Techniques for getting the most from an evaluation: Review of methods and results for attributing progress, non-energy benefits, net to gross, and cost-benefit

Lisa A. Skumatz, Ph.D.
Skumatz Economic Research Associates, Inc.
Superior, CO

Keywords
attribution, net to gross, evaluation, cost-benefit, evaluation, non-energy benefits, program theory, best practices

Abstract

As background for several evaluation and attribution projects, the authors conducted research on best practices in a few key areas of evaluation. We focused on techniques used in measuring market progress, enhanced techniques in attributing net energy impacts, and examining omitted program effects, particularly net non-energy benefits. The research involved a detailed literature review, interviews with program managers and evaluators across the US, and refinements of techniques used by the authors in conducting evaluation work. The object of the research was to uncover successful (and unsuccessful) approaches being used for key aspects of evaluation work. The research uncovered areas of tracking that are becoming more commonly used by agencies to assess progress in the market.

In addition, detailed research by the authors on a number of impact and attribution evaluations have also led to recommendations on key practices that we believe comprise elements of best practices for assessments of attributable program effects. Specifically, we have identified a number of useful steps to improve the attribution of impacts to program interventions. Information on techniques for both attribution / causality work for a number of programs are presented – including market transformation programs that rely on marketing, advertising, training, and mid-stream incentives and work primarily with a network of participating mid-market actors. The project methods and results are presented and include:

- Theory-based evaluation, indicators, and hypothesis testing;
- Enhanced measurement of free riders, spillover, and other effects, and attribution of impacts using distribution and ranges of measure and intervention impacts, rather than less reliable point estimates;
- Attribution of program-induced non-energy benefits;
- Net to gross, benefit cost analysis, and incorporation of scenario / risk analysis of results;
- Comparison of net to gross results across program types to explore patterns and important differences.

These extra steps improved the reliability and robustness of the results of the causality analysis and provided a better foundation to guide benefit-cost analysis and program and investment decisions – an important goal of an evaluation. The paper highlights benefits and impacts of these approaches, and provides comparisons, contrasts, lessons learned, and highlights successful approaches that may be transferable to other locations.

Finally, the authors suggest that there are important impacts from programs that are often omitted from evaluations – non-energy benefits. Suggested approaches for assessing and measuring these hard-to-measure impacts are presented, along with order of magnitude results that have been estimated in previous research. The research suggests that cost-benefit analyses incorporating scenario analyses related to credible ranges for NTG estimates, and incorporating subsets of NEBs may provide the best information on program impacts. Combined with good quality information, tracking market progress provides a strong basis for evaluat-

ing program effects and understanding the progress the program has made in transforming the market.

Context for Evaluation: Attribution and Net To Gross

Evaluating the performance and outcomes of energy efficiency programs serves several important purposes, including assuring that public funds are being spent appropriately, providing information to help design and refine programs, and helping assess investments in programs to a degree that is similar to guiding other uses to which the funds might be put – e.g. investment in generation or other applications. While very large amounts of money may be invested in evaluation, a guiding principle should be that the level of accuracy needed for the evaluation should be that which is needed to avoid making a wrong decision for the investment of public funds – potentially no more, and no less. The investment in evaluation should be related to the cost of a wrong program investment. However, the evaluation and analysis should explicitly recognize aspects of uncertainty, risk, and scenario analysis and other criteria that are commonly applied to analysis of other investments. These last analyses are not commonly performed in association with energy conservation as they are with supply facilities.

The evaluation seeks to identify and quantify those impacts that occur because of the program – specifically excluding those effects that would have occurred naturally, in the absence of the program. This comparison becomes complicated because the status quo changes; the program occurs over time in a constantly changing (energy) environment. The issue of a dynamic baseline makes comparison and attribution of program impacts a considerably more complicated endeavor than comparisons assuming simply that all changes after implementation of the program are due to the program. The issue of a changing baseline, along with myriad other complexities associated with attribution, are among the reasons that economists[1] indicate that causality (at least in the real world)[2] cannot generally be proven – however, it can be indicated through a preponderance of reasonable or hopefully overwhelming and consistent evidence, and that is the goal of the evaluator.

It has become standard practice to develop some degree of program theory and logic for efficiency programs –specifically assessing the market, actors, and interactions and identifying specific barriers and conduits to address barriers to adoption of energy efficient equipment. Theory-based evaluation is a useful enhancement to the historical evaluation approach. Furthermore, it can help reduce the evaluation problem. Theory-based evaluation can potentially break the large evaluation and attribution problem into a number of smaller – more measurable – problems. Program theory and logic assess the market and context in which a program operates and attempt to identify the actors, supply chains, actions, links, and supply chains that decompose the program into a number of steps. For example, measuring the effect of an increase in awareness among participating distributors of a particular energy efficiency measure can be a considerably easier task than to design a survey around the overall end-goal of "moving the market forward", increasing sales or other end indicators. Furthermore, these interim effects allow the program evaluators to identify the point at which the program theory or logic broke down, or the point at which market progress ceased – allowing planners to address the logic, modify the incentives, or address the issue in other ways.

Assessing Market Progress and Tracking

As a precursor to evaluation efforts for several evaluation and attribution projects, the authors conducted background research on tracking, evaluation, and net to gross methods being used elsewhere. The tracking issue is complicated, and data can be difficult and expensive to obtain. The object of the research was to uncover successful (and unsuccessful) approaches being used for market share tracking, program indicators, and evaluation methods, identify lessons learned, and assemble comparables for the results of attribution and net-to-gross analyses to avoid pursuing tracking methods that had already failed. We queried more than 60 professionals involved in tracking and evaluating market transformation and resource acquisition programs, and asked detailed questions about approaches to tracking aspects of market progress, market causality / attribution methods, and net to gross analysis.

The interviewees included utility staff, program managers, consultants, regulators, interveners, and others with evaluation experience and experience trying to track market shares and market progress for a diverse set of programs and sectors. We identified positive experience with some market share tracking efforts, and significant caveats associated with others. In addition, we found distinctly different approaches and philosophies used in different parts of the country, especially in the development of indicators. The survey of attribution studies identified limited patterns in net-to-gross (NTG) results by program type, measures, and sectors – and provided ranges for comparison.

The interviews focused on gathering feedback on both: 1) indicators of market progress (including awareness, knowledge, stocking practices, incremental price), and 2) market share / sales tracking efforts. We asked about:

- tracking efforts and market progress indicators in the areas of: awareness, market share, stocking behaviors, price increments, and impacts;
- efforts for a variety of sectors, including: residential, commercial, low income, new construction, renewables, and outreach programs;
- successful attribution and causality approaches and results.

We found significant differences in the design of indicators used to evaluate net program progress. While some agencies work to assemble exhaustive lists of indicators reflecting market factors to signal changes in awareness, knowledge,

1. Granger, among others.
2. Author's caveat

Table 1. Key Tracking Topics / Indicators of Energy Efficiency Market Progress Used by Agencies in the US.

Topic	Tracking methods / indicators
Awareness / knowledge	Awareness and knowledge of program, features, efficiency, logos, and other specific program aspects / interventions among market actors gathered through surveys (Changes in) attitudes about programs, features, efficiency; perceptions of future trends
Product service / availability and practices	Actors undertaking program activities / enrollment, etc. Specifier practices (builders, contractors, A&E, owners, etc.) Stocking and sales practices (retailers, contractors, etc.); product availability; perception of the role of efficient equipment in business success
Market share	Sales or market share of relevant equipment tracked through shipments or sales garnered from third parties (AHAM, D&R, etc.) Tracking through (labour-intensive) calls / relationships with distributors (like California); In-store model counts / comparisons by mystery shoppers or circuit riders; Sales reporting that may be required of participating retailers; and Other methods.
Energy savings	Gross savings measured through program records Net savings through analysis of inspections, realization rates, and Net to Gross analysis of free riders and spillover (discussed below)
Incremental cost	Indicators of incremental cost, competitiveness, price premiums, often from surveys or program records.
Other effects	Perceived and measured benefits of EE (including NEBs) Other indicators of market progress, dependent on program design.

market share, stocking behaviour, availability, prices, savings, and other metrics, others focus on a few indicators as proxies for broader market progress. For example, while a number of the California utilities collect a wide array of metrics, other agencies in Wisconsin concentrate on tracking just a few per program.[3] Those agencies tracking fewer metrics felt it allowed them to track each metric more carefully, and cost less than more exhaustive approaches. Of course, a trade-off is comprehensiveness.

Interviewees reported a variety of levels of effort applied to collection of tracking data. These included: 1) extensive collection of sales data; 2) collection of state-level shipment data; and 3) random survey methods. Progress has been made in detailed market share tracking – particularly on the residential side. Early efforts had difficulty getting retailers or manufacturers to report sales consistently; California's more recent efforts are finding greater success with labor-intensive efforts that rely on establishing on-going telephone relationships with retail outlets and reminding them to submit periodic sales data. This is an expensive approach, but with a good retailer sample, it can provide high quality data[4].

Wisconsin purchases state level "shipments" data for key appliances. These data are compared to national data that have been stripped of the impact of the fairly aggressive sales in California, the west, and northeast. This approach is significantly less costly than detailed sales tracking, but may only be applicable in some areas of the country. Wisconsin notes that its situation differs significantly from some other states; there is relatively little cross-border sales or re-shipping – a problem that would be much more significant in other locations. This approach is more difficult for tracking lighting or commercial measures, where sales data are less readily available from secondary sources.

Finally, some states use consumer or business surveys to collect data on sales. On the residential side, several states and utilities (including Wisconsin and others) use large-scale surveys of residences to estimate the share of appliance sales that are energy efficient or ENERGY STAR®. Wisconsin has collected data on washers, refrigerators, and other appliances. New York has used this approach to collect data on residential lighting and appliances for several years. Depending on the sample size, this approach can provide good information on market share itself, and can also be used to confirm or leverage data from shipments and other sources. While telephone or mail surveys are not very expensive, collecting this information can be relatively more expensive because only a share of respondents purchased items of interest to the program, and only a share of those selected energy efficient equipment. These data are even more difficult to collect on the commercial side, because specifics about equipment may best be known by persons that are not readily available (including an array of contractors or architects and engineers).

We assembled lists of scores of specific indicators used by programs from around the country. These included metrics for residential, commercial, renewables, mid-market, outreach, and other programs. In general, the state of the art in market progress indicators – emanating from and linked to

3. For example, for one year, an agency in Wisconsin tracked market share of ECM motors, increased certification of builders (into program) compared to baseline, and increase in builder and trade ally events and initiatives.
4. Some utilities and states also mentioned gathering data from participating retailers as part of partner agreements. These data can be considerably less expensive to collect; however, they represent sales from partner stores (which do not represent all sales, but can be useful as an indicator). The quality of the data depends on the consistency of this reporting.

program theory and logic – tended to fall into several key topic areas. These included awareness / knowledge, availability and practices, market share, savings, incremental cost, and other effects (often specific to the program). Table 1 lists the types of tracking that were conducted for each of these topic areas.

These tracking practices may be useful to evaluators, provide ideas for fruitful approaches, and possibly help develop methods that are practical elsewhere.

Examining Net To Gross

Translating gross program-tracked energy and demand savings into just that share that can be specifically attributed to the program is a complex problem in evaluation. Regulatory agencies call for the use of net program impacts in the benefit-cost analysis. [5] Net program impacts reflect gross changes adjusted to account for the combination of two main effects:

- Net effect: a reduction in the gross effect for "free ridership", or that share of program participants that would have undertaken the efficiency behavior or implemented the efficiency measure(s) even without the influence of the program or its market interventions.
- Market effect: an additive adjustment to gross impacts, accounting for the positive impact and increases in efficiency that the program may have on market actors and actions above and beyond direct program participants.

Given that the evaluation is attempting to measure changes due to the program, and specifically effects above and beyond what would have happened without the program, free ridership (or net effects) are a key component. Free ridership addresses the set of program participants that would have purchased the energy efficient measure, or adopted the behavior, even without the influence of the program – that is, the program was not instrumental in the participant adopting the desired change. The easiest case to illustrate is that of a customer that got a rebate (paid through the program – a participant) but they would have purchased exactly the same measure even without the rebate. Given that the smallest this factor can be is zero, this factor always reduces the gross savings attributable to the program.

Spillover, on the other hand, attempts to measure the impacts that the program caused in the market through changes in behavior and purchases that were not included in the program records or assessment of gross savings. There are at least three kinds of spillover:[6]

- *Inside spillover:* Inside project spillover is defined as additional measures installed or practices used within the program building that were not necessarily required by or incented directly by the program or its incentives – but were installed because of the influence of the program. These are measures not incorporated into program records or accounted for in program savings computations. In some cases, behavioral effects may also be incorporated, if relevant.
- *Outside spillover:* Outside project spillover is defined as energy efficiency measures or practices installed by participating builders or owners at other buildings that did not qualify for the program or were not participating in the program. This factor accounts for the increase in efficient measures / practices adopted because of the influence of the program even without direct incentives.
- *Non-participant spillover:* Non-participant spillover in this definition refers to actions (installation of energy efficiency measures or practices) taken by builders or owners or other actors who were not participants in the program in buildings that are not participating buildings. This may include, for instance, use of energy efficient practices or increases in efficiency of equipment installed because of market pressure caused by participating builders, or demand by owners, etc. We are not including changes in stocking practices as an end in itself, but only the impacts related to additional installations that can be attributed to the program's influence (e.g. outreach, market competition).

Methods of combining free ridership and spillover to produce an NTG ratio vary. Some evaluators subtract spillover from free ridership, while others use a spillover multiplier (i.e., the value [1 + spillover]) to adjust the attribution factor (i.e., the value [1 – free ridership]).

Both difference-of-differences and self-report are imperfect methods of measuring net impacts attributable to programs. Our approach is based on the self-report method. However, the quality of analyses from the self report approach can be improved through the application of several enhancements to self report surveys. The approach uses multiple steps and refinements to attribute progress in market indicators and energy savings to program influences – effects above and beyond what would have occurred naturally without the program. A key element is to recognize that the purpose of the analysis is to provide information for program decision-making, and specifically for significant financial investments. As a result, it is important to provide results that meet a reasonable burden of proof at several stages – operation of technology, measurement of effects, and attribution.

The approach we use is robust in that it incorporates several enhancements on the basic survey approach.[7]

Improved Methods of getting Feedback from Relevant Actors

- *Clarifying Definitions and Groups Upfront -- Participants vs. Non-participants:* Explicitly defining program participants and non-participants is an important step that simplifies the work and analyses that follow. While in some cases it may be obvious, in other cases – particularly market transformation programs and programs without rebates, etc. – participants are not trivial to identify, and there

5. Standard Practice Manual: Economic Analysis of Demand-Side Management Programs, California Energy Commission, December 1987.
6. It is also important to examine behavioral changes in both residential and non-residential programs as appropriate for the program.
7. This work represents additions and enhancements built over the process of a number of projects over the years, including work conducted with Summit Blue Consulting, PG&E, Seattle City Light, NYSERDA, Wisconsin FOE, and other projects.

may be a host of "levels" of participation. Clarifying these groups up front, and structuring the appropriate interview guides and stratification – is an important step in the study design. Furthermore, it is critical to identify the appropriate baseline for the analysis; for example, will code be used as the baseline, or will the surveys need to incorporate information on "standard practices" (or will both be considered in the scenario analyses).

- *Feedback from multiple perspectives and actors:* We gather information on effects (specifically free ridership and elements of spillover) from several different actors involved in the program. For the residential sector this may include both builders and owners, or contractors and owners. For non-residential buildings it may include architects / engineers and owners. Each has useful information about both free ridership and spillover. In addition, while owners may be able to provide information on inside spillover, they are generally not able to provide very useful information related to non-participant spillover and other decision-makers must be relied on for this feedback.
- *Feedback from several points in the decision process:* For many years, we and other evaluators have recommended to clients that data collection for evaluation should be incorporated as part of the forms and submittals for program participation and incentives. In the cases in which this recommendation has been adopted, this has proved to be a valuable source of data for measuring elements of net to gross. It provides several advantages. It is inexpensive to implement (added as a form or incorporated into existing forms for programs). If incorporated into existing forms, it garners high response rate. The data are collected close to the point in time decisions are made – not a year or two after participation in the program. Finally, this method provides on-going data collection, so evaluations may be conducted more regularly with lower cost.
- *Input from experienced staff and program implementers:* We also gather detailed information on free ridership and spillover from program staff and/or implementers. They have years of experience and knowledge about the program, and this information is rarely integrated into the analysis in a formal way.

Improved Questions

- *Incremental / multi-stage questions:* We ask about free ridership in three steps, so that the results are consistent, and so that if the respondent cannot answer the more detailed questions at the end, useful partial feedback is still provided and the observation is not lost.[8] We ask 1) whether equipment works or performs as expected and whether the effect exists, 2) whether that effect was large vs. small, and 3) a more quantitative assessment of the size of the effect. By asking in stages, we obtain confirmation on the earlier issue – even if the respondent "drops out". We also gain information confirming that the effect exists, its relative size, and other information that can be used to benchmark and confirm other information provided. Multi-stage questions can help increase the proof of the impact and its size.
- *Ranges:* We ask questions in a way that walks the respondent through the thought process associated with free ridership, specifically asking for minimum and maximum values as well as a most likely value.[9] We have also always been careful to avoid limiting responses or estimates to point estimates. This uses more of the information gathered during data collection, provides a band of confidence around results, and more fully reflects the range of impacts induced by the programs. This approach allows the numbers to be "bounded", more reliable, and more consistent with the respondent's intent. It supports scenario analysis that is more related to responses given, than to arbitrary scenario settings.
- *Related Questions:* We gather information from several angles: that is, we ask percent of savings or measures that would have been installed without the program[10], as well as the likelihood that some of the measures would have been installed, and the likelihood that all the measures would have been installed, what share at the same efficiency, timing issues, etc. This provides feedback for the variety of situations encountered by different program designs (and measures)– including the "all or nothing" (0%/100% of measures) situation, where a set of program measures would vs. would not have gone forward (in total or in part) depending on the presence of program.[11]
- *Corroborating Information / Influencing Factors:* We gather "corroborating" information to help assess the responses on key free ridership questions. The corroborating information (other aspects of program influence on decision making, previous plans, effects on timing of installations, changes in energy efficiency, etc.) is used to interpret and adjust free ridership values in logical and consistent ways to improve the reliability and consistency of the results. It also provides indicative feedback on whether the respondent fully understood the concept behind the free ridership or spillover questions.

Data Collection Improvements

- *Experienced Interviewers and Probing:* We use experienced staff to conduct interviews – staff that know exactly how the data will be used and can probe to clarify responses. It is important to set up data collection procedures that assure the goals of attribution of net observed effects to

8. In this way, the partial information can often be used through interpolation based on other respondents. We report information on the percent that believe the effect exists, as well as the final size / value of the effect in the final report. We believe both are relevant.
9. Use of a three step approach asking explicitly about minimum, maximum, and best guess for the free ridership figure was an enhancement suggested by Summit Blue Consulting. See also Skumatz, Violette, and Woods, "Successful Techniques for Identifying, Measuring, and Attributing Causality in Residential Programs", American Council for an Energy Efficient Economy Summer Study on Buildings held in Asilomar, CA, ACEEE, Washington DC, August 2004.
10. And if survey time allows, ask the inverse.
11. Time factors are also incorporated (consistent with suggestions in the literature, examining whether or not the measures would have been installed within a year, although this is less relevant for new construction programs.

the program can be accomplished – accounting for free riders, snapback, and several types of spillover to assure net effects are properly attributed. Callbacks at least 5 times at various times of day and days of week are important, and interviewers are sensitive to busy schedules for interviewees. We make appointments, conduct half the survey at two separate times, and other methods to get the best information possible. We also ask if the respondent will be willing to have a callback – we find this extremely valuable in case there are confounding data identified in the analysis phase. Therefore, we do not need to "guess" at interpretations of responses.

- *Better data from fewer respondents:* Large sample sizes are not always best. We believe it is more important to gather good data than much data (within limits). Large scale surveys using staff that do not understand how the data will be used provide data that have oftentimes been problematic. The statistics establishing the sample sizes needed to achieve various accuracy levels for confidence intervals are predicated on the questions being accurately answered. Complex free ridership and spillover questions do not lend themselves to quick answers without explanation so a large number of this type of response may well be less useful than fewer, more well-understood responses. Using the available budget to poll smaller samples with experienced staff that understand the concepts and how the data will be used can be important in achieving good quality estimates of net to gross.

Analysis and Results

- *Comparisons and Feedback:* Comparing interim results with similar programs elsewhere is an important step in the analysis. For example, it provides an opportunity to benchmark results and examine where and why results differ from other programs – were there design or targeting differences that led to different results, or is one program older, which might tend to lead to higher free ridership and higher spillover. This analysis can provide additional confidence in the results. Examining differences may also identify possible program design or other changes that may provide suggestions to improve performance from the program. In addition, reviewing draft results with program staff provides important feedback on omitted issues, or caveats concerning results. Staff have been following the program for some time, and their feedback provides important context that might otherwise leave important issues unexplored.
- *Presentation of Results:* The results are provided in terms of a point estimate, but also as a distribution to help decision-makers understand the level of confidence in the information and to specifically incorporate uncertainty in the work. A point estimate that everyone knows is "wrong" is less useful than an interval that has a high degree of confidence and can be used to support scenario analysis and provide very useful information for program decision-making. This is not simply a computed statistical confidence interval; rather, we provide information on the percent of responses that were in different levels, and estimates of high and low bounds supported by the data. This provides a much fuller understating of the market effects than a simple point estimates. Scenario analysis of the cost-benefit analysis based on ranges for net to gross, and also related to attributed non-energy benefits (see later section of this paper) provide the most credible analysis of program effects.

While it is true that a basic attribution questionnaire can be relatively short, this enhanced version can be quite long.[12] This affects response rates and cost, and can lead to survey fatigue and lower quality responses. The surveys can be broken into two, incentives can be used, or other methods employed to improve survey responses. One key is that the survey can probably not be loaded down with a number of questions regarding awareness, sales, and other issues. A dedicated NTG / attribution questionnaire (or NTG/NEB questionnaire) is probably necessary separate from other market tracking information desired from these same audiences.

RESULTS FOR C&I AND RESIDENTIAL NTG RATIOS ACROSS THE US

The research indicated that administrators of energy efficiency programs throughout the country have applied a variety of methods to quantify the energy savings attributable to their programs and to characterize the related market impacts. Although the methods and programs vary, a review of program evaluation studies provides a useful landscape of typical findings regarding free ridership, spillover, and overall attribution as measured by a net-to-gross ratio that adjusts for these factors. Our research reviewed more than 60 residential and non-residential reports from which we catalogued information on NTG ratios, and free-ridership or spillover methodologies or results. This information was analyzed to identify patterns in results for common target sectors, end uses, and measures – and to provide comparisons to research efforts by the authors.

At least two states, Massachusetts and California, have developed standardized methods and assumptions for attributing energy savings to program efforts. California's 2001 Energy Efficiency Policy Manual stipulates NTG ratios to be applied to a variety of residential, non residential, and new construction programs.[13] In Massachusetts, work has proceeded on developing standardized methods for measuring free-ridership and spillover. Preliminary free ridership and spillover estimates for several programs have recently been developed by agencies in Massachusetts, but it is too early to determine how the new, standardized methodology will impact future attribution findings from Massachusetts utilities in general.

The differences in program focus, such as the incentives and assistance provided and the end-uses targeted, make comparison of attribution results between utilities difficult.

12. especially if non-energy benefits questions are incorporated (see section below)
13. A new evaluation framework is currently under development in California that may provide utilities more flexibility in determining appropriate net-to-gross ratios for program attribution of energy savings.

Table 2. Ranges of Energy Savings Attribution for Non-Residential New Construction and Retrofit Programs.

	Free-ridership	Spillover	Net-to-Gross Ratio*
New Construction	0.17 to 0.41	0.05 to 0.37	0.62 to 0.94
Retrofit**	0.15 to 0.27	0.08 to 0.17	0.81 to 0.94

* As mentioned in the text, there are several methods of computing NTG currently being used. Some evaluators subtract spillover from free ridership, while others use a spillover multiplier (i.e., the value [1 + spillover]) to adjust the attribution factor (i.e., the value [1 – free ridership]).

** Retrofit programs delivered through performance contracting are not included in these figures. Some performance contracting programs have NTG ratios in the range of 50% or less.

Table 3. Ranges of Energy Savings Attribution for Various Residential Programs and Measures.

	Free-ridership	Spillover	Net-to-Gross Ratio
Air conditioners			0.3 to 0.9
Cooking equipment			0.2 to 0.6
CFL	0.1 to 0.5	0 to 0.3	0.8 to 0.9
Clothes washers	<0.1		0.9 to 1.0
Dryers			0.4 to 0.7
Whole house		<0.1 to 0.2	0.8 to 0.9
HVAC			0.2 to 1.0
Lighting equipment	<0.1	<0.1 to 0.3	0.8 to 1.1
Torchieres	<0.1	<0.1	0.9 to 1.1
Low Income	<0.1	0	0.9 to 1.0
Refrigerator incentive programs			0.8-0.9
Spare refrigerator			0.4-0.5

However, it is apparent that findings can vary greatly by program type (e.g. new construction versus retrofit) and by end use. In general, new construction programs have higher free ridership rates than retrofit programs, but also higher spillover; the net effect is that the NTG ratios of new construction programs have somewhat higher variability (Table 2).

Free ridership and spillover vary greatly by end use, regardless of the program type. One efficiency rebate program, for example, indicates free ridership of approximately 4% for lighting and motors and 35% for HVAC systems. Our analysis of commercial sector results also indicated that free ridership was higher for lighting, HVAC, and motors than other measures.

On the residential side, studies providing net to gross information on more than 125 measures and programs were gathered and reviewed. In some cases, separate information for free ridership and spillover was also provided.[14] The results are grouped and presented in Table 3.

The research on net to gross figures showed indicative patterns and variations by sector, program, and measure type. The analysis also showed:

- *Consistency between studies:* While some measures showed fairly similar results between studies, others varied widely. Net to gross (NTG) results for similar program types for central air conditioners and other HVAC systems varied, but NTG figures for CFLs, clothes washers, and lighting equipment tended to be more consistent between studies.
- *Influences by program types:* As expected, NTG figures for spare refrigerator programs tended to be considerably lower (0.4-0.6) than for other refrigerator programs (>0.8). Low income programs showed high NTG figures.
- *Results by equipment type:* NTG figures for clothes washers, CFLs, and lighting measures were relatively high, with results over 0.9. Results were lower for dryers and cooking appliances.
- *Spillover and free rider results:* Information on free ridership and spillover was not reported in many of the reports. Clothes washer programs had few free riders, and the studies indicated low free ridership for lighting equipment and for low income programs as well.

A Step Beyond: Omitted Impacts and Value from Non-Energy Benefits

While energy savings, awareness, market share and other metrics provide direct indicators of program effects, a significant body of work has developed around recognizing and measuring net non-energy benefits (NEBs). This includes any and all impacts that are not directly the energy and bill savings resulting from the program. Previous work shows that these benefits are significant in relation to the energy savings, and are highly valued by participants. In some cases, the analysis suggests that the primary value from the program was non-energy benefits, rather than energy-related bill savings. Previous work also indicates that market actors – specifiers like builders, architects, engineers, contractors – also recognize these benefits and use them in "selling" energy efficiency.

14. Skumatz, Lisa A. 2004. "Leveraging and Review of Indicators and NTG Results from US Programs", Skumatz Economic Research Associates, Inc. Report 2004-04, Superior, CO.

NEBs include a variety of impacts that result from the program. Although the literature calls them non-energy benefits, they include the "net" of both positive and negative effects that may be attributable to the program. The convention has been established to separate these benefits into three "perspectives": [15]

- *Utility NEBs*: These include utility/ratepayer-type benefits result in reduced revenue requirements, including savings in a variety of administrative and carrying costs related to arrearages, service terminations, and related changes, as well as reductions in T&D losses when fewer kWh are distributed through the system. The changes attributable to these impacts are mostly valued at utility avoided costs for the relevant labor category, etc.
- *Societal NEBs*: Societal benefits include the value of reductions in emissions, economic stimulus, and similar public benefits. The values associated with these program-caused changes vary with the type of impact.
- *Participant NEBs*: Participant impacts include effects above and beyond energy savings, and include improvements in comfort, lighting quality, resident satisfaction, equipment maintenance benefits, safety issues, and a wide variety of other NEBs. While many of these indirect benefits may be difficult to measure, they can ultimately be translated into dollar terms, and incorporated as net program benefits accruing to participants.

Typical categories of benefits based on past work follow in Table 4 below. This list is not comprehensive, and obviously some benefits can cross categories. We tend not to include tertiary type benefits like tax –related impacts, as we prefer to be more conservative. Whether specific benefits are included or excluded from the analysis tends to depend on which measures are included in the program. The list of benefits to be included in the program attribution analysis is usually refined in collaboration with the program staff.

Note that several benefits arise in multiple categories. For example, having fewer bill-related calls to the utility benefits both the utility / ratepayers AND the households making or receiving those calls. This is not double-counting benefits – rather, it recognizes that some effects have multiple beneficiaries and each is valued at the appropriate tailored valuation method. For example, this saved time from calls may be valued at the marginal labor cost for customer service staff for the utility's benefit, and at the minimum wage rate for low income households. Benefits are recognized and realized by both groups; whether they are included in specific computations depends on their appropriateness to the application.

Attribution of utility and societal NEBs can be measured using a combination of primary and secondary data. There is an extensive literature measuring the arrearage impacts of programs (particularly low income programs), as well as many others of these impacts. Detailed examination of the program impacts – or the literature– may be needed to estimate the impacts on reconnections and other factors that may be affected by the program.[16]

Societal impacts also have a significant literature and indeed, the two key components, environmental and economic impacts – have a very high degree of volatility depending on the data sources and valuation methods used. Impacts on greenhouse gases (GHG) are increasing in importance and have been estimated in the literature. These impacts are a "slippery slope" – they can be estimated in a simplistic way, or if health impacts are to be measured in detail, then issues related to specific microclimates and time of day and zones are important. For some programs, average generation mix should be used to assess emissions; for others (e.g. a peak load reduction program, residential air conditioning programs, etc.) emissions from marginal peak load plants should be used to estimate changes in emissions from the energy savings. Valuations are the source of considerable debate in the literature as well.[17] There exists a literature estimating economic impacts from energy efficiency programs. Some of the literature are flawed in that they estimate the job creation and economic multipliers of a *gross* expenditure on the economy when instead they should be measuring the net impact of a switch from, say, the sectors included in electricity generation, into the economic sectors affected by the weatherization or other program.[18]

The most challenging portion of non-energy benefits work is assessing the participant portion of the benefits. The authors have spent considerable time on this issue, and have developed several credible methods of estimating these "hard to measure" (HTM) impacts. The authors have conducted scores of projects to measure net NEBs for both residential and commercial programs. We have pioneered a number of different approaches, and have had the opportunity to evaluate a number of them with respect to:

- Ease of response by respondent / comprehension of the question by respondents
- Reliability of the results / volatility
- Conservative / consistent results
- Computation clarity

We have pioneered and tested approaches including:

- Willingness to pay (WTP) / willingness to accept (WTA) / contingent valuation (CV)
- Alternative methods of comparative or relative valuations
- Direct computations of value to owner,
- Ordered logit, and

15. This convention is developed in Skumatz, 1997. "Recognizing all Program Benefits: Estimating the Non-Energy Benefits of PG&E's Venture Partner Pilot Program", 1997 IEPEC Energy Evaluation Conference, Chicago, Illinois.
16. See for example, Hall, Skumatz, and Megdal, "Low Income Public Purpose Test: Non-Energy Benefits for Low Income Weatherization Programs", prepared for PG&E, 2000 for an extensive discussion of these estimation methods.
17. For some clients, there are values that have been agreed upon by the regulators. For others, we used specific values included in the literature, or averages of valuations from many sources. Which valuations are most appropriate depends on not only the location, but also the use to which the work will be applied.
18. For an extensive discussion of the environmental and economic impacts, see Imbierowicz and Skumatz, " The Most Volatile Non-Energy Benefits (NEBs) – New Research Results "Homing In" On Environmental And Economic Impacts", American Council for an Energy Efficient Economy Summer Study, held in Asilomar, CA, ACEEE, Washington, DC, August 2004.

Table 4. Net Non-Energy Benefits (NEBs) Categories.

Utility Benefits		
Reduced carrying cost on arrearages (interest) Bad debt written off Shutoffs Reconnects Notices Customer calls / bill or emergency-related Other bill collection costs	Emergency gas service calls (for gas flex connector and other programs) Insurance savings Transmission and distribution savings (usually distribution only) Fewer substations, etc. Power quality / reliability Reduced subsidy payments (low income) Other	
Societal Benefits		
Economic benefits – direct and indirect multipliers Emissions / environmental (trading values and/or health / hazard benefits) Health and safety equipment Water and waste water treatment or supply plants Other		
Participant Benefits*		
Single Family Participants	**Multifamily Building Participants**	**Commercial/Industrial Participants**
Net program rebate (If relevant) Water / wastewater bill savings Equipment maintenance (labor and cost) Equipment performance / features Equipment lifetime Shutoffs Reconnects Property value benefits (Bill-related) calls to utility Aesthetics / appearance Fires / insurance damage (from gas-related audits/fix) Indoor air quality Moving costs / mobility Illnesses and lost days from work / school Transactions costs (complicated, not critical) Comfort Noise Safety Lighting / quality of light Feeling of greater control over bill (if relevant) Improved understanding of energy use / (if relevant) Feeling others "care" (low income only) NEGATIVES: Installation hassles / mess from installers – rest are mostly negative values for other factors above.	Water / wastewater bill savings Operating costs (non-energy)** Equipment maintenance Equipment performance (push air better, etc.) Equipment lifetime Tenant satisfaction / fewer tenant complaints Comfort Aesthetics / appearance Lighting / quality of light Noise Safety, insurance Health issues Ease of selling / leasing Labor requirements (separate from equipment O&M) Indoor air quality Doing good for environment Reliability of service / power quality Savings in other fuels or services (as relevant) Feeling of greater control over bill / understanding of energy use (residents if relevant) NEGATIVES (usually incorporated into above) some may have worse maintenance, parts may be harder to get, greater training needs for maintenance staff, etc.	Water / wastewater bill savings Operating costs (non-energy)** Equipment maintenance Equipment performance (push air better, etc.) Equipment lifetime Productivity Tenant satisfaction / fewer tenant complaints Comfort Aesthetics / appearance Lighting / quality of light Noise Safety Ease of selling / leasing Product losses (mostly refrigeration at grocery) Labor requirements Indoor air quality Health / lost days at work Doing good for environment Reliability of service / power quality Savings in other fuels or services (as relevant) NEGATIVES include: Production disruption during installation. Others are included above (some may have worse maintenance, etc.)

* Positive and negative impacts, estimated using participant surveys for many of the NEBs.

** Sometimes omitted if likely to double count with the next two categories.

- Other approaches.

These measurement methods can be complex to implement, and we have worked hard to refine the techniques. These techniques have been applied to the full gamut of benefits categories that are listed in Table 1. Our research over 10 years of performing these analyses has found that generally, comparative or relative valuations[19] perform substantially better than other methods. Willingness to pay (WTP) can often provide very volatile numbers and respondents have an extremely difficult time understanding the concept of stating a dollar amount they would be willing to pay for these benefits. We have incorporated multiple measurement methods into the same studies, and have found that on average, WTP is volatile (and less conservative), at least in a series of residential projects.[20] In addition, while non-residential respondents occasionally have information on the direct value of some of the benefits, they do not have

19. Methods pioneered and adapted by the authors, based on the academic literature; see descriptions in Skumatz, "Comparing Participant Valuation Results Using Three Advanced Survey Measurement Techniques: New Non-Energy Benefits (NEB) Computations of Participant Value", Proceedings of the 2002 ACEEE Summer Study on Energy Efficiency in Buildings held in Asilomar, CA, ACEEE, Washington, DC, August 2002.

Table 5. Orders of Magnitude Results / Rules of Thumb for Residential and Non-Residential NEBs* (Source: Skumatz Economic Research Associates research on more than 40 programs)

Topic	Tracking methods / indicators
Awareness / knowledge	Awareness and knowledge of program, features, efficiency, logos, and other specific program aspects / interventions among market actors gathered through surveys (Changes in) attitudes about programs, features, efficiency; perceptions of future trends
Product service / availability and practices	Actors undertaking program activities / enrollment, etc. Specifier practices (builders, contractors, A&E, owners, etc.) Stocking and sales practices (retailers, contractors, etc.); product availability; perception of the role of efficient equipment in business success
Market share	Sales or market share of relevant equipment tracked through shipments or sales garnered from third parties (AHAM, D&R, etc.) Tracking through (labour-intensive) calls / relationships with distributors (like California); In-store model counts / comparisons by mystery shoppers or circuit riders; Sales reporting that may be required of participating retailers; and Other methods.
Energy savings	Gross savings measured through program records Net savings through analysis of inspections, realization rates, and Net to Gross analysis of free riders and spillover (discussed below)
Incremental cost	Indicators of incremental cost, competitiveness, price premiums, often from surveys or program records.
Other effects	Perceived and measured benefits of EE (including NEBs) Other indicators of market progress, dependent on program design.

See Skumatz, "Non-Energy Benefits (Nebs) – A Comprehensive Analysis And Modeling Of Nebs For Commercial & Residential Programs", Proceedings from the 2001 Association of Energy Service Professionals (AESP) Conference, Ponte Vedra, Florida, December 2001.and Skumatz, "Non-Energy Benefits Including Productivity, Liability, Tenant Satisfaction, and Others: What Participant Surveys Tell Us About Designing and Marketing Commercial Programs", Proceedings of the 2002 ACEEE Summer Study on Energy Efficiency in Buildings, Asilomar, Washington, DC, August 2002.

information on the value of other benefit categories, leading to incomplete assessments.

We have found that relative valuations are by far the superior and comprehensive measurement method for many hard-to-measure effects. Respondents can readily answer whether these other benefits are more valuable or less valuable than energy savings or another benchmark.[21] This is important. If respondents have t puzzle a long time over a method for responding, the answers they give are likely to be guesses. Our field experience with thousands of interviews and surveys has shown that this is not a problem for the relative responses – most respondents can talk at length about these answers. In addition, asking value relative to savings helps eliminate the "time" aspect of the values – we do not have to worry about figuring out what discount rates they may be using if we ask total value or some other valuation approach. Using relative responses tends to reduce the number of "outliers" who generally have no idea where to start with WTP responses. Direct respondents' valuations of benefits (based on studies) are also fine methods, but many respondents have not conducted these studies (e.g. residential), or have conducted them on only a few side benefits, so we end up with too many missing values to be useful.

To provide credible estimates of the NEBs actually attributable to the program, the results must be "net" in several key ways.

- *Net positive and negative:* Despite the historical name for these impacts (non-energy benefits), both positive and negative impacts must be incorporated. The term we use is "net non-energy benefits" (NNEBs) but we will refer to them as "NEBs" in this paper. We generally ask questions in three levels – 1) whether they can name any benefits (to get at existence), 2) whether they experienced any impact from each of a set of NEB categories -- positive, negative, or no impact, and 3) the size of the benefits using one or more of several approaches mentioned above.
- Compare efficient to standard equipment: To attribute the impact due to the program, the respondents need to be asked about the NEBs for the new efficient equipment relative to the base non-efficient equipment that would otherwise have been purchased. The appropriate comparison is generally not the new efficient equipment but the old equipment that was in place.[22]
- Net of free riders: Similarly, if there are free riders that would have purchased the same equipment without the program, then the NEBs associated with that equipment should not be attributed to the program.

These nuances have not always been incorporated in NEB research.

20. For an analysis of comparative, willingness to pay, and labelled magnitude scaling methods, see Skumatz, Lisa A., Ph.D., "Comparing Participant Valuation Results Using Three Advanced Survey Measurement Techniques: New Non-Energy Benefits (NEB) Computations of Participant Value", Proceedings of the 2002 ACEEE Summer Study on Energy Efficiency in Buildings, held in Asilomar, CA, American Council for an Energy Efficient Economy, Washington, DC, August 2002.

21. This was a key development – this lever or relative approach (see "Recognizing All Program Benefits: Estimating the Non-Energy Benefits of PG&E's Venture Partner Pilot Program (VPP)", 1997 Energy Evaluation Conference, Chicago, August 1997. Skumatz (1996). We find that if we can ask value relative to some other item for which we can assign a dollar amount separately, we can then apply the "relative" figure to the value for our benchmark to compute the total value of the NEB. The lever we use varies by program.

22. However, some caveats are needed, depending on how the work is to be used. It may be that in the case of residents that would not have purchased new equipment at all without the program, a case may be made that for participant NNEBs, they recognize all the change from old equipment to the new efficient equipment. Also, if the measures would not have been installed for a period of time, the full NNEBs may be appropriately credited (as should the savings) during the interim. However, these are fine points on the principles discussed above.

NEBS RESULTS

Table 5 below provides overall findings related to the magnitude of NNEBs from residential and non-residential programs. NEBs are hard to measure, which has led to exclusion of these benefits from program evaluation and the decisions that derive from the evaluations. However, that practice leads to the implicit use of the value of "zero" as the NEB. The results below show how wrong that simplification can be.

These values are presented in terms of relation to energy savings, because dollar values would vary based on the size of the program / measures. This provides a better / more consistent basis for comparison.

Certainly the results and currency valuations vary by type of program, measures included, weather, and program targets. For instance, we find programs targeted at households with chronically ill inhabitants to lead to higher benefits from measures that address comfort. We find including gas measures can increase the NEBs associated with the program. Comfort-related NEBs are higher (especially in currency value) in climates with higher heating (or cooling) needs. However, the table provides some information on the range of omitted NEB values.

USES OF NEBS

One application for the NEBs analysis is benefit cost analysis. While perhaps not the primary use of NEB work, these measurement efforts assess (and value) another set of impacts of the programs above and beyond what would have been realized without the program. The individual categories of NEBs for a variety of types of programs were listed in Table 4. The results of the dollar NEBs valuations for each of these individual NEB categories varies based on the program's design, measures included, sectors, targets, area of the country, and other factors. Estimated properly, these represent attributable impacts; however, only a subset of these impacts may be appropriate for inclusion into a benefit cost analysis or into specific regulatory tests. For example, some of the societal benefits may belong in societal tests, and a number of the participant benefits may be appropriate for inclusion in tests for low income programs, because these programs often have reductions in hardship and bill-payment improvements as specific program goals.

Identifying the particular subset of NEBs to be included in a particular computation for a program depends on the application / use of the computation. NEBs that should be included for various applications are not appropriate for other computations, and rarely is it appropriate to include all the NEBs in a computation. On the other hand, perhaps it should be equally rare to include none of the NEBs in a computation of program effects.

Early applications of NEBs research applied to estimating savings to utilities (e.g. lower arrearages, etc.). This expanded to the use of NEBs for improving benefit cost analysis. However, NEBs are not only useful in assessing value from the program, but we have also found that this analysis provides a more sophisticated method of analyzing benefits and barriers.[23] Further, NEBs provide exceptional guidance for program targeting, marketing, and design. We find assessment of these net non-energy benefits is critical to understanding the full range of benefits provided by programs. They have a variety of uses:

- Benefit cost analysis (using subsets or pieces of the NEB analysis[24])
- Measuring barriers to adoption of programs.
- Program design: Selecting measures and selecting target groups for participants that will maximize the program impact to one or more audiences / perspectives (utilities, participants, etc.), potentially based on program goals. This can maximize the program "bang" given a fixed budget.
- Marketing / targeting programs to provide maximum benefits or target groups receiving high benefits and designing programs to achieve greatest total value related to program goals. Proctor & Gamble doesn't sell households Tide based on "buy this because it gives us greatest profits", and the implication of the NEB results has shown that selling efficiency programs on energy efficiency / conservation which is important to program design is a poor approach – these are often not the highest valued benefits participants derive from programs, and energy efficiency may not be the most appealing to advertise. The NEB results also indicate that it may not be the most important feature people want to buy. The single most common quote we get in our interviews is "well, we may have gotten some energy savings from the program, but what we really noticed is…". People are sceptical about the savings, and they "wash out" for commercial buildings in the midst of all the other things that change year to year.
- Marketing programs to appeal to participants based on the types of benefits that they actually value to improve the "bang for the buck" in outreach expenditures.

Benefit Cost Enhancements

Applying some of the methods presented here supports an improved program benefit cost analysis. The NTG information derived includes not only point estimates, but ranges that better reflect the translation of the program's net energy and demand benefits. Combined with the NEB information, the results provide the basis for a robust analysis of program impacts, and scenario analysis for both benefit cost (B/C) and regulatory tests. For example, B/C results may be provided incorporating a subset of NEBs, or scenarios may be constructed using alternate proportions of the NEBs – figures on the order of 10%, 25%, and 50% have been used by some agencies to reflect the NEBs but to be conservative and acknowledge that the NEBs may not have been included in the initial goals for the programs. In addition to reflect-

23. Dollar values of importance are a much more useful way of assessing barriers than a 1-5 "importance" scale.
24. For example, the authors were involved in identifying those categories of net NEBs most appropriate to be included in a revised public purpose test for low income programs in California, and our NEB work was used to establish NEB "adders" 'in states in New England, etc.

ing risk in the results through use of NEB ratios and ranges, the ranges from the NTG elements can also be used to construct decision trees or other methods of addressing risk. These techniques are well supported using the data collection approach suggested in this paper. Most importantly, use of scenarios or methods of incorporating risk are more credible than point estimates to multiple decimal points, are more robust, and makes it clearer that the measurement of program impacts is not an exact science and a derived point estimate is not the "answer". Rather, the ranges are more likely to include the correct estimate, and the scenario work tends to put the treatment of program benefits and costs on a more parallel track with supply resources.

Summary and Conclusions

The study conducted interviews with evaluators, utility staff, regulators, and others from agencies around the country. We gathered information on market progress tracking, indicators, and NTG results. The results showed that there are significant different philosophies in tracking – some states are tracking laundry lists of market progress indicators, and others endorse a more minimalist approach, selecting just a couple indicators by which to monitor progress. Across the states, tracking is accomplished via detailed data collection from distributors or retailers; purchase of secondary data; end-user surveys, or other methods. One example is sales of residential appliances. Some states use shipments (which are easier to obtain), especially if they are fairly isolated and do not have reasons to suspect significant shipments across state lines, but others have worked to develop sources for true sales tracking.

Interviews with staff and consultants tasked with developing indicators provided information on metrics in a number of key areas, including sales and market share; awareness and knowledge; specification, stocking, and sales practices; attitudes; incremental costs; non-energy benefits, and other program-specific indicators.

We presented suggestions for enhancements to the attribution of energy savings impacts, and the analysis of net to gross effects. The methods work within the theory-based evaluation framework, but also work toward providing a burden of proof on par with the criteria for other public and private investments. The section addressed additional steps to improve: methods of obtaining feedback from relevant actors; question development and content; data collection; and analysis and results.

Specifically, the work suggests using the distributions of the measure and intervention impacts, rather than relying on point estimates. In addition, the work described enhanced method to address free riders, spillover, and free drivers to allow for partial free riders and used indicator methods to provide evidence on program-induced effects on spillover. We also directly addressed the issue of uncertainty and risk in the attribution work using scenarios to "bound" the effects. Thus, the attribution work examined several methods of addressing risk – an important component of using causal results. This approach uses much more of the information gathered during the data collection, and more fully reflects the range of impacts induced by the program. Although causality may not be strictly "proveable", these extra steps improved the reliability and robustness of the results of the causality analysis and provide a better foundation to guide program and investment decisions – one of the most important goals of an evaluation.

The research on net to gross figures showed indicative patterns and variations by sector, program, and measure type. In the commercial sector, free ridership was higher for lighting, HVAC, and motors than other measures, and NTG was often higher for retrofit than new commercial new construction. Among residential programs, low income programs had high NTG ratios, and NTG ratios tended to be higher for efficient clothes washers, CFLs, and lighting measures than others. This research may provide useful information to other agencies attempting to benchmark NTG results from other programs.

Non-energy benefits (NEBs) are an often-ignored, but important set of benefits provided by energy-related measures and features in residential and non–residential buildings. These NEBs can be measured, and they represent effects that are attributable to programs. Utilities may run energy conservation programs to reduce energy use, and builders may build homes and commercial buildings that include energy saving features, measures, and designs. However, energy savings may not be – and appear not to be – the highest valued outcome of these measures and features to buyers / participants. These NEB results have several applications.

- Benefit/Cost analysis: some share of these effects may be appropriate for inclusion in program benefit cost analysis or regulatory "tests".
- Program design: NEBs can be incorporated into initial decision-making about which measures / features to include in new / remodelled buildings (or into programs) and computations of costs and benefits from investment in energy using equipment. They can also be used to identify (and modify) key program barriers.
- Outreach / advertising to attract buyers to these efficient buildings or to energy conservation programs that incorporate these measures. The value of these other benefits – may be stronger selling points for the measures than energy savings – and these benefits should be used as key sales messages in program outreach.[25]

25. See Bicknell and Skumatz, "Non-Energy Benefits (NEBs) in the Commercial Sector: Results from Hundreds of Buildings", Proceedings of the 2004 ACEEE Summer Study, Asilomar, CA, August 2004.