qualitative conclusions will be drawn and presented to the PSC and the PSC can consider them in deciding how much credit to give a program |
| | 4. When SS and SR results are integrated, important to acknowledge the resulting uncertainties. |
| **Integration Approaches** | Different categories of integration of SS results with end-user SR: |
| | 1. Making no changes to end-user self-report result |
| | • May be appropriate when supply-side results constitute leading indicators, or are not sufficient to conclude there are energy savings |
| | 2. Altering NTG approach for the next round |
| | • Use when there is evidence of supply-side effects, but not practical |

to incorporate the results into the current NTG analysis

- Particularly appropriate when current energy savings are likely to be modest, but seem likely to increase over time
- May be appropriate if supply-side research is exploratory, but yields results suggesting that more rigorous analysis may provide improved NTG estimates

3. Using supply-side results to refine end-user self-report battery

- Example: program is found to have reduced incremental cost of energy efficient measures; ask end-users about WTP at higher price

4. Disentangling Supply- and Demand-Side Impacts

- Use analytical techniques to eliminate or correct for overlap in the two net savings estimates, then adding them together.

5. Altering the self-report interpretative algorithm

- Appropriate when supply-side results suggest significant changes in supply-side conditions of which end-users are unlikely to be aware.
- Can override certain end-user SR responses based on SS results

6. Adjusting the self-report net-to-gross result

- Prime example: supply-side research yields strong evidence of non-participant spillover effects, but does not call into question validity of end-user self-report responses

7. Not integrating the demand- and SS results

- Evidence of SS effects, but none of the above approaches are appropriate

Rathbun, Pam, Carol Sabo, and Bryan Zent. PA Consulting Group. *Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised).* Prepared for the Massachusetts Utilities, June 13, 2003.

| Year(s) implemented | 2003 |
|---|---|
| Sponsoring agency/utility | National Grid, NSTAR Electric, Northeast Utilities, Unitil, Cape Light Compact: |
| Sector(s) | Commercial, Industrial |
| Goal | Develop standardized methods to be used by all the sponsors to determine free-ridership and spillover factors for C&I programs. |
| Background | Previous studies in MA had used independent evaluation approaches, with varied survey instruments, analysis techniques and assumptions. This report represents a collaborative effort between sponsors and evaluators in developing standardized sampling techniques, data collection approaches, survey questions, survey instruments and an analysis methodology. |
| | The evaluation methodology is an assessment of the annual program impacts including disaggregated values for free-ridership and spillover. |
| Methodology | Recommendations for standardization were made in the following areas: |

**PRE-SURVEY PREPARTION**

- Program application data should be maintained in an electronic database
- Additional data should be collected at project closure to make it easier to identify the appropriate decision maker(s) and vendors.

**SAMPLE DESIGN**

- Samples should be designed to achieve a minimum of +/- 10% precision level at the 90% confidence level at the end-use measure category level.
- A census of measures (i.e., all measures) should be included for end-use measure categories with less than 50 installations in order to achieve the minimum +/- 10% precision level. Precision levels worsen dramatically as the number of measures in the population decreases from 50
- For categories with more than 50 installations, a stratified sampling strategy should be implemented, sampling all top 10% savings sites plus a random sample of the remaining sites to achieve the minimum precision levels.
- In all cases, a 60% response rate should be considered as a minimum goal when conducting the surveys.
- Customers with multiple measure categories should only be asked about their decisions regarding two measures, with priority given to the more rare measures and/or measures with the largest savings.

**SURVEY IMPLEMENTATION**

- Surveys should be conducted within a year of participation
- Surveys should be administered via telephone (though large, custom or industrial projects may require on-site surveys) by professional interviewers.
- An advance letter on Sponsor letterhead explaining the study should be mailed to all sampled program participants prior to the survey.
- The key decision-maker(s) must be identified when conducting the surveys.
- When evaluating a customer's free-ridership rate, it may be necessary to interview the design professional or vendor involved in the project.

| | |
|---|---|
| | **ANALYSIS:** <br> • Free-ridership and spillover estimation should be conducted annually after the end of the program year. <br> • Free-ridership and spillover estimation should be conducted at the specific end-use measure category level. <br> • Completed surveys must be weighted to account for disproportional sampling probability and non-response so the results represent the population of measures. That is, statistical expansion methods appropriate to the sampling process are required. <br> • With standardized methods to estimate free-ridership and spillover, Sponsors can also calculate net program savings in a consistent manner: <br><br> *Net Savings = (Gross Savings) * (RR) * (1-FR+PS+NPS)* <br> *=(Gross Savings) x (RR) x (1-FR+SO), where* <br><br> *RR is the realization rate (evaluated/tracking), FR is the free-rider fraction, PS is the participant "like" spillover fraction, and NPS is the non-participant spillover fraction. SO (total spillover) is the sum of PS and NPS.* Variations on survey wording may be necessary to fit key decision-making groups that vary by program type. |
| **Timing of measurement** | Should be conducted annually after the end of the program year |
| **Free ridership questions for customers** | • Identification of key decision maker <br> • Project and decision-making review questions:  Intended as warm-up/context questions <br> • Reminder of what incentive and services (e.g., technical assistance) were received <br> • Timing:  whether or not any measures would have been implemented within one year without incentive <br> • Quantity:  whether they would have purchased exact same quantity in absence of incentive <br> • Program Efficiency:  what percent of installed measures would have been of same efficiency without incentive <br> • Cost: whether the company would have paid for the same installed measures in absence of program <br> • Consistency checks for measures initially assigned free-ridership of 0% or 100% <br> • Technical Assessment Study impact question <br> • Past program participation impact questions |
| **Free ridership questions for vendors** | Confirmation that key decision maker was design professional. <br> Rest of questions are parallel to the customer questions |
| **Free ridership algorithm** | Free-ridership (both pure and partial), using a customer survey: <br> • Definition: customer who received an incentive who would have installed the same or smaller quantity on their own within one year if the program had not been offered <br> • Calculation addresses the full range of total free-ridership (0% to 100%), based on the quantity and efficiency of any equipment that would have been installed outside the program <br> • Other factors such as utility-sponsored technical assessments and the influence of past program participation are considered in the free-ridership rate calculation |
| **Spillover questions for customers** | Measured only like spillover--questions probe on recent purchases and the similarity to measures installed from the program, and the influence past installation through the program had on the decision. |

| Spillover questions for vendors | Four steps are used to determine non-participant "like" spillover:<br><br>1. For each design professional/vendor, the survey determines the percentage of all program-eligible equipment sold/installed outside the program in the Sponsor's service territory.<br><br>2. For each design professional/vendor, the survey determines whether the sale or installation of program-eligible equipment outside the program was due to the program (non-participant spillover).<br><br>3. For each design professional/vendor, savings associated with this "non-participant spillover" equipment are determined by examining the participant database and quantities installed.<br><br>4. Non-participant spillover savings are then extrapolated from the survey to the total program savings in the year. |
|---|---|
| Spillover algorithm | Participant like-measure spillover, using a customer survey:<br><br>• Definition: customer who installed equipment through the program in the past year and then installed additional equipment of the same type ("like")<br><br>• Calculation assigns weights based on the source of the spillover – experience with the equipment, participation in any past program, or recommendation from design professional<br><br>Non-participant like-measure spillover, using a survey of participating design professionals and vendors:<br><br>• Definition: energy efficiency measures installed by program non-participants due to program influence, based on responses from design professionals and vendors participating in the program<br><br>• Calculations based on fraction of equipment receiving no program incentive and the program's influence on the design professional/vendor's decision to recommend equipment<br><br>• The program tracking system database can be used to attach kWh savings estimates to non-participant spillover if the database contains this information |

Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches.* October 15, 2007.

| Year(s) | 2007 |
|---|---|
| Sponsoring agency/utility | California Public Utilities Commission (CPUC) |
| Key Issues Addressed | Provides basic methodological guidelines that are considered best practice in the social science and engineering communities that evaluators should use to assess net impacts and spillover. |
| Background | The CPUC adopted the *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professions* (TecMarket Works, 2006) for the measurement and evaluation of energy efficiency programs. In certain situations, the Protocols allow for use of self-report approach (SRA) to estimate NTG for the basic, standard, and enhance levels of rigor. The protocols recognize that using both quantitative and qualitative data can be used to assess causality.<br><br>However, the Protocols are silent regarding basic methodological guidelines that are considered best practice. |
| Free Ridership Methodology | The SRA deviates from the standard approach to assessing causality (experiment or quasi-experiment), which are not always desirable or possible. The SRA guidelines are as follows:<br><br>1. Timing of the interview—as soon after installation as possible.<br>2. Identifying the correct respondent(s)—is critical. For large C&I is complicated as different actors have different and complementary pieces of information about the decision making; decisions may be being made in regional or national headquarters; decision making may be done by commissions, committees, boards or councils; and there may be both a technical and a financial decision maker.<br>3. Set up Questions—Need to adequately establish the context and sequence of events that led to decisions.<br>4. Use of multiple questions—uses both quantitative and qualitative questions to measure a construct such as free ridership is preferable as reliability is increased.<br>5. Validity and reliability—should be assessed for each question used. Also the internal consistency of multiple-item scales should be tested. For large savings sites, multiple members of the evaluation team should review the results to ensure consistency.<br>6. Consistency checks—set up checks for inconsistencies and establish rules for handling inconsistent responses while the respondent is on the phone.<br>7. Make questions measure-specific<br>8. Partial free-ridership—should explore cases where participant would have installed something more efficient than program assumed baseline but not as efficient as program equipment.<br>9. Deferred free-ridership—need to measure program's impact on acceleration of installation. Should use more than one question and use preponderance of evidence approach.<br>10. Scoring algorithms—can impact results and must be documented and tested using sensitivity analyses. Preponderance of evidence approach is better than relying solely on an algorithm.<br>11. Handling "don't knows" and non-responses—need to determine in advance how these will be handled. Make a special effort to avoid don't know responses.<br>12. Weighting the NTGR—Need to take into account the size of the savings impacts at the customer or project level.<br>13. Ruling out rival hypotheses—need to ask open-ended questions |

regarding other possible reasons for installing the efficient equipment.

14. Precision of the estimated NTGR—need to report this but is complicated when have multiple sources of information or multiple respondents. In this case need to take into account the propagation of errors in the relative precision.

15. Pretesting the instrument—always pretest to reveal ambiguous wording, faulty skip patterns, leading questions, faulty consistency checks and incorrect sequencing of questions.

16. (Large Savers, complex decision making)  Incorporation of additional quantitative and qualitative data in estimating the NTGR.

- use multiple respondents
- use other site- and market-level data
- establish rules for data integration
- analysis method (case studies are one method for assessing both quantitative and qualitative data, content analysis to identify coherent and important these and patterns in the data). Use of multiple evaluators to independently review the data.

17. Qualified interviewers—For complex situations, engineers familiar with the more complicated technologies should be trained to collect the data.
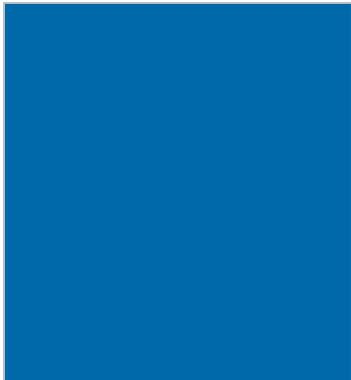
Ridge, Richard, Ridge & Associates, Phillipus Willems, PWP Inc, Jennifer Fagan, Itron, Inc., and Katherine Randazzo, KVD Research Consulting. *The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio.* IEPEC. 2009.

| Year(s) implemented | 2009 |
|---|---|
| Sponsoring agency/utility | Energy Program Evaluation Conference, Portland |
| Key Issues Addressed | While the SRA approach has been used for over 30 years, because it does not involve any formal comparison groups it has been criticized as inherently biased and unreliable. This paper discusses the role of the self-report approach (SRA) within the larger evaluation framework, improvements to the SRA over time in both its internal validity and reliability, and responds to more common criticisms of the California SRA. |
| Background | Which technique is used to estimate net energy and demand impacts depends on a number of factors, including time, money, data availability, and effect size. For example, the expected magnitude of savings for a given program might not warrant the investment in a billing analysis or discrete choice analysis. Or, key stakeholders might not want to wait for a billing analysis, which typically requires up to 12 months of post consumption data. And with a small signal to noise ratio, the sample sizes necessary for the required statistical power can be prohibitively expensive. |
| | In addition, there are situations in which the standard quantitative approaches involving comparison groups are not always possible. For example, in the industrial sector there are 3 barriers—low signal to noise ratio in a part./nonpart. billing analysis, industrial customers have participated in energy efficiency programs in prior years making it difficult to find true non-participants, and large industrial customers are each unique making it unlikely that one could find a matching nonpart. group. In new construction programs, many of the large res. and nonres. developers, architects and engineering firms have also been contaminated by prior participation in energy efficiency programs. The authors believe that discrete choice, difference-of-differences, and econometric modeling approaches have become over time increasingly unreliable and implausible in these cases and that SRA is the most appropriate for evaluating this complex and diverse program and market. |
| | In 1993, California formally recognized that methods involving comparison groups were not always feasible. The1993 California Protocols allowed the SRA as one way to estimate the NTGR. The 2005 Protocols expanded on the SRA approach and explicitly required triangulation for programs assigned the enhanced level of evaluation rigor understanding that there is error associated with any single method. Triangulation uses a variety of research methods and data sources to reduce the risk of systematic biases. |
| Free Ridership Methodologies | The establishment of a causal connection between the program and customer behavior is at the core of arguments for and against research methods aimed at establishing program impacts. One strategy is to develop and assess rival hypotheses to guard against threats to validity. Sound program theories and logic models can provide valuable assistance in identifying the plausible rival hypotheses. Approaches to demonstrating causality using non-experimental designs could include case studies or SRA. |
| | In 2007, the Energy Division published the *Guidelines for Estimating Net-to-Gross Ratios Using the Self-Report Approaches* which contained 17 recommendations for further improving the validity and reliability of the CA-SRA. In 2007, the CPUC also formed two groups (the Residential and Non-Residential NTGR Working Groups) of nationally recognized evaluators to consolidate the lessons learned over the last 15 years in order to make further improvements in the CA-SRA. |
| | The CA-SRA involves asking one or more key decision makers a series of closed and open-ended questions about their motivations for installing the program-eligible equipment, about what they would have done absent the program, as well as questions that attempt to rule out rival explanations for the installation. In the |

| | |
|---|---|
| | simplest case (e.g., residential customers), the CA-SRA is based mainly on quantitative data. In more complex cases in nonresidential programs, the CA-SRA is strengthened by including additional quantitative and qualitative data (e.g., in-depth open-ended interviews, direct observation, review of customer and program records). |
| **Response to criticisms of SRA** | The authors note a number of criticisms of the SRA and provide their responses to those criticisms.<br><br>1. Legitimacy—CA-SRA is a legitimate social science tool for establishing causality.<br><br>2. Turbulent environment—In any evaluation, as the number of alternative hypotheses grows, the task of teasing out the single intervention effect becomes more challenging. This is the case regardless of what method is used.<br><br>3. Nonlinear approach—the CA-SRA recognizes that the route is nonlinear by attempting to identify other parties most important in a customer's decision to participate and uncover various ways the program might have influenced these market factors.<br><br>4. Recall—Interviews should be conducted with the decision maker as soon after the installation of equipment as possible.<br><br>5. Subjective—the CA-SRA collects a variety of qualitative and quantitative evidence so it is not merely subjective.<br><br>6. Treating ordinal data as interval—there is strong support in the social science literature that treating ordinal scales as interval data yields results that are both meaningful and useful to decision makers, and there is no reason to think that this measurement is not randomly distributed.<br><br>7. The meaning and calculation of NTGR—algorithms and weights must be developed by experienced professionals who understand that these have to be transparent, plausible and defensible and must be subjected to sensitivity analysis.<br><br>8. Socially desirable responses—Methods have been developed to address this potential source of bias. These methods have been incorporated into the CA-SRA.<br><br>9. Stated intentions—The CA-SRA collects a variety of information to measure the counterfactual, get at the main reasons for installing the efficient equipment, and establish the temporal precedence of the program. For more complex projects additional information is gathered from vendors and file review. |
| **Conclusions** | The paper concludes by saying it doesn't make sense to compare all SRA approaches equally as some conform to best practices (CA-SRA) and others don't.<br><br>For projects with substantial savings that have been assigned the enhanced level of rigor, the 2005 Protocols require that two or more approaches of the available three (discrete choice with a comparison group, billing analysis with a comparison group, and the CA-SRS) must be used. For programs that have been assigned the standard or basic level or rigor and for which methods involving comparison groups are impossible, the CA-SRA can provide sufficiently rigorous estimates of NTGR.<br><br>Any set of rewards and penalties should never require a level of accuracy that exceeds the ability of any evaluators to provide. This is unreasonable burden on evaluators and results in a continuing, contentious and unproductive relationship among implementers and regulators. |

Saxonis, William P. New York State Department of Public Service. *Free-Ridership and Spillover:*
*A Regulatory Dilemma.* IEPEC. 2007.

| Year(s) | 2007 |
|---|---|
| Sponsoring agency/utility | Energy Program Evaluation Conference, Chicago |
| Key Issues Addressed | The paper examines free rider and spillover results from energy efficiency programs administered by NYSERDA, which found that free rider rates for C&I programs ranged from 10-67% and spillover rates ranged from 19-168%. For residential programs, the free ridership ranged from 2-28 percent and spillover from 5-48 percent. Paper looks at FR and SO measurement in a historical context, compares NYSERDA results to other states, and concludes with practical recommendations. |
| Background | In NY, NTG=(1-FR)+SO<br><br>Regulators need reliable estimates for 3 reasons:<br><br>1.  Protect ratepayers economic interests<br>2.  A secure supply of electricity.<br>3.  Environmental<br><br>Spillover measurement trails free ridership measurement in the level of research attention and the level of confidence in the reliability of the results. |
| Methodology | Due to size and scope of the NYSERDA program portfolio there may be some variations in the following general FR approach used by NYSERDA:<br><br>• Directly ask participants if they would have implemented the same energy efficiency measure without program assistance<br>• Ask quantitative and open-end questions regarding program influence<br>• Score open end responses using an established formula to capture the degree of FR based on factors such as the timing of installation, quantity, and efficiency.<br><br>The spillover approach:<br><br>• Multi-question survey approach similar to FR methodology with participating customers/vendors.<br>• For some programs, non-participants were surveyed to determine any influence of the program on their energy efficiency related behavior.<br><br>For some evaluations, used an "integrated data collection process" to gain participant feedback in near real time to supplement retrospective survey efforts. Participants were asked to complete an abbreviated survey containing questions related to program attribution soon after their participation. This approach is useful for identifying trends and confirming FR and SO values in between major evaluation cycles. |
| Analysis | Despite the high FR and SO rates in NYSERDA programs (especially compared to other programs in the region), the impact on NTGR is virtually non-existent. |
| Conclusions | 1.  Improve data reliability—has been little research to quantify FR and SO results using multiple approaches in the same study. This would help increase confidence in the data if they produce similar results.<br><br>Need to increase methods to triangulate the data, increase the precision and CI of surveys, employ more long term and comparative analysis (especially for SO), conduct studies that compare adoption of ee products in regions with and without intervention programs, and develop more probing questions that |

|  | go beyond questions related to specific actions. |
|  | 2.   Leverage FR/SO data to maximize value—need to understand the change in FR/SO levels as economic conditions and markets evolve so that programs meet today's needs. Monitor program application rates, process evaluations, product baselines, etc. Link FR/SO data with results from questions on demographics, attitudes toward environment and energy efficiency, reasons for program participation, shopping preferences and status of economy and conduct longitudinal studies to see how rates change over time and under what conditions. |
|  | 3.   Increase collaboration—need to look at attribution in both regional and national forums. By doing this on a group basis, could defray costs. |

Stoops, John, KEMA, Inc., The Cadmus Group, Inc., Itron, Inc., and Nexus Market Research, Inc. *Non-Residential New Construction (NRNC) Programs Impact Evaluation.* California Public Utilities Commission Energy Division. February 08, 2010.

| | |
|---|---|
| **Year(s) implemented** | 2006-2008 Program Years |
| **Sponsoring agency/utility** | CPUC |
| **Sector** | Commercial and industrial new construction (plus agriculture new construction for PG&E) |
| **Goal** | Savings by Design provides design assistance and financial incentives to improve the energy efficiency of commercial new construction |
| **Timing of Measurement** | Projects completed in 2006-2008 PY, but may have been started in earlier PYs. |
| **Eligible respondents** | Program participants – decision makers including building owners/managers, architects and engineers |
| **Type of measures** | System Shell, System Lighting, System HVAC + Motors, System Refrigeration, Whole building |
| **Free ridership questions for customers** | Most SBD have multiple measures, for which the levels of free-ridership may vary across the measures. Consequently, the survey questions were asked for every incented measure in the tracking database (systems approach) or identified in the project file (whole building approach). <br><br> The decision maker survey included the following questions as the NTG battery: <br><br> • On a scale from 0 to 10, where 0 means not influential whatsoever and 10 means extremely influential, How influential was Savings by Design, including the incentives, design assistance, design analysis and interactions with SBD representatives and consultants in the implementation of <measure description>? <br><br> • How did Savings by Design influence the implementation of <measure> (choose all that apply)? <br><br> • On a scale from 0 to 10, where 0 means that this measure would have been installed exactly the same regardless of interaction with Savings By Design regarding this project and 10 means that the measure would definitely not have been installed without SBD influence and interaction, what is the likelihood that this measure would not have been installed with SBD interaction? Why? |
| **Free ridership questions for vendors** | None |
| **Free ridership algorithm** | Since most projects involved multiple measures with potentially a range of free-ridership values, the estimation of the NTG ratios incorporated the responses for all measures. <br><br> The responses to the "influence of the program" and "absent the program" questions were used to generate measures' net-savings scores which were used to estimate measures' NTG ratio. <br><br> The "how influential" question response was multiplied by 0.1 to assign points for that response. The "how did SBD influence the implementation" question was assigned a score of 0 to 2 points, based on the response given. The "absent the program" question was assigned points by multiplying the answer by 0.3. <br><br> The cumulative score for each measure was compared to the max value of 6 to determine the degree of free-ridership. A score of 6 indicates that the measure was completely influenced by the program, and a score of zero indicates the measure would have been installed without the influence of the program. <br><br> The responses from the decision maker surveys were reviewed along with the |

| | |
|---|---|
| | program file to assess for consistency. |
| **Spillover questions for customers** | None |
| **Spillover questions for vendors** | None |
| **Spillover algorithm** | N/A |

Winch, Rick and Tom Talerico, Glacier Consulting Group, Bobbi Tannenbaum, KEMA Inc., Pam Rathbun, PA Consulting Group, and Ralph Prahl, Prahl & Associates. *Framework for Self-Report Net-to-Gross (Attribution) Questions.* July 2, 2008.

| Year(s) implemented | 2008 |
|---|---|
| Sponsoring agency/utility | Public Service Commission (PSC) of Wisconsin |
| Goal | Develop a framework to guide the revision of existing survey instruments used for determining attribution. |
| Key Issues Addressed | Revise existing surveys to improve consistency across Focus program areas and provide transparency for the approaches used. |
| Background | A working group of the FOE Evaluation team reviewed the existing self-report attribution batteries. The reviewers identified four areas of information that ideally would be collected as part of a self-report battery: <br><br>1. Context: recollection of past events, the sequence of these events and the how these events affected the participation process. <br>2. Decision-making: Having participants discuss their Focus project-related decision making; identification of factors that contributed to the process, and what decision-makers were involved. <br>3. Direct Attribution: Assess the impact of the program on the timing, efficiency level and quantity of technology installed. <br>4. Consistency Checks: Identified both during the interview and in the analysis stage, level of effort to resolve varies by importance of case. <br>   a. CATI survey implementation: changes to initial calculation of attribution are made by the analyst or project manager in the analysis phase <br>   b. In-depth project review by senior staff: Interviewer should be provided guidance on where to anticipate inconsistencies and relied upon to make a judgment of the degree of influence the program had on the customer responses. |
| Free ridership methodology | Specific questions were not developed, but types of information or issues to be considered are provided: <br><br>***Context*** <br>*Information classified as "Key Information"* <br><br>1. Confirm or determine whether the project involves new construction, building expansion, replacement of existing equipment, or modification to existing equipment. <br>2. Confirm type of equipment installed, date, reward amount, and other items deemed relevant. <br>3. Confirm evaluator's information regarding key services, rewards and assistance provided by Focus as well as the type and amount of vendor/implementer involvement. <br>4. Determine when and how respondent first heard about the services/rewards/assistance available through Focus. <br><br>• Explore possibility that new equipment was already installed before hearing about the services/rewards/assistance available from Focus. <br>• Explore any plan(s) to purchase or install equipment before learning about the services/rewards/assistance available through Focus. <br>   a. Understand existing plans. <br>   b. Understand point in planning process that respondent/organization (1) became aware of Focus and (2) begin discussing plans with Focus representative(s). <br>   c. Understand qualitatively the impact/changes necessitated by |

Focus involvement.

5. Discuss the working condition of replaced equipment (Probe: planned replacement/upgrade, failure, estimated remaining useful life, repair history)

*Information classified as "Supporting Information"*

1. Explore what first made respondent (organization) start thinking about installing/replacing equipment at (home/this facility).

2. Age of equipment that was replaced.

3. Explore previous Focus participation.

### Decision-Making

*Information classified as "Key Information"*

1. Organizational policies that specify factors considered when purchasing new (replacing old) equipment/ (Probe: payback, return on investment, guidelines on efficiency levels)

2. Major obstacles/barriers faced when seeking approval for project. (Probe: budget, time constraints, other priorities, disruption of production, etc.)

3. Role of contractor(s)/vendor(s) in project.

- Making respondent aware of Focus (or vice versa).

- Decision to participate.

- Recommendation to install certain type/energy efficiency level of equipment.

- Influence of contractor/vendor involvement on decision to install equipment at this time.(If not available from database)

4. Explore the percentage of the total costs—that is, all financial assistance plus the costs not covered by financial assistance—of installing improvements that were covered by Focus.

*Information classified as "Supporting Information"*

1. Budgeting process for new/replacement equipment. (Probe: size projects budgeted for, budget planning cycle/length)

2. Who within organization is responsible for recommending the purchase of new/replacement equipment.

3. Who within organization is responsible for approving the purchase of new/replacement equipment.

### Attribution (Timing, Efficiency, Quantity)
**Timing**
*Information classified as "Key Information":*

[Note: Remind respondent of ALL services/rewards/assistance from Focus—spanning from a facility assessment to educational materials to rebates/rewards].

1. Explore whether or not, in absence of any assistance from Focus, respondent would have replaced equipment (purchased new equipment) at the same time.

- If no, determine when it would have taken place. (If they cannot give number of years/months, then probe with mutually exclusive response categories.)

This is a critical point for a consistency check with responses in the context and decision making sections.

### Efficiency
*Information classified as "Key Information"*
(If operate multiple facilities)

1. Explore whether or not (before installing this equipment) the

organization had installed equipment of the same energy efficiency level at this or another facility without receiving services/rewards/assistance like those from Focus.

2. Explore whether or not respondent, when considering the purchase of this equipment, was aware of *a range of* efficiency levels that could have been chosen. Explore the range.

- Explore respondent understanding of efficiency levels available prior to Focus involvement.

- Explore when/how respondent first became aware of the energy-efficiency options available.

- Explore role of Focus (Focus representatives) in helping respondent understand the range of efficiencies available.

- If respondent has difficult time answering previous three questions, then: Explore whether or not all available efficiency options presented to respondent qualified for a reward/incentive through Focus (i.e., did respondent have both rebated and non-rebated option?)

- Without the services/reward/assistance from Focus, ask if respondent (organization) would have installed less efficient equipment. (Probe: If efficiency range available, ask more pointed question about what the efficiency level would have been).

*Information classified as "Supporting Information"*

1. Explore respondent's awareness of how the efficiency level of the old equipment compares to the efficiency level of the equipment that replaced it. (If new equipment must meet a minimum governmental standard, probe to see if respondent understands what that standard is.)

**Quantity**
*(ASK QUANTITY MODULE ONLY OF RELEVANT MEASURES WHERE THERE IS A VARYING LEVEL OF QUANTITIES—E.G., LIGHTING, MOTORS, VSD)*
*Information classified as "Key Information"*

1. Determine whether or not respondent (organization) would have installed the same quantity of equipment, fewer, or more at that time without the program.

2. If more or less, explore what percentage of (measure) would have been installed without the program assistance.

***Consistency Checks***
*(Questions that can be used to check consistency were included in the Context, Decision-Making and Direct Attribution sections. The questions below work best at the end of the attribution section of a survey.)*
*Information classified as "Key Information"*

1. Explore importance of services/rewards/assistance received from Focus on decision to install (measure)?

2. Ask respondent to describe, in their own words, what they (their organization) would have done if they had not participated in Focus (i.e., not received the Focus-related services/awards/assistance that they did).

| | |
|---|---|
| **Free ridership algorithm** | Responses to Direct Attribution are compared to Context and Decision-making responses to identify inconsistencies.<br><br>Level of effort for processing a single case during the analysis phase depends on the importance of the case to the overall attribution rate.<br><br>Analysis of self-reported battery varies by:<br><br>    1. Complexity of project |

|  | 2. Number of decision-makers |
|  | 3. Role of suppliers |
|  | 4. Evaluation budget |

For example, a program may have a large number of participants, relatively small or simple projects, single decision-makers and little involvement of suppliers. The analysis may be straight-forward, using direct attribution questions along with a few consistency checks. The data may be collected through CATI surveys.

Conversely, a program or project may have a small number of large, complex projects, multiple decision makers and multiple suppliers. The analysis would likely involve reviewing information from multiple sources to form an overall picture of the program or project. This information would be considered to make an informed decision as to the likely action (or lack thereof) that would have taken place absent the program.

# APPENDIX C: BIBLIOGRAPHY OF REFERENCES

Cook, Gay, Summit Blue Canada Inc. *Attribution Methodology Wars: Self-Report Methods versus Statistical Number Crunching—Which Should Win?* Paper delivered at ACEEE. 2008.

Cooney, Kevin, Beth Baker, Timea Zentai, and Adam Knickelbein, Summit Blue Consulting, LLC, *Gas Furnace Market Transformation Model Development and Market Research.* Submitted to Energy Trust of Oregon. August 5, 2009. Presented by Fred Gordon at AESP Brownbag. 2010.

Dohrmann, Donald, ADM Associates, Inc. John Peterson, Athens Research. Steve Westberg, Hiner and Partners. John Reed, Innovologie. *Evaluation Study of the 2004-05 Statewide Residential Appliance Recycling Program.* April 2008.

Efficiency Vermont. *Technical Reference User Manual (TRM), No. 2005-37, Measure Savings Algorithms and Cost Assumptions Through Portfolio 37.* Sent to Vermont Department of Public Service. November 29, 2005.

Erickson, Jeff and Mary Klos, Summit Blue Consulting and Valy Goepfrich, WPPI Energy. *Free-Ridership: Arbitrary Algorithms vs. Consistent Calculations.* IEPEC. 2009.

Fagan, Jennifer, Mike Messenger, and Mike Rufo, Itron, Inc. and Peter Lai, CPUC Energy Division. *A Meta-Analysis of Net-to-Gross Estimates in California.* AESP. 2009.

Goldberg, Mimi L., J. Ryan Barry, Erika Morgan, Ben Jones, Joshua Horton, and Nicole Buccitelli, KEMA, Inc. *Business Programs: Additional Looks at Attribution.* Prepared for the Public Service Commission of Wisconsin. February 26, 2010.

Goldberg, Mimi, KEMA Inc, Rick Winch and Tom Talerico, Glacier Consulting Group, Ralph Prahl, Prahl & Associates, and Bryan Ward, PA Consulting Group. *Treatment of Accelerated Savings.* Memorandum prepared for the Public Service Commission of Wisconsin. July 2, 2008.

Goldberg, Miriam L., J. Ryan Barry, Tammy Kuiken, Ben Jones, Paulo Tanimoto, Nicole Buccitelli, Colin Rickert, and Darcy DeAngelo-Woolsey; KEMA, Inc., *Business Programs: Acceleration Treatment and Life Cycle Net Savings.* Submitted to the Public Service Commission of Wisconsin. March 10, 2010.

Goldberg, Miriam, KEMA Inc., Oscar Bloch, Wisconsin Department of Administration, Ralph Prahl, Prahl & Associates, David Sumi and Bryan Ward, PA Consulting Group, and Rick Winch and Tom Talerico, Glacier Consulting Group. *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs.* Prepared for Public Service Commission of Wisconsin. March 16, 2006

Itron, Inc. and KEMA. *2004/2005 Statewide Express Efficiency and Upstream HVAC Program Impact Evaluation.* December 31, 2008.

Keating, Kenneth M., PhD. *Free-Ridership Borscht: Don't Salt the Soup.* IEPEC. 2009.

KEMA, Inc. *Evaluation of the 2004-2005 Statewide Multifamily Rebate Program–Volume I.* March 16, 2007.

Keneipp, Marshall, Floyd Keneipp, and Jeff Erickson, Summit Blue Consulting, LLC and Bill Norton, Opinion Dynamics Corp. *APS Measurement, Evaluation, & Research (MER) Report, Consumer Products Program (CPP).* APS. September 30, 2008.

Klos, Mary and Joan Huston, Summit Blue Consulting, LLC. *Impact Evaluation of 2007 CFL Buy-Down Pilot.* Prepared for Progress Energy—Carolinas. May 20, 2008

Lee, Allen, Cadmus, and KEMA, Inc. *Codes & Standards (C&S) Programs Impact Evaluation.* April 9, 2010.

Lori Megdal, Rick Ridge, Pam Rathbun, PA Consulting. Mimi Goldberg, KEMA. Ben Bronfman Quantec. Stuart Schare, Summit Blue; Nick Hall, Ken Keating. *Joint Simple Net of Free-Ridership and Participant Spillover SRA Battery.* April 19, 2008.

Megdal, Lori, Megdal & Associates, LLC, Yogesh Patil, Energy & Resource Solutions, Inc., Cherie Gregoire and Jennifer Meissner, New York State Energy Research and Development Authority, and Kathryn Parlin, West Hill Energy & Computing, Inc. *Feasting at the Ultimate Enhanced Free-Ridership Salad Bar.* IEPEC. 2009

National Action Plan for Energy Efficiency (2007). *Model Energy Efficiency Program Impact Evaluation Guide.* Prepared by Diane Munns and Jim Rogers. <www.epa.gov/eeactionplan>

Nonresidential Net-To-Gross Ratio Working Group. *Methodological Framework for Using the Self-Report Approach to Estimating Net-to-Gross Ratios for Nonresidential Customers.* May 8, 2009.

Peters, Jane S. and Marjorie McRae, Research Into Action, Inc. *Free-Ridership Measurement Is Out of Sync with Program Logic… or, We've Got the Structure Built, but What's Its Foundation?* ACEEE. 2008.

Peters, Jane S., Ph.D. and Ryan E. Bliss, Research Into Action. *Fast Feedback Pilot: Existing Buildings and Production Efficiency Programs.* Prepared for Energy Trust of Oregon. March 10, 2010.

Prahl, Ralph, Prahl & Associates, Goldberg, Miriam and Bobbi Tannenbaum, KEMA Inc, David Sumi and Bryan Ward, PA Consulting Group, and Tom Talerico and Rick Winch, Glacier Consulting Group. *Integrating Supply-Side Results with End-User Net-to-Gross Self Reports.* Memorandum prepared for the Public Service Commission of Wisconsin. July 2, 2008.

Rathbun, Pam, Carol Sabo, and Bryan Zent. PA Consulting Group. *Standardized Methods for Free-Ridership and Spillover Evaluation—Task 5 Final Report (Revised).* Prepared for the Massachusetts Utilities, June 13, 2003.

Residential/Small Commercial Joint Simple Net-to-Gross (Self-Report) Committee, *Response to Overarching Comments Regarding the Use of Self-Reported Net-to-Gross (NTG) and the Residential and Small Commercial Self-Report Approach NTG Method*, January 28, 2010.

Ridge, Richard, Ken Keating, Lori Megdal, and Nick Hall. *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches.* October 15, 2007.

Ridge, Richard, Ridge & Associates, Philippus Willems, PWP Inc, and Jennifer Fagan, Itron, Inc. *Self-Report Methods for Estimating Net-to-Gross Ratios in California: Honest!.* AESP. 2009.

Ridge, Richard, Ridge & Associates, Phillipus Willems, PWP Inc, Jennifer Fagan, Itron, Inc., and Katherine Randazzo, KVD Research Consulting. *The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio.* IEPEC. 2009.

Saxonis, William P. New York State Department of Public Service. *Free-Ridership and Spillover: A Regulatory Dilemma.* IEPEC. 2007.

Skumaz, Lisa, SERA, Sami Khawaja and Jane Colby, Cadmus Group. *Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence of Energy Efficiency Behavior.* November 2009.

Stoops, John, KEMA, Inc., The Cadmus Group, Inc., Itron, Inc., and Nexus Market Research, Inc. *Non-Residential New Construction (NRNC) Programs Impact Evaluation.* California Public Utilities Commission Energy Division. February 08, 2010.

TecMarket Works, et al. *The California Evaluation Framework.* June 2004.

Titus, Elizabeth and Michals, Julie, Northeast Energy Efficiency Partnerships. *Debating Net Versus Gross Impacts in the Northeast: Policy and Program Perspectives.* AESP. 2008.

Winch, Rick and Tom Talerico, Glacier Consulting Group, Bobbi Tannenbaum, KEMA Inc., Pam Rathbun, PA Consulting Group, and Ralph Prahl, Prahl & Associates. *Framework for Self-Report Net-to-Gross (Attribution) Questions.* July 2, 2008.

Wright, Roger and Stoops, John, RLW Analytics. *An Evaluation of the 2004-2005 Savings By Design Program.* October 2008.