Attribution Methodology Wars: Self-Report Methods versus Statistical Number Crunching—Which Should Win?

Gay Cook, Summit Blue Canada Inc.

ABSTRACT

Several methods have been used to assess the impact of free riders and spillover on both residential and nonresidential energy efficiency programs. Those methods include self-reported program response, econometric methods such as billing analysis, econometric or discrete choice models, and market sales based methods such as saturation data analysis. Each approach has benefits and drawbacks, each is appropriate in different circumstances, and each has its own advocates. However, it is often hard to know which approach is best for a given program. How are program evaluators to know which approach they should use in what circumstance? What are policy makers to believe when told that "this method is the best?" How can these approaches be placed in context and be assessed objectively?

This paper explores these fundamental questions and provides the background to place the approaches in a broader perspective. It describes the various approaches and presents their pros and cons, and also provides some practical examples of how to determine which approach is most suited to specific circumstances. The paper is based on free ridership and spillover research done on a variety of programs in the United States and Canada and a thorough literature review.

Methods to Assess Free Riders and Spillovers

Net impacts from energy efficiency programs are primarily determined by adjusting claimed savings for those savings that would have occurred in the absence of the program (free riders) and those savings that resulted from the program but were not counted in claimed results (spillover). Many jurisdictions assess the annual impact of free riders and spillover applying methods which range from assuming a fixed (or deemed) net-to-gross (NTG) ratio such as in Iowa which assumes a NTGR of 1.0 to triangulation of several methods such as for the enhanced level of rigor required in the California evaluation protocols (Quantec 2002).

Evaluators use several methods to assess the NTG ratio; it may not always be clear what would be the most appropriate method to apply for a given program. This section compares and critiques the key methods to assess net program impacts—self-report surveys and econometric methods, and discusses market share approaches. The most common methods used are described briefly below and in more detail in the rest of the section.

- *Self-report methods* rely on responses to a battery of survey questions asking participants and nonparticipants what they would have done in the absence of the program support. Enhanced self-report methods include surveys of vendors and trade allies, and reviews of program documentation, market data, etc.
- *Econometric methods* apply statistical tools and techniques to economic issues and data. Billing analysis determines the net impact by analysis of billing data from participants and nonparticipants. Econometric models compare participants' and non-participants' energy and demand patterns, adjusting for external variables that could account for

changes in use and patterns. Discrete choice analysis uses data on equipment or practice choices by participating and nonparticipating customers together with other information about customers to model choices participants would have made in the absence of the program.[1]

- *Market share methods* include comparing aggregated sales volumes of a particular technology in a specific location with a baseline estimate of the volume that would have been sold in the absence of the program (market sales approach), and using observations at two points in time of the share of existing equipment stock that is high efficiency (saturation data analysis).

Self-Report Surveys

Generally, the simplest and lowest cost NTG method is using the survey-based stated intentions method with a telephone survey for data gathering. Although research has shown that this method can provide biased results, coming at the question of what the participant would have done in the absence of the program from a variety of different perspectives (directly asking, decision-making criteria, where they were in the process, etc.) and assessing these together is one way the survey methods have used to triangulate on the correct construct.

Much work has been done to develop, implement, and refine and standardize self-report methodology to assess free riders and spillover in energy efficiency programs. In 2002, Massachusetts regulators asked for a study to create a standardized free ridership survey method to be used by all Massachusetts utilities for program evaluations (PA Consulting, 2003)**.** The objective was to develop standardized sampling techniques, data collection approaches, survey questions, survey instrument(s), and an analysis methodology that each of several Sponsors[2] can use to determine free-ridership and spillover factors for C&I programs. This standardization project was designed to provide a methodology to meet the regulatory requirements to report annual program impacts (along with disaggregated free-ridership and spillover values). In 2006-2007 the CPUC developed both an Evaluation Framework (TechMarket Works, 2004) and a set of protocols (TecMarket Works, 2006). The Evaluation Framework documents note that NTG can be expected to vary depending upon maturity of equipment or service, type of delivery in the program, maturity of the program, and customer sector. The California guidelines for self-report method specify the following steps: identify the correct respondent; use multiple questions; assess validity and reliability of each question; include consistency checks; make the questions measure-specific; include and document partial free-ridership; assess deferred free-ridership; develop scoring algorithms; explain handling of non-responses and "don't knows"; weight the NTG for size of impacts; report precision of the estimated NTG; pre-test the questionnaire; use multiple respondents and consider third-party influence. A set of standardized questions for both free riders and spillover has been developed for use in evaluating the current programs in California.

Schare and Ellefsen (2007) discuss the approach used to estimate free ridership for several New York State Energy Research and Development Authority (NYSERDA) programs; a

[1] Delphi methods which collect judgmental estimates from a panel of experts and develop a consensus or central range estimate are typically used only if more objective methods are not available.

[2] The Sponsors are National Grid (Massachusetts Electric, Nantucket Electric), NSTAR Electric, Northeast Utilities (Western Massachusetts Electric), Unitil (Fitchburg Gas & Electric Company), Cape Light Compact).

method that had evolved from previous NYSERDA evaluations and work done in California and New England. This method was extended to the impact evaluation of the Shared Savings Program in 2007. Estimates of savings attributable to the program were based on responses of participants to carefully derived questions (building on established work within the field) regarding prior intentions, the importance of factors such as financial incentives, and the likelihood the same actions would have been taken without the program. Free ridership was asked in both direct questions—aimed at obtaining respondent estimates of the appropriate (partial) free ridership rate that should be applied to them—and in supporting, or influencing, questions that could be used to verify whether direct responses are consistent with participants' views of the program's influence. Each response was assessed in a systematic manner to produce estimates of free ridership and spillover.

The method overcomes a key limitation of self-report—the difficulty of systematically converting opinions of participating customers into quantifiable free ridership values. The approach is based on participant self-reports and offers unique benefits of a clearly defined and repeatable method to quantify free ridership, while also incorporating qualitative information from program participants often used only as supporting illustration. The approach includes:

- Direct estimation of free ridership from the perspective that is most appropriate for the project and to which the respondent can best relate his program experience. This takes the form of either the likelihood at the high-efficiency measures would have been installed without the program, or the share of high-efficiency measures that would have been installed without the program.
- Separate estimation of free ridership addressing the complete project across all measure types and, alternatively, addressing decisions to install specific measures. The dual line of questioning allows respondents to provide a big-picture view of the program's influence on the project as well as to focus on specific measures, which may have been influenced by the program to varying degrees.
- Quantitative incorporation of qualitative responses based on interviewers' probing for details and causality. This aspect of the approach relies on experienced interviewers who are able to apply appropriate judgment to assign influence scores reflecting the degree to which the program affected equipment-purchasing decisions.

To improve the quality of NTG ratios drawn from self-reported survey responses, the evaluation can rely on multiple data sources for the decision to install or adopt energy efficiency measures or practices; this is referred to as *enhanced self-report* methods. Enhanced self report methods involve calibrating other sources of information, such as interviews with trade allies, review of program documentation, or analysis of market-based sales data, with the survey results. Self-report methods are primarily used to determine if participating end users would have implemented program measures without the program. Enhanced self-reports methods can also determine what additional efficiency improvements participating customers have made outside the program, how participating vendor sales practices would have been different without the program, and how nonparticipating vendor and customer practices have changed since the advent of the program.

California's new evaluation protocols for NTG impact evaluation rely heavily on self-report methods but require triangulation of methods for the enhanced level of rigor. California has three levels of rigor that can be applied to NTG analysis—basic, standard, and enhanced.

Participant self-report through surveys is the required method for the basic level of rigor; for the standard level of rigor, one of three methods can be used (billing analysis, self-report, econometric/discrete choice). The enhanced level requires triangulation using more than one of the methods in the standard rigor level. This must include analysis and justification for the method for deriving the triangulation estimate from the various methodologies used.

Econometric Methods

These methods consist of statistical models that compare participants' and non-participants' energy and demand patterns, their knowledge about efficiency options, and/or the trade-offs they are willing to make between efficiency options and the costs of purchasing and installing them. They include billing analysis, econometric models, and discrete choice models and often include survey inputs as well as other non-program-related factors such as weather and rates.

Billing analysis uses multivariate regression models with historical utility billing data (kW and kWh) to calculate annual demand and energy savings. In general, billing analysis is used with complex equipment retrofits and controls projects and provides retrofit performance verification for projects where whole-facility baseline and post-installation data are available. The net billing model specification incorporates both participants and nonparticipants into one model, and the resulting sample is not randomly determined. In particular, participants self-select into the program and therefore are unlikely to be randomly distributed; the unobserved characteristics that influence the decision to participate must be accounted for in the model to avoid producing biased coefficient estimates. The Inverse Mills method which includes a ratio in the model to account for self-selection was developed to correct for this bias but has several limitations: large customers can exert such a significant influence that they overly bias results; the usable sample is reduced by the need for good historical billing data for each customer; and the method does not estimate spillover, rendering it an incomplete model of net impact.

Econometric models are used to analyze co-relational relationships, usually with the hope of determining causation. They are used to estimate macroeconomic trends and in microeconomics to estimate virtually any sort of social relationship (much as metric models, involving these same regression techniques, are used in other social sciences). The use of statistical/econometric models to estimate net impacts can avoid both the concern over the potential for bias and cognitive dissonance issues with survey research by analyzing participant and non-participant actions, characteristics and attitudes to predict free ridership and spillover. The disadvantage of this method is its inability to estimate spillover upstream in the distribution channel. A robust statistical analysis includes surveys designed to minimize self-reporting bias while collecting data on other program and participant characteristics. This level of sophistication requires a relatively large expenditure on evaluation, which can impact the cost-effectiveness of a marginal program.

Another method is the two-stage discrete choice model which simulates the decision to purchase various types of commercial equipment. Once estimated, the model is used to determine the probability of purchasing high-efficiency equipment in the absence of the program. The probability of purchasing any given equipment option A is expressed as the product of two probabilities—the probability that a purchase is made multiplied by the probability that equipment option A is chosen given that a purchase has been made.

Market Share

Market share methods (saturation data analysis, market sales) can also be used to estimate free riders and spillover. For saturation data analysis, translating successive observations into incremental attributable sales requires information (estimates or assumptions) about equipment turn-over rates, stocking practices, and changes that would have occurred over the time period without the program. Collecting reliable saturation data is typically expensive and not repeated frequently. The market sales approach is generally used to assess transformations of markets and depends on completeness and accuracy of sales data and the validity of the baseline estimate. It is often difficult to collect sales data due to vendor concerns about releasing competitive data and comparison of data by region, e.g. manufacturers may track by sales region (east/west) rather than by state or province, much less by utility franchise area.

A study done for Wisconsin Focus on Energy (KEMA, 2006) describes a way to apply screening criteria to determine whether to use market sales or self-report methods to assess NTG results. Taken together, such factors can indicate an overall preference for one method or another. In some cases, the preference will be clear-cut. In others, the two methods may be nearly equally good—or nearly equally poor. Following are the criteria to assess whether market sales can substitute for self-report methods. For the first two criteria, the quality of available data depends in part on the details involved in data collection which in turn depends on resources available.

Table 1. Screening Criteria for Self Report versus Market Share NTG Approaches

Screening Criteria	Example Screening Questions
Sales Data Availability: The availability of current and baseline market sales data enables estimating free ridership based on such data.	Are current and baseline data readily available? Are the data comprehensive and complete? Able to supplement/overcome shortcomings in data with other data collection techniques? Is the baseline estimate reliable?
Accuracy of Self-Reports: The ability of end users and vendors to report accurately what would have occurred in the absence of the program enables the use of program-response self-report methods.	Can end users/vendors accurately report what would have occurred without program? Supply-side actors can comment on programmatic versus non-programmatic influence on market? Has program altered the supply side in ways a participant would not be able to recognize?
Likelihood of Large Nonparticipant Market Effects: The likelihood of substantial nonparticipant market effects may indicate a need for applying methods for adequately capturing such effects.	Is the scale of program large relative to overall market? Are primary sales driving components (promotions, incentives) available at a consistent level throughout the year? Does the program have broad reach across market niches? Does program theory predict significant non-participant effects?
Narrowness of Technology Definition: A market data approach is suggested if the technology is a single type and well-defined, versus encompassing multiple categories, types, or wide variations.	Does program offer "custom" solutions (broad definition) or "prescriptive" measures (narrow definition)? Does program target specific technologies (narrow definition) or a broad range of technologies (broad definition)?
Uniformity of Unit Savings: The choice of method is guided by whether savings per unit is sufficiently consistent across types of units & customers to adequately quantify in terms of total units sold, or needs information on unit characteristics by customer type.	Do units promoted through the program come in widely varying size ranges/savings levels? Is an engineering estimate of necessary? Large variation in customer application of measures? Do savings per unit vary by customer application? Expect savings to vary widely by customer?
Source: Goldberg M.L. et al *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs*, March 2006.	

Choosing Appropriate NTG Methodology

The survey approach is the most straightforward way to estimate free ridership and spillover and is usually the lowest cost approach. As noted by the National Action Plan Guidelines (NAP, 2007)…*"survey methods can be used with any program regardless of the number of participants"* whereas econometric methods "c*an only be used with programs with large numbers of participants because the models need large amounts of data to provide reliable results".* In California, econometric methods are preferred in situations with enough participants and comparable non-participants, and when the program is large enough to justify the expense. However, programs with either a very small number of participants or nonparticipants or where comparability is a severe problem (such as industrial plants with unique facilities) are not amenable to these methods and need to rely on a survey-based method. Market share methods are generally used to assess market transformation programs or in situations where participation is not well defined. Table 2 has an overview of the pros and cons of the methods discussed in the previous sections.

Table 2. Comparison of Free Rider and Spillover Methodologies

Approach	Pros	Cons
Self-Report	Simpler and less expensive than all other approaches. Can use all data points and can be used in a variety of situations. Directly addresses the behaviors the program is seeking to affect. Flexible enough to take into account the complexities of program-participant interaction. Enhanced self-report increases accuracy by triangulating estimates from other sources like vendor interviews.	Potential for non-response bias, limited respondent recall of program influence on decision-making, and potential investigator bias in translating responses into free ridership values. Tends to underestimate spillover. It is virtually impossible to define a precision target and a statistically valid sample size. This challenge in surveying comes from the nature of collecting both qualitative and quantitative data from various participants and non-participants involved in the decision to install energy efficiency measures.
Econometric	Analysis provides quantitative estimates of magnitude of net impacts from statistically valid methods based on historical billing data. Modeling can provide more accuracy because tests for bias and precision can be included.	Includes participants and non-participants in one model; sample not randomly determined due to self-selection. Could produce biased coefficient estimates if unobserved characteristics which influence decision to participate not accounted for. Large customers can overly bias results. Econometric models need good historical data for each customer and this can reduce number of data points. For discrete choice models it is difficult and costly to get accurate data on types and efficiency levels of existing equipment. Neither method includes trade ally effects.
Market Share	Addresses trends in the entire market for equipment. Can estimate net energy impacts for program where participation is not well defined.	Even if discrete pieces of equipment can be identified, obtaining relevant and adequate market sales information can be very difficult as well as costly.

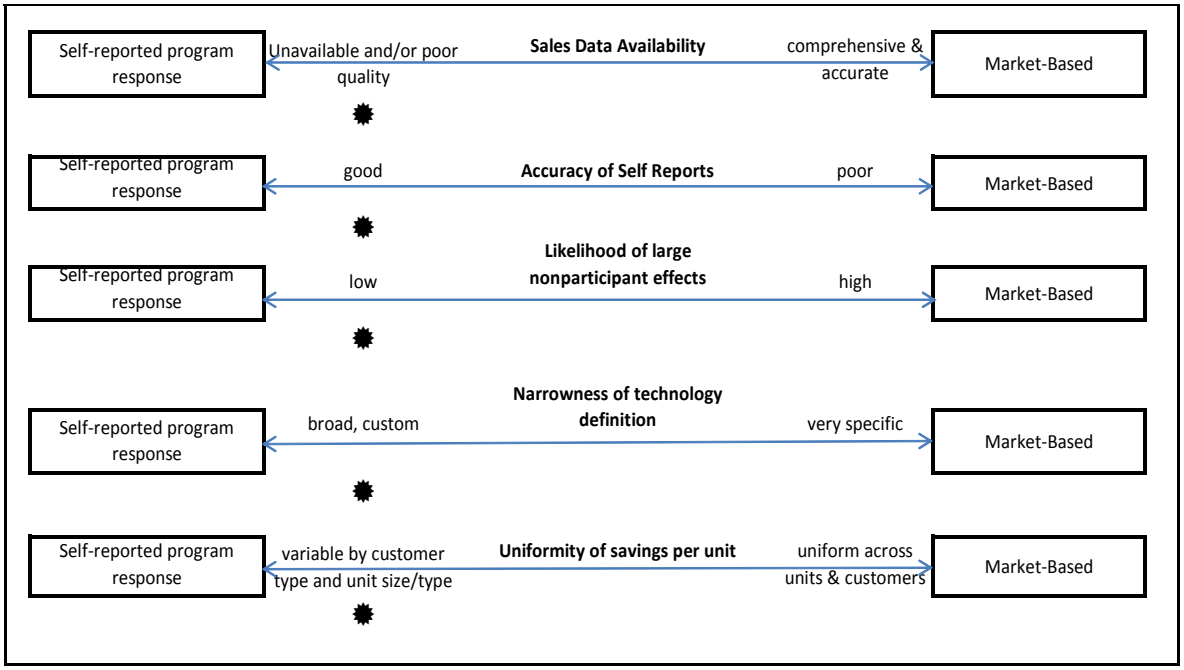
The selection of the best approach to evaluating NTG ratios depends on the objectives of the program being evaluated, evaluation budget and resources, and specific aspects of measure and program participants. For some programs, such those involving custom projects for commercial and industrial (C&I) customers, it can be quite clear what methodology to use. Custom C&I projects programs are targeted specifically at large commercial and industrial customers and target complex and unique systems rather than offering prescriptive rebates. In addition, it is possible that in some segments such as agriculture, most eligible customers participate, making the selection of a non-participant group problematic. The self-report method is more appropriate for this research than econometric methods as shown in Table 3 which compares the two methods based on relevant program characteristics.

Table 3. Compare Self-Report to Statistical Models for C&I Custom Programs

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.

And applying the Wisconsin Focus on Energy method selection criteria to the custom projects program, as shown in Figure 1, clearly indicates that the self-report method is preferred over the market share approach. For Custom C&I programs the enhanced self-report is the best approach to determine free riders and spillover. This would involve conducting customer surveys, trade ally and vendor surveys, assessing information from project documentation, holding discussions with account representatives, and including other industry information such as standard payback requirements for a specific segment, etc.

Figure 1. Market Sales or Self Report Criteria Applied to Custom C&I Programs

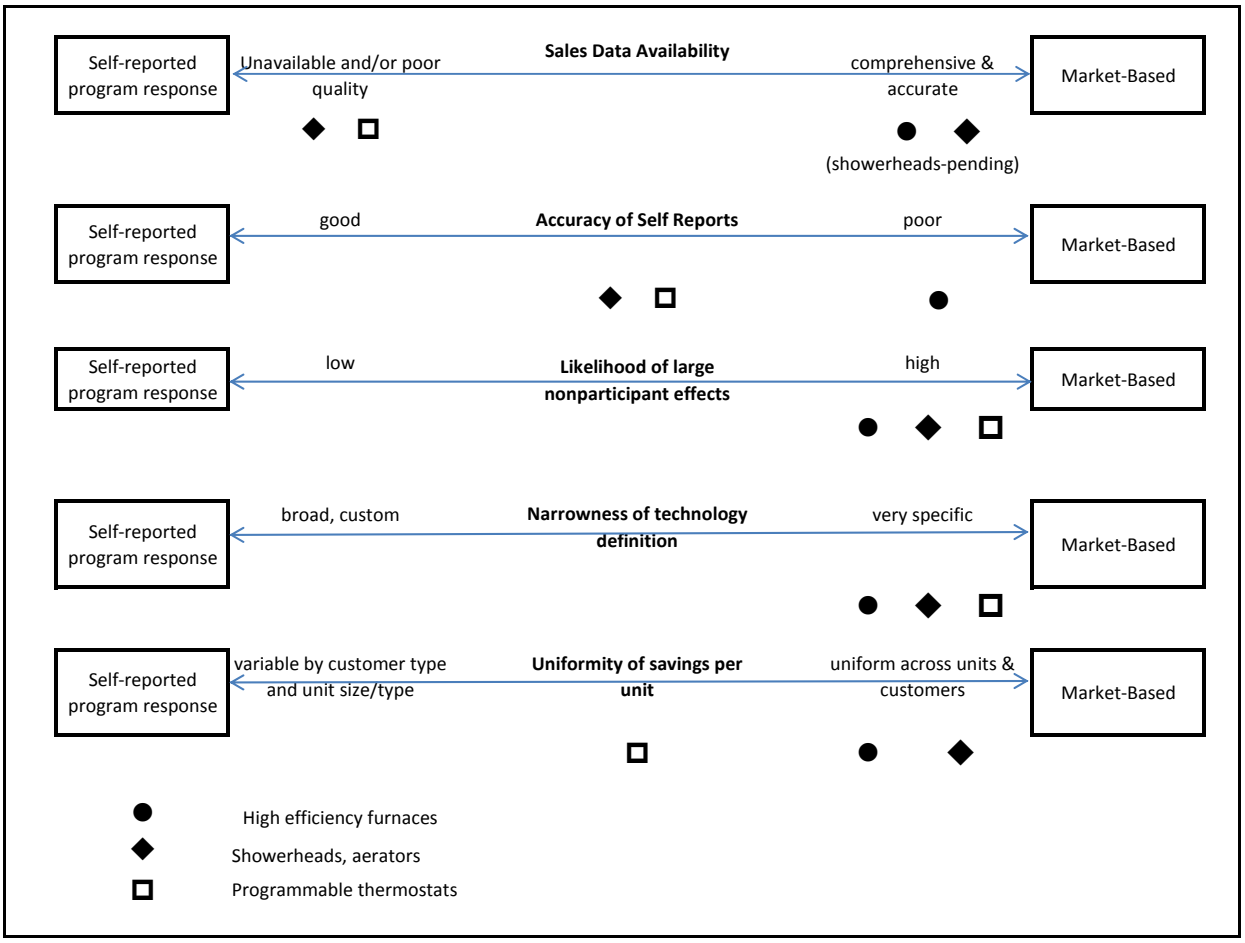


Source: Goldberg M.L. et al *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs*, March 2006 and Summit Blue analysis.

In other situations, it is not as simple. For example Figure 2 below shows how some selected residential gas measures (high-efficiency furnaces, low-flow showerheads and faucet aerators, programmable thermostats) were scored with the screening criteria in terms of self-

report compared to market sales data. The assessment suggests market data are preferable to self-report methods; however, in this example market data were not available for three of the four target measures. Non-participant data was not readily available to use in regression models and there were also cost considerations. To determine NTG ratios, self-report surveys with participants were conducted for all measures and for high-efficiency furnaces, these surveys were enhanced by interviews with furnace vendors and market sales data on natural gas sales.

Figure 2. Applying NTG Screening Criteria to Selected Residential Measures



Source: Goldberg M.L. et al *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs*, March 2006 and Summit Blue analysis.

Which Should Win – Self-Report or Statistical Methods?

What are best—self-report or statistical methods or should market share methods be applied? Market share methods are applicable when there is not good information on participants and/or the goal is to assess transformation of a market. Surveys with participants and non-participants should be done whether the intention is to estimate NTG with self-report methods or with statistical ones. The best approach is to use the guidelines for self-report surveys, e.g. frameworks, questionnaires, etc. that have been developed by the evaluation industry, to enhance these self-report surveys with interviews with other market actors, market share data, industry

trends, etc., to use the survey results in statistical modeling where appropriate and to select the approach that best suits the available resources and meets the program evaluation goal.

In addition, it is not unusual for combinations of approaches to be used to determine NTG ratios. For example, rigorous econometric methods may be used every three years with self-reported or deemed NTGRs used for the other program years. Ultimately it comes down to the judgment of a qualified evaluator. The theory behind attribution analysis is that only impacts caused by the program should be included in net savings estimates; however, absolute proof of causality is unattainable since one can never observe what would have happened in the absence of the program. Consequently, causality "must be justified or rationalized on the basis of *a priori* argument, outside evidence, intuition, theory, or some other informal means."(Moffitt 2003). The necessity of this approach to attribution analysis, relying in part on intuition and outside assumptions, is supported by Heckman (2000) in his argument that "there is no mechanical algorithm for producing a set of 'assumption free' facts or causal estimates based on those facts." In the context of energy efficiency program evaluation, Wirtshafter and Sorrentino (1994) add that "any proof [of program influence] available will be, at best, circumstantial."

References

California Public Utility Commission NTG Working Group. 2008, *Proposed Net-to-Gross Ratio Estimation Methods for Nonresidential Customers* (Draft).

California Public Utility Commission NTG Working Meeting. 2007, *Overview Summary of Direction and Agreement*.

Goldberg M.L., Bloch, O., Prahl, R., Sumi, D., Ward, B., Winch, R. & Talerico, T. 2006. *Net-to-Gross Method Selection Framework for Evaluating Focus on Energy Programs*.

Kandel, A. 2002. *Theory-Based Estimation of Energy Savings from DSM, Spillover, and Market Transformation Programs Using Survey and Billing Data.* Program Measurement & Evaluation.

KEMA. 2005. *Final Report: Phase 2 Evaluation of the Efficiency Vermont Residential Programs*, prepared for the Vermont Department of Public Service.

Itron. 2006. *2003-2004 Home Energy Savings Program Residential Impact Evaluation,* prepared for the Energy Trust of Oregon.

National Action Plan for Energy Efficiency. 2007. *Model Energy Efficiency Program Impact Evaluation Guide*.

PA Consulting Group Inc. 2003. *Standardized Methods for Free Ridership and Spillover Evaluation – Task 5 Final Report (Revised)*.

Quantec. 2002. *Assessment of Energy and Capacity Savings Potential in Iowa Volume 2: Free Riders and Spillover – A Look Back, A Path Forward,* prepared for the Iowa Utility Association.

RLW Analytics and KEMA, Inc. 2006. *Final Report: Phase 2 Evaluation of the Efficiency Vermont Business Programs*.

Schare, Stuart and Ellefsen, Jennifer. 2007. *Advancing the "Science" of Free Ridership Estimation: An Evolution of the Self-Report Method for New York Energy $mart℠ Programs.*

Stone, Arthur A., Jaylan S. Turkkan, Christine A. Bachrach, Jared B. Jobe, Howard S. Kurtzman, and Virginia S. Cain. 2000. *The Science of the Self-Report: Implications for Research and Practice*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Summit Blue Consulting. 2007. *New York Energy $mart Program Evaluation and Status Report for the Year Ending December 31, 2006*, New York State Energy Research and Development Authority.

Summit Blue Consulting. 2006. *Shared Savings Decision-Making Process Evaluation Research Results,* prepared for Wisconsin Power & Light.

Summit Blue Consulting & Quantec, *Commercial/Industrial Performance Program Market Characterization, Market Assessment and Causality Evaluation*, prepared for New York State Energy Research and Development Authority, April 2006.

TecMarket Works, *California Energy Efficiency Evaluation Protocols: Technical, Methodological and Reporting Requirements for Evaluation Professionals*, April 2006.

TecMarket Works, *The California Evaluation Framework,* Southern California Edison, 2004.

Torok, C., Cavalli, J. and O'Drain, M. *Any Way You Slice It: Issues of Behavior and Influence in Net Impact Analysis,* 1999.