Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution: Energy Savings, Net to Gross, Non-Energy Benefits, and Persistence of Energy Efficiency Behavior

Prepared by:

Principal Investigator and Primary Author
Lisa A. Skumatz, Ph.D.
Skumatz Economic Research Associates (SERA)

Contributing Authors
M. Sami Khawaja, Ph.D.
Jane Colby
The Cadmus Group

Funded by:
California Public Utilities Commission

Prepared for:

CIEE Behavior and Energy Program
Edward Vine, Program Manager

California Institute for Energy and Environment
2087 Addison, St., Second Floor
Berkeley, CA 94704

November 2009

DISCLAIMER

This report was prepared as an account of work sponsored by the California Public Utilities Commission. It does not necessarily represent the views of the Commission or any of its employees except to the extent, if any, that it has formally been approved by the Commission at a public meeting. For information regarding any such action, communicate directly with the Commission at 505 Van Ness Avenue, San Francisco, California 94102. Neither the Commission nor the State of California, nor any officer, employee, or any of its subcontractors or Subcontractors makes any warranty, express or implied, or assumes any legal liability whatsoever for the contents of this document.

ABSTRACT

This white paper examines four topics addressing evaluation, measurement, and attribution of direct and indirect effects to energy efficiency and behavioral programs:

- Estimates of program savings (gross);
- Net savings derivation through free ridership / net to gross analyses;
- Indirect non-energy benefits / impacts (e.g., comfort, convenience, emissions, jobs); and
- Persistence of savings.

Evaluation and attribution methods have reached a point that they must evolve in order to provide credible results for the next generation of programs. Two primary factors have complicated the methodologies that have been applied to energy efficiency programs:

- Transition to more behavioral, outreach and other non-measure-based programs (education, advertising), making it especially hard to "count" impacts, and
- Increased chatter in the marketplace, in which consumers may be influenced by any number of utility programs by the host/territorial utility (the "portfolio") as well as influences from outside the territorial utility (national, neighboring programs, movies/media).

We[1] reviewed hundreds of conference papers and interviewed scores of professional researchers to identify improved techniques (and associated policy issues) for quantifying the share of direct and indirect effects that can be attributed to the influence of program interventions above and beyond what would have occurred without the intervention – either naturally or due to the sway of other market influences or trends. We reviewed evaluation methods from around the US and Canada and examined evaluation practices in different states. We analyzed: issues / problems / gaps from current approaches; priority applications for the results and potential alternatives proposed or considered (and associated data needs); and proposed next steps in a research agenda. Finally, we also present near- and long-term implications for program design, evaluation, outreach, and benefit-cost for programs across the US; and best practices for key elements of evaluation of direct and indirect energy efficiency and behavioral program effects.

[1] The author wishes to thank the following for assistance in preparing this document: D. Juri Freeman, Dana D'Souza, and Dawn Bement (Skumatz Economic Research Associates), Dr. Carol Mulholland, Jamie Drakos, and Natalie Auer (Cadmus Group), and Gregg Eisenberg (Iron Mountain Consulting)..

ORGANIZATION OF THE REPORT

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

Introduction

In conducting this project for the California Public Utilities Commission (CPUC) and the California Institute for Energy and Environment (CIEE), the authors were tasked with identifying current and improved techniques – and associated policy issues – related to:

- **Gross Effects**: Measuring the broad array of impacts caused, or potentially caused, by program interventions – measure-based, market-based, education or other interventions. This includes the measurement of gross energy savings and non-energy impacts.
- **Net Effects Attribution**: Identifying the share of those effects – direct and indirect – that can be attributed to the influence of the interventions undertaken – above and beyond what would have occurred without the intervention – either naturally or due to the sway of other market influences or trends.

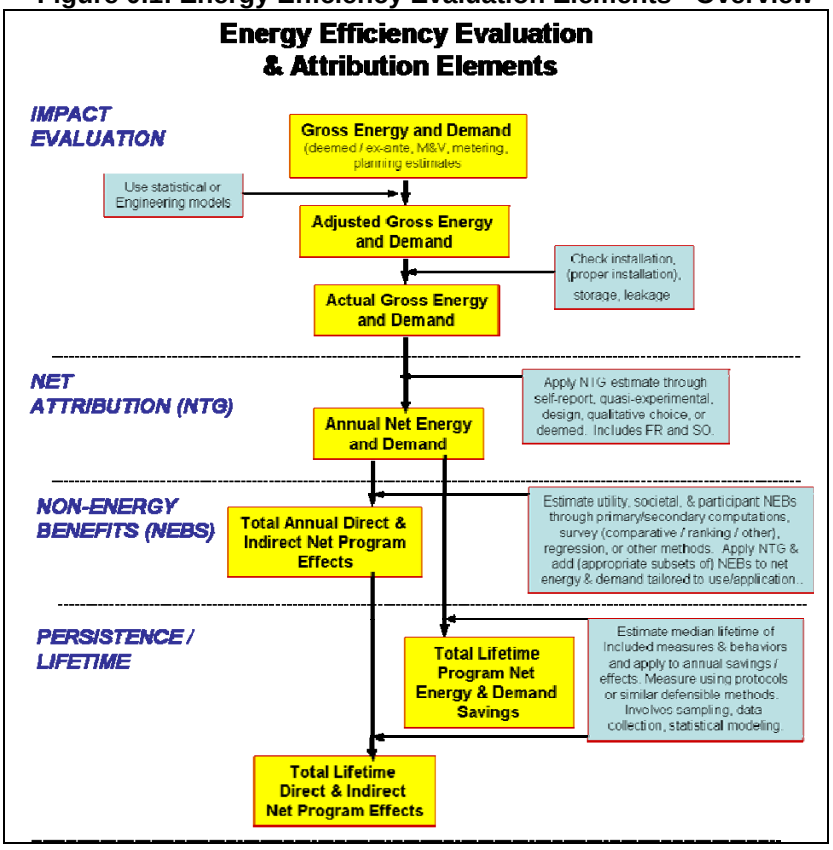
Using the current terminology, this boils down to examining four key topics in evaluation: impact evaluation; attribution / free ridership / net to gross; non-energy benefits; and persistence. The data and outputs from these evaluation topics are used for an array of applications, including:

- Measuring progress in the market – most often using share of sales / installation of energy efficiency (EE) equipment compared to standard equipment;
- Benefit-cost analysis for programs – generally using standardized regulatory tests;
- Attributing savings, and shareholder benefits, to entities investing in (specific) programs – applying gross impact evaluation values modified by net to gross attribution ratios consisting of free ridership, and some share of spillover;
- Comparing savings from EE to market needs and supply sources to assure energy demand needs are met – reviewing the cost per unit for EE vs. new supply, and the size and reliability of the kWh or therms;
- Program decision-making, marketing, and program design – using results of process, impact, free ridership, and other program evaluation elements to improve and optimize the program offering.

In each case, there are significant investment dollars at risk or associated; hence, the need to revisit methods and approaches. Further, as programs have evolved, evaluation has become more complex. Programs have moved away from “widget”-based programs toward education, advertising, and upstream programs that make it harder to “count” impacts. In addition, there is an increasing number of actors delivering these programs – leading to market “chatter” and increasing difficulty in identifying which among all the deliverers of the EE “message” are responsible for the change in energy efficiency behaviors, actions, or purchases. The increased chatter in the marketplace allows a situation in which consumers may be influenced by any number of utility programs by the host / territorial utility (the “portfolio”) as well as influences from outside the territorial utility (national, neighboring programs, movies / media, etc.). Attributing or assigning responsibility for changed behaviors, adoption of EE measures or similar

effects is muddied. Thus, separating out program influences has become more and more complex.

Figure 0.1: Energy Efficiency Evaluation Elements - Overview



There is considerable debate over precision – or lack (presumed) thereof – in association with a number of specific aspects of evaluation research. For example, many criticize the accuracy of free ridership or net to gross ratios, or deride the estimates of non-energy impacts. The 2003 Nobel-award winning economist, W.J. Granger, summarized the overall purpose of evaluation as '…*research designed to help avoid making wrong decisions (about programs)*'. As this relates to energy efficiency programs, perhaps the three most important potential "wrong decisions" might relate to the following topics:

1) Assuring public dollars are being responsibly spent;

2) Apportioning dollars and efforts between alternative strategies; and

3) Helping identify the appropriate time for exit strategies (or program revisions).

Perhaps this overriding principle is worth keeping in mind as we consider our standards for evaluation in energy efficiency. If this principle is accepted at least for some applications), then it becomes clear that the level of accuracy applied to evaluation research can be flexible, based on the value (cost) of the possibility of a wrong decision coming out of the particular advisory research. Identifying the cost of a "yes/no" decision about going ahead with a program or intervention may allow a much less accurate estimate for input information than a decision about the precise level of shareholder dollars that should be allowed for a particular agency, should that be a desired outcome to be supported by the evaluation exercise.

Finally, although we note multiple specific uses of the results of the analyses throughout the paper, we note several key uses of the results of the evaluation work and expected level of accuracy:

- **Program Planning**: Providing estimates of savings attributable to a program that can be used for program planning purposes – including potentially as elements of a programmatic benefit-cost test or other criteria used for program approval. This requires moderate to high accuracy.

- **Program Marketing and Optimization**: Providing quantitative feedback that helps to inform the design, delivery, marketing, or targeting of programs, including revisions to incentives, outreach, exit timing, or other feedback. The evaluation information can be used to understand tradeoffs, benefit-cost analysis, and decision making. This requires low to medium accuracy.
- **Integrated Planning, Portfolio Optimization, and Scenario Analysis**: Providing savings and other feedback across and between programs that helps optimize program portfolios. This requires medium accuracy and confidence intervals and alternative values for assessing risk.
- **Generation Alternative**: Providing an estimate of energy savings attributable to a program, which is, to some degree, suitable for comparison with energy delivered from a power plant and which supports confidence in generation deferral. This requires high accuracy and confidence intervals at the portfolio level.
- **Performance Incentives**: Providing estimates of savings attributable to a program that may be used to compute incentives to various agencies in return for efforts in program design, implementation, and delivery. This requires high accuracy and confidence intervals.

Certainly the level of accuracy associated with each may differ, but each of these is an application to which the types of measurements that we discuss in this paper have been used or have been proposed for use. We will refer to these uses throughout the paper.

This paper represents the preliminary results of research that involved outreach to more than 100 researchers in the energy evaluation and related fields, as well as review of more than 100 papers and reports representing research in the key topics covered by this paper. Although the topics certainly warrant even more work, budget constraints limited the scope and outreach for the paper. The work does, however, attempt to identify the state of the art and its strengths / weaknesses, potential improvements and how they relate to behavioral issues, and recommendations on next steps and next research directions.

Gross Energy Savings Measurement

The first step in the attribution of program effects from an energy efficiency intervention is developing an estimate of gross energy savings.

- **Standard Impact Evaluation methods:** Impact evaluations apply at least one of five general methods. 1) **Measurement and Verification (M&V),** which involves metering or estimating key parameters from a sample of participants and applying it to all participants.**2) Deemed Savings**, which involve applying "deemed" or agreed-upon savings obtained from other evaluations or manufacturers' data to all program participants.**3) Statistical Analyses**, which involve applying statistical regression models to utility billing or metering data of all program participants.**4) Market Progress / Market Share**, which uses information from sales, shipments, or other similar data to develop estimates of changes in sales (and implied usage) of program-recognized energy-efficient equipment relative to non-program equipment. Estimates of the associated energy

(and/or demand) savings are then calculated.[2] **5) Surveys**, which are often needed to estimate the savings-related changes from behavioral / educational / social marketing programs, perhaps in concert with the market progress methods described above. While there can be difficulties linking back to direct savings (and some simply don't try to count or evaluate these programs), experimental design with random assignment to test and control groups of adequate size can provide estimates.

These approaches have generally served to provide gross estimates of programs, even if there are a few issues arising because of the switch toward market and behavioral programs. Interviews with leaders in the field and review of the literature indicated a number of issues associated with the application of these methods to the evolving generation of programs:

- **Problems and best practices suggestions for (program design and) impact evaluations**: Our study indicates that an up-front understanding of *program goals* against which progress is being measured is not always available, thereby complicating evaluation. In addition, the field should consider more regularly conducting *market assessments* – up front – so it becomes clearer what actions are needed in the market and when a program should exit the market – and to allow better understanding of the market, identify needs, and provide a baseline for program evaluations. As part of that baseline work, market and appliance / equipment *saturation surveys* need to be re-introduced to allow better understanding of the market, identify needs, and provide a baseline for program evaluations.
- **Gaps and methodological improvements for impact studies***:* The study indicates that *logger studies* are needed for some types of measures (e.g., lighting) to improve the reliability of impact studies. There has been a gap in detailed assessment of behavioral programs, and the modeling approaches used for assessing behavioral programs could be improved.
- **Baseline and overlap issues**: There is a significant problem in using *program records* for establishing a baseline: this type of information is collected to support rebates and not evaluation, so that useful baseline data are not collected up-front. To date, no studies have identified revelatory methods of isolating impacts for individual programs from "noisy" markets (markets with multiple programs influencing behavior). Estimating the impacts from one program is difficult – many suggest it may only be possible to estimates market effects from entire portfolios of programs.
- **Adaptations for educational / behavioral programs**: Education and behavioral program evaluations have been evaluated, but tend to require *tailored*, rather than prescribed, *evaluation methods*. Impacts may be indirect in some cases, but direct and indirect impacts can be measured for many programs with up-front experimental design methods and sufficient sample sizes. Work in developing creative adaptations to better fit behavioral programs would be valuable.

[2] One innovative approach indirectly measures market share by estimating the effect on a decomposed price differential and tracking the size of the coefficient for the efficiency features of the measure(s). See Skumatz 2007 and Skumatz 2009.

Net Effects – Free Riders and Net to Gross (NTG)

Estimating the effects of the program above and beyond what would have happened without the program involves another step – identifying the share of energy-efficient measures installed / purchased that would have been installed / purchased without the program's efforts.

Thus, the following elements need to be considered: **1) Free riders (FR)**: Some purchasers would have purchased the measure without the program's incentive or intervention. They are called "free riders" – they received the incentive but didn't need it. 2) **Spillover (SO)**: Others may hear about the benefits of the energy-efficient equipment and may install it even though they do not directly receive the program's incentives for those installations. These are called "spillover" – attributable implementation of measures that were not recorded directly in the program's "count" of installations. 3) **Net-to-gross (NTG)**: Free ridership and spillover are estimated for calculating the "net to gross" (NTG) ratio, and are applied to the "gross" savings to provide an estimate of the attributable "net" savings for the program.

- Standard methods of treating NTG or its main components: For planning, incentives, and other purposes, the NTG, or its components, have been addressed in four main ways: 1) "**Deemed" (Stipulated) NTG**, where a particular NTG is assumed (1, 0.8, 0.7, etc.) that is applied to all programs or all programs of specific types. This is generally negotiated between utilities and regulators or assigned by regulators. **2) NTG adjusted by models with a dynamic baseline**: in this case, a baseline of growth of adoption of efficient measures is developed, and the gross computation of savings is adjusted by the estimate from the baseline for the period. 3) **Paired comparisons NTG**: Saturations (or changes in saturations) of equipment can be compared for the program (or "test") group vs. a control group. The control group is similar to the test area in all possible ways, but does not offer the program being studied – or those particular customers do not receive the program. Pre- and post- measurement in both test and control groups are ideal to allow strong "net" comparisons. 4) **Survey-based NTG**: In this approach, a sophisticated battery of questions is asked about whether the participant would have purchased the measures / adopted the behavior without the influence of the program. Those participating despite the program are the free ridership percentage. These are then netted out of the gross savings. Similarly, spillover batteries can also be administered to samples of potential spillover groups (participants, non-participants).

- **Including or excluding spillover or free ridership in program computations**: Spillover is more complicated than free ridership to measure, and as a consequence, a number of utilities that include free ridership never estimate spillover. Free ridership emanates from the pool of identified program participants; the effects from spillover are not realized from the participating projects and, in many cases, not even the entities that participated. Identifying who to contact to explore the issue of spillover and associated indirect effects can be daunting. However, given that many of the benefits from outreach and educational programs are realized from "spreading the word" (and the behaviors that follow), developing reliable and trusted methods of including free ridership in program computations should be a priority for future research.

- **NTG in regulatory applications**: There is a considerable – and growing - controversy regarding the use of net to gross, particularly in regulatory applications. NTG ratios can have large fiscal effects in some states in which utilities may receive financial awards for running programs and running them well. The argument is that the program carefully

estimates (gross) savings that were delivered, but then the savings (and, directly, the associated financial incentives to the agency delivering the program) are discounted by a free ridership factor measured by potentially less-than-reliable means. The controversy arises from concerns about error and uncertainty; cost; baselines; separation of program effects from marketplace chatter; and risk. Concerns arise that using measured NTG or free ridership ratios introduces a great deal (to some, an unacceptable level) of risk into the potential financial performance metrics for the program, and, as a consequence, leads program investments toward "same old / same old" programs, reducing innovation in program offerings. This controversy has only been fed by the fact that only a small minority of free ridership, spillover, or NTG studies report any confidence ranges or even discussions of uncertainty. Until these issues are addressed, given the financial implications, it is unlikely much additional progress will be made in more comprehensive treatment of FR, SO, or NTG in the regulatory realm. Because of their spillover implications, this puts educational (and potentially behavioral) programs at a disadvantage in portfolio development and rewards / incentives.

- **Uses for FR, SO, NTG – and errors from omission:** The literature indicates there are a number of other uses to which the free ridership, spillover, or net to gross ratios are relevant. Free ridership helps to identify superior program designs and helps to identify program exit timing. Spillover helps to assess the performance of education / outreach /behavioral programs, and it helps to identify program exit timing. Not examining free ridership and spillover *ex post* will make it impossible to distinguish and control for poorly designed / implemented programs, as well as for programs that may have declining performance over time and may have outlived their usefulness, at least in their current incarnation. Some interviewees said 'deemed savings are ridiculous' for this reason.
- **Accuracy, reliability, and incentive issues**: Reasonable reliability is needed to provide useful information. To provide the best chance for optimal programs, the following are needed. NTG or FR and spillover (SO) estimates that are as reliable and precise as needed for the particular use – with greater precision needed for the calculation of program or portfolio incentives vs. quasi-quantitative / qualitative uses. NTG or FR and SO estimates that provide replicable results and are based on credible, defensible estimation methods suited to the accuracy needed are a critical step in getting NTG results included in design and evaluation. Methods suited to different levels of accuracy for estimates of NTG, FR and SO at reasonable cost levels would help optimize expenditures where they are most needed, and balance the tradeoffs of program funds vs. evaluation expenditures. Similarly, there should be flexibility in the application of NTG, FR, and SO results depending on type of program (whether programs are new / innovative / pilot; "same-old-same-old"; cookie cutter; custom; information-based; etc.). Finally, it is critical that the application of NTG results is conducted in ways that avoid discouraging the development of new and creative and potentially effective programs. NTG should be applied in ways that properly assess program performance, but makes the risk of fiscal investment in (especially, new) programs manageable and reasonably predictable.
- **Defining acceptable NTG options**: The goal is to encourage good design and performance, but avoid stifling program innovation, and do so in a way that isn't too burdensome (analytically or budget-wise). The goal is to provide an approach that will address practicality in both how NTG elements are estimated and how they are used / applied. A case might be made that the most "accurate" metric is pure *ex post* measurement especially when those estimates are used for planning and reward

purposes. If the main "rub" arises when NTG elements are part of the computations of financial reward or program approval, there are several possible options for the short term (until a "grander" solution is identified). Short term deemed values (1-2 years of a new program that differs from traditional offerings) could be identified, allowing time for development and refinement of new, creative programs without punishing fiscal consequences. The program could be dropped if performance doesn't meet the offerer's expectations, and the method avoids an innovation penalty. True-up at some point is necessary to assure that the field learns about the performance of different types of programs and to assure that ineffective programs are not rewarded indefinitely. Deemed spillover values may be especially needed for programs targeted at education. Long term deemed values could be allowed for well-known program types based on measured NTG from programs around the nation, check program performance every 3 years, and penalize programs that perform more poorly than the norm, or require program comparisons against "best practices" periodically (every 3 or so years). Again, periodic true-up is needed.

- **Additional analyses needed**: Reliable measurement methods are available that suit many program types, but more work remains in the following areas:
    - Enhanced NTG, FR, and SO methods incorporating partial free ridership and corroborating information.
    - Experimental design including random assignment for participants and non-participants should be used for as many program types as feasible.
    - Comprehensive market assessment work for baseline support, on non-participant spillover, and modeling of decision-making. This is particularly important for many training, education, and behavioral programs.
    - Data collection approaches that introduce a real-time data collection element piggybacking on program handouts / materials / forms and to allow periodic reviews of performance in time to refine programs.
    - Discrete choice and other modeling methods, and statistical techniques to help address issues of imperfect control groups, unobserved factors, etc., to allow for improved estimates of attributable impacts.
    - Results on elements of NTG should be accumulated in a database and continuously updated with new research and evaluations, so comparisons and tracking are facilitated.

Non-Energy Benefits (NEBs)[3]

Non-energy benefits (NEBs) represent the positive and negative effects beyond energy savings and energy bill savings that are attributable to energy efficiency programs. Strictly speaking, NEBs are "omitted program effects" – impacts attributable to the program, but often ignored in program evaluation work. After years of research, more and more utilities and regulators are considering these effects in program design, benefit / cost analysis, marketing, and other applications. Research over the last 20 years has identified a wide range of NEBs, and sorted the constituent effects into three classes based on "beneficiary" or "perspective". These are: 1) utility-perspective NEBs realized as indirect costs or savings to the utility – and its ratepayers

[3] Also titled non-energy impacts in more recent literature, but there is no difference in definition or the effect being measured.

(like bill payment improvements, infrastructure savings, etc.). 2) Societal-perspective NEBs represent indirect program effects beyond those realized by ratepayers / utility or participants, but they accrue to society at large. 3) Participant-perspective NEBs accrue to the program participants. This is where factors like operations and maintenance, comfort, productivity, "doing good for the environment," and others arise.

Methodological basics / best practices in NEBs: While there are certainly measurement issues associated with estimating "hard to measure" (HTM) effects like NEBs, credibility also suggests that some basic methodological considerations be considered in assessing and attributing NEB effects to energy efficiency (EE) interventions. Best practices require addressing a number of methodological issues in NEB research as "standard practice". Attributable NEBs represent NET effects – positive and negative –beyond those that would accrue from standard efficiency equipment (with the possible exception of low income measures on this last point), and, ideally, net of free ridership / effects associated with the program. In addition, analyses should work to avoid overlap in definition of NEB categories within a perspective.

- **Progress in NEBs**: In the last decade, significant progress has been made in estimating several key categories of NEBs: emissions / GHG impacts, and economic development / job creation. Modeling approaches in GHG have improved dramatically, partly owing to the attention coming from implications for carbon trading and other applications. The literature shows three main methods, each representing an increase in accuracy and also cost: "grid or system average" values (average fuel mix for the entire year across the territory); (2) marginal operations (varying the emissions per kWh by type of fuel mix for peak / off peak and similar variations depending on the program and measures); or (3) hourly dispatch, examining 24/7 adjustments. To support use in trading schemes, the analyses need to address three measurement issues – additionality; program vs. project attribution; and error / risk / uncertainty issues. While each issue has been raised by many papers, none of the papers forwarded solutions, and the debate continues on the international stage. Third party modeling for economic impacts has improved substantially, providing feasible tools for examining and attributing credible estimates of job creation to energy efficiency programs. The literature available to date shows significant differences in job impacts based on program types – findings that have important potential implications in deciding among similarly-effective programs within a portfolio, especially at a state level. Other than these two topics, the greatest attention has focused on participant NEBs (discussed next).

- **Measurement of participant NEBs**: A large share of the literature in the last decade has focused on bringing more maturity to the methods for measuring participant-side NEBs. Because these rely on self-report surveys, and represent "hard to measure" benefit categories (comfort, etc.), significant work was needed. The literature has explored more than a dozen measurement approaches with grounding in the academic literature, and work proceeds on trying to identify methods that are accurate, but also feasible to implement. Each method has pros and cons, and a few studies have compared the performance of different measurement methods. The main purpose of each is to develop monetized estimates of the indirect impacts that can be assigned to the program. One key class of methods is "leading the pack", focused on variations in comparative contingent valuation approaches (as discussed in the chapter). Additional studies incorporating comparison of the performance of the key measurement methods are much-needed to improve confidence in participant NEBs. Only a few of these studies currently exist.

- **Programs with NEB results**: NEB studies have been applied to a wide variety of programs – including entire utility portfolios. NEB results are available for the wide variety of initiatives in the residential, commercial, and multifamily sectors, as well as for renewable, real-time pricing, commissioning, and low income weatherization programs. The results tend to show that utility benefits are fairly low, and the dollar value of benefits are realized from the societal (especially environmental and job creation) and participant perspectives. Several commercial studies report negative NEB values - and significant concerns – especially related to the maintenance of new, cutting edge energy-efficient equipment. The negative NEBs can be considered indicators of "barriers" to programs or measures. The computed values for just the participant perspective often exceed the value of the energy savings from the program measures. Although the papers varied in their estimation methods, all argued that the impacts were real, and were significant and merited continued analysis. The most common positive, highly valued NEBs related varied somewhat by programs and measures (especially on the commercial side). Highly valued residential NEBs tend to include comfort, operations and maintenance, ability to "do good" for the environment, and water savings. Highly valued positive effects for commercial programs tended to include comfort, operations / maintenance / lifetime, "doing good" for the environment, productivity, and performance issues.

- **NEBs for educational and behavioral programs**: NEBs have been applied liberally to behavioral and education programs – and it has been suggested these represent some of the key values of the programs. These include a variety of ENERGY STAR™ programs, weatherization and education programs, commercial training, and schools programs. The literature has also explored NEB values toward a more robust understanding of program participation and decision-making for direct participants and actors along the chain of delivery programs and measures.

- **Uses of NEBs**: Studies point out that, internally, in program design and evaluation, NEBs can be used for several key purposes: marketing and targeting to maximize the bang for the budget dollar; crafting the marketing message to "sell" the program or measures based on the features that most appeal to potential participants; identifying "negative" NEBs; examining the degree to which differences in the valuation of NEBs affects the actions of supply chain actors toward recommending / purchasing energy efficient equipment; selecting among measures to include in the program; examining tradeoffs in terms of measures with higher NEBs to provide maximum value for participants; estimating appropriate program incentives; and benefit-cost assessment. A review of current treatment of NEBs in regulatory tests finds evidence of utilities using NEBs in program marketing, in scenario analysis, as a project screening device, and as a program screen (but none are currently using it formally as a program screen in regulatory applications). NEBs may reflect some of the most important effects from energy efficiency measures and programs, and may especially represent some of the main outcomes of educational and behavioral programs.

- **Use of NEBs in regulatory applications**: While most utilities and regulators are not treating NEBs formally, some are examining them for marketing purposes. A few include "easily computed" or "readily measured" NEBs in formal analyses (e.g., soap and water savings for washing machine programs). One utility includes the percentages of NEBs in various scenarios it presents to the regulators. Although NEBs have been applied in less formal ways, they have been used only sparingly by utilities and regulators largely because of concerns about measurement uncertainty. For instance, many believe that

some NEBs (environmental and elements of participant benefits) should appropriately be introduced into the total resource cost (TRC) or societal test – an inclusion that would be consistent with the intent of the test and better represent attributes that differ between programs. Current regulatory tests, by omitting these impacts, may serve to discourage adoption of these programs. Although more than 10 years of research have measured NEBs, it remains unclear how quickly regulators or others may begin to incorporate NEBs into the program review process. Perhaps an important near-term step may be to report program metrics including various proportions of NEBs, which would demonstrate differences in the performance of different programs (for program selection), and might better reflect some of the differential values associated with education and behavioral – and other – programs. If these indicators can be allowed to influence some program choice, this may help avoid making suboptimal program choices.

- **Key NEB categories needing research:** Health and safety impacts have been very sparsely studied, even though the impacts on the health care system (including incidence of chronic illnesses) and productivity may, in fact, be quite large. Infrastructure (water and power) and national security impacts are gaining some attention.

- **Gaps and research needs in NEBs**: Although there are numerous large-scale studies of NEBs, additional work in gaps and in overlapping categories is needed to improve the field and confidence in results. The most pressing gap includes an assessment of NEBs related to peak and demand, not just energy. This is especially important for several categories of utility-perspective NEBs including avoided capacity / deferred construction (and possibly power quality) and line losses. Other gaps, some of which may be addressed in on-going work on a statewide project in California, include:
    - Utility perspective: updates to address kW and peak/off-peak NEB impacts; line losses; health and safety; and capacity building/ deferral values.
    - Societal perspective: health and safety; tax credit considerations; national security; and neighborhood preservation.
    - Participant perspective: non-energy operating costs; financial computations for maintenance and lifetime effects; fires / safety methodology; mobility, hardship / family stability, and others.

Persistence and Measure Lifetimes

Measure lifetimes are another critical element in the computation and attribution of savings to programs – computations that are important in credibly assessing remaining energy generation needs, as well as rewards and incentives for providers of programs. The measure lifetime analysis literature and methodology is fairly robust. More than 100 studies have been conducted, examining *in-situ* median lifetimes for residential and non-residential measures. This chapter reviewed the literature and status of work on measure lifetimes and provided information on a number of key topics in persistence. The research found the following:

- **Problems and best practice suggestions for effective useful life (EUL) studies:** Our study addressed some of the key issues that have hampered EUL studies in the past. Of particular note are the following: the need to assure that implementation databases are better structured to support evaluation research; use of appropriate

sampling approaches when bundled programs are implemented; use of phone data collection only when measures are unique or memorable; use of panel surveys if possible; more enhanced modeling that supports the incorporation of tests of multiple model specifications; and, most importantly, benchmarking of the results against the findings for earlier years of the program and for similar programs around the nation.

- **Results and gaps in EULs**: A review of results from measure-based EUL studies around North America showed that measure lifetimes exist and are fairly consistent for many measure-based programs in commercial, residential, and industrial (?) sectors. Relatively similar EUL values are being assigned by utilities across the country – perhaps with not enough recognition of the variation in operational hours by climate zone. The review also shows a lack of depth in studies in process equipment; some shell measures; and specific end-uses like cooking, refrigeration, and air compressors.

- **Technical degradation**: The issue of technical degradation was discussed, and there is a shortage of primary research on this topic. Certainly, engineering-type studies can help to identify research priorities to some extent, noting which technologies have undergone engineering, mechanical, or process changes that will more likely significantly change their performance relative to standard equipment. However, equipment with significant changes in behavioral (operational or upkeep) elements may also see changes in performance. Priority-setting for new research on this topic should take both factors into account (mechanical and behavioral), and resulting figures should be verified periodically.

- **RUL issues**: Regarding the topic of remaining useful lifetimes (RULs), some utilities argue RULs are critical to certain programs; others don't feel the estimation complexity is a worthwhile expenditure. The jury is still out on the policies to be applied broadly, but if a program is designed as early replacement, a credible case could be made that its savings pattern is significantly altered from end-of-lifetime programs. Perhaps in the short run, presenting benefit-cost figures including and excluding the enhanced savings could be presented to identify whether the programs are moving decisions forward enough to make a difference. There are potentially cases in which this analysis would also be applied to behavioral programs.

- **Retention of behavioral changes results and needs**: Of particular note is the virtual absence of studies addressing retention or persistence of education / outreach / behavioral programs. This is an important gap, as behavioral and market-based programs have become a larger and larger share of utility / agency portfolios. Further research in best practices for the array of behavioral programs or "types" would be a useful addition to the literature, and agencies should consider requiring new behavioral programs to conduct retention assessments every year or two for a period reaching on the order of three or more years out. This may be the only way to gain enough information to develop credible estimates of the persistence of savings from behavioral programs and to allow more serious consideration of them as reliable resource substitutes. The issue of retention of behaviors and savings for "upstream" education and training programs is particularly troublesome, and, to the degree that these programs are part of portfolios, retention work is needed where there currently is none. Finally, EUL measurement approaches will need to be tested and applied to a variety of behavioral programs. Some may parallel traditional EUL estimation best practices, but

the application of statistical approaches to some programs may be challenging. This research should be a priority for the near term.

Measure lifetimes are a key element in the computation of program savings. It is important to assure that new programs are developed – including creative programs and programs that encourage new measures and behaviors and are not the "same old same old". However, if measure lifetimes, technical degradation factors, and other factors are known for some programs and unknown up front for others, there will be a bias away from developing new (more uncertain) programs. Risk is an issue affecting investment and development.

Risk needs to be considered from two perspectives – providing up-front information on computational elements encourages program development. "True-up" is needed for credibility and reliability of savings estimates for EE relative to generation capacity. One suggestion may be that new programs are assigned a deemed lifetime by general "type" up front, and then after 1-2 years, a true-up is prepared that does not readjust program incentives retroactively, but does refine the estimate of future savings from a resource perspective.

Identifying the lifetimes or EULs of behavioral or information programs is complicated as more media messages on behaviors and education bleed across territories. This affects retention of the messages and behaviors because behaviors originally attributable to the program may be "refreshed" from other sources. It may not be possible to separate these out cleanly; research is required to determine the extent of this problem. The priority depends on the ranking of estimated savings and costs from these programs. In addition, results on measure lifetimes, and any remaining useful lifetime (RUL) and technical degradation factor (TDF) research should be accumulated in a database and updated continuously so comparisons and tracking are facilitated.

Conclusions and Recommendations

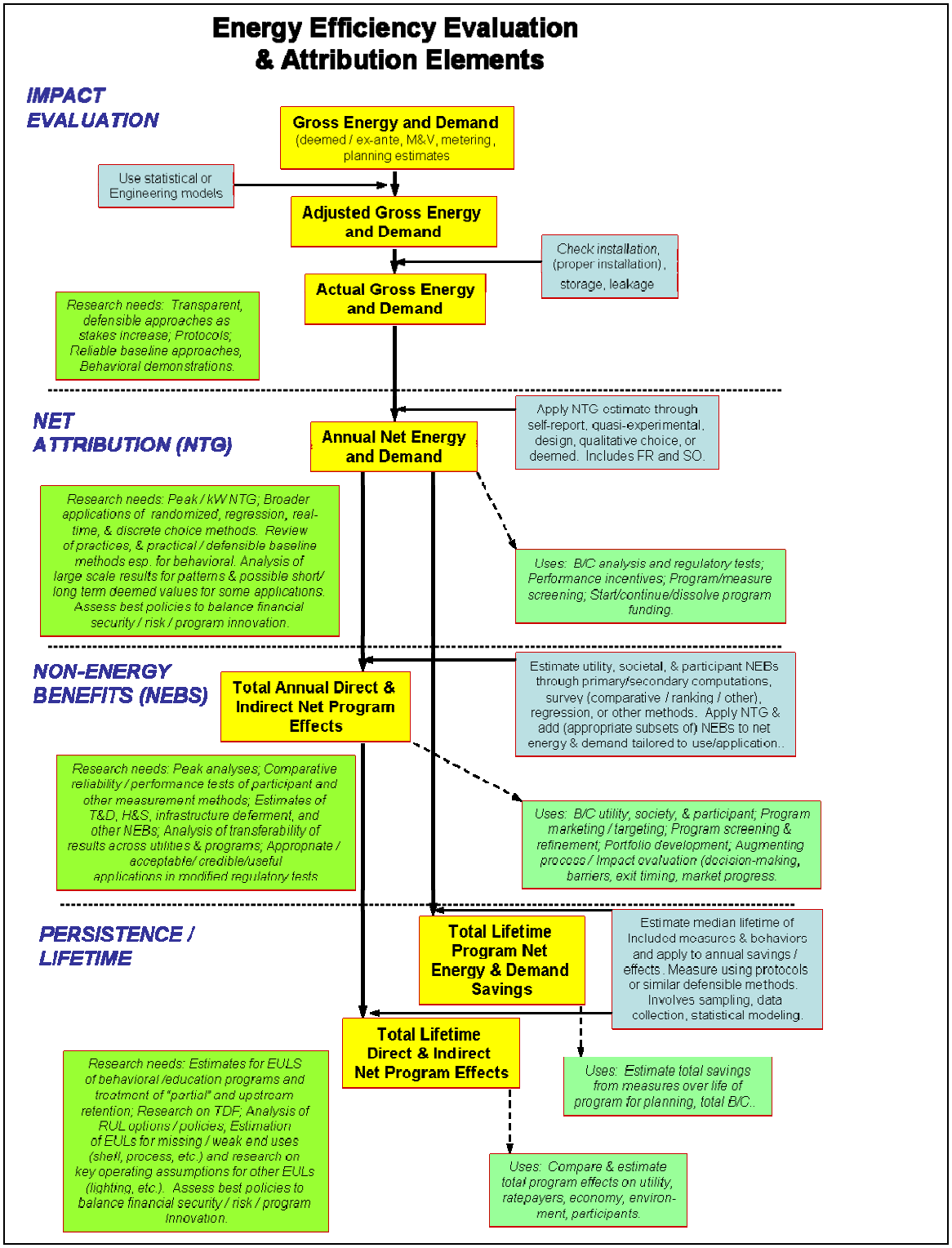
New program generations have complicated evaluation. Education, outreach, training, and market-based approaches make it harder to count "widgets" and assign savings for energy efficiency programs. New and multiple actors providing programs and outreach within utility territories increases the influence "chatter" and make it harder to isolate the impacts associated with one agency's program, or even the influence of one vs. another program from one utility or entity. These important evaluation complexities have become harder to ignore.

Some have argued that traditional evaluation approaches are failing and not worth conducting. Others have proposed modifications and patches. It may be the case that varying and evolving programs may not be suited to "one size fits all evaluation protocols" and need tailored evaluations, but, to paraphrase, not measuring is not the best answer. The best programs will not be identified – or valued and taken seriously by system planners and regulators – unless they are measured and verified.

A review of the state of evaluation in these areas – gross and attributable net savings, and non-energy benefits – suggests some lessons are old lessons (up-front evaluation design and random assignment may seem difficult, but there is no reliable "after the fact" substitute). Some are new possibilities (for example, reflecting market share through price decomposition, revisions to the regulatory tests to incorporate NEBs). Some concessions to chatter and overlaps may be needed (portfolio-level decision-making or scenarios may be an appropriate

evolution). There needs to be more up-front market assessment and baseline attention (saturation studies, perhaps augmented with behavioral aspects) to support evaluation of effects at least at the portfolio level. In some cases, deemed estimates associated with template program types may be appropriate if they are updated based on periodic measurement. Most importantly, evaluations need to continue and to loop back to program design to assure that the public dollars are being well-spent and “wrong” program decisions are avoided.

Figure 0.2: Efficiency Evaluation Elements Overview, Uses, and Research Needs



1. BACKGROUND / PROJECT SCOPE / DEFINITIONS / GOALS

This white paper is concerned with identifying current and improved techniques – and associated policy issues – related to the following:

- **Gross effects**: Measuring the broad array of impacts caused or potentially caused by program interventions – either measure-based, market-based, education, or other interventions. This includes the measurement of gross energy savings and non-energy impacts.
- **Net effects attribution**: Identifying the share of those effects – direct and indirect – that can be attributed to the influence of the interventions undertaken – above and beyond what would have occurred without the intervention – either naturally or due to the sway of other market influences or trends.

Using the current terminology, this boils down to examining four key topics in evaluation, with a focus on how they relate to the evaluation of "behavioral" programs. These topics include the following:

1. Impact evaluation
2. Attribution / free ridership / net to gross
3. Net non-energy benefits (NEBs) / non-energy impacts
4. Persistence of savings

The data and outputs from these four evaluation issues are used for an array of applications, including the following:

- Measuring progress in the market – most often using share of sales / installation of EE equipment compared to sales / installation of standard equipment;
- Benefit-cost analysis for programs, generally using standardized regulatory tests;
- Attributing savings, and shareholder benefits, to entities investing in (specific) programs – applying gross impact evaluation values modified by net to gross attribution ratios consisting of free ridership and some share of spillover;
- Comparing savings from EE to market needs and supply sources to assure energy demand needs are met – reviewing the cost per unit for EE vs. new supply, and the size and reliability of the kWh or therms;
- Program decision-making, marketing, and program design – using results of process, impact, free ridership, and other program evaluation elements to improve and optimize the program offering; and

In each case, there are significant investment dollars at risk / associated; hence, the need to revisit methods and approaches.

These white papers are needed because two primary factors have complicated the methodologies that have been applied to this field:

- The evolution of energy efficiency programs from "widget" or measure-based programs (direct install, etc.) toward an increasing focus on programs based on outreach, education, training, and efforts to change behaviors. Transition to more non-measure-based programs (education, advertising) has made it especially hard to "count" impacts.

- The increasing number of actors delivering these programs – leading to market "chatter" and increasing difficulty in identifying which among all the deliverers of the EE "message" are responsible for the change in energy efficiency behaviors, actions, or purchases. The increased chatter in the marketplace allows a situation in which consumers may be influenced by any number of utility programs by the host / territorial utility (the "portfolio") as well as influences from outside the territorial utility (national, neighboring programs, movies / media, etc.). Attributing or assigning responsibility for changed behaviors, adoption of EE measures or similar effects is muddied.

Separating out program influences has become more and more complex.

1.1 Scope and Conventions of the Paper

This paper is intended to provide a review of the state of the literature on the four topics, identifying gaps, and where possible, suggesting possible strategies for addressing the gaps. The paper notes cases in which literature is absent, but it is beyond the scope of this paper to develop significant new research. The intended audience is evaluators in the energy efficiency field.

We include several key conventions and interpretations associated with this paper below:

- Behavior: We interpret the term "behavior" and behavioral programs broadly. We interpret these efforts to incorporate behavior in terms of the participation decision, equipment acquisition, and use / operation / and maintenance of the equipment, as well as associated energy behaviors.
- Consumer: We assume that the broad range of consumers is relevant to the project, including residential and non-residential (commercial, industrial, institutional, and R&D / renewable) actors under a wide range of programs.
- Energy consumption: We interpret energy to include both electricity and gas, and we assume that measurement of both energy and demand is relevant.
- Interventions / programs: We interpret our task to cover the range of interventions, including measure and broad-based EE interventions, education, advertising, and other approaches for influencing behaviors and purchases. We use EE interventions in most of the paper, but use "programs" and other terms as well.

1.2 Purpose of Evaluation

There is considerable debate over precision – or lack (presumed) thereof – in association with a number of specific aspects of evaluation research. For example, many criticize the accuracy of free ridership or net to gross ratios, or deride the estimates of non-energy impacts.

The 2003 Nobel-award winning economist, W.J. Granger, summarized the overall purpose of evaluation as '…***research designed to help avoid making wrong decisions (about programs)***'.[4]

Based on the background for this project, perhaps the three most important potential "wrong decisions" might relate to:

[4] Luncheon speech, Western Economics Association Meetings, Denver, CO, 2004. Parenthetical clarification added by author.

1) assuring public dollars are being responsibly spent;
2) apportioning dollars and efforts between alternative strategies; and
3) helping identify the appropriate time for exit strategies (or program revisions).

Perhaps this overriding principle is worth keeping in mind as we consider our standards for evaluation in energy efficiency. If this principle is accepted (at least for some applications), then it becomes clear that the level of accuracy applied to evaluation research can be flexible, based on the value (cost) of the possibility of a wrong decision coming out of the particular advisory research.[5] Identifying the cost of a "yes/no" decision about going ahead with a program or intervention may allow a much less accurate estimate for input information than a decision about the precise level of shareholder dollars that should be allowed for a particular agency, should that be a desired outcome to be supported by the evaluation exercise.

The CIEE project considered several other issues. Given the current regulatory / utility / program environment (especially in California), there are significant investment dollars at risk; hence, the need to revisit methods and approaches. Further, there is a need to look at the following concerns:

- What are the goals and applications – what are we trying to measure? General progress in EE or specific progress? Is it necessary to provide measure- or program-based metrics? Are measures of market progress sufficient? Can we measure the array of impacts that are appropriate?

- What degree of effort is appropriate for measuring these effects? Does it need to be at the program basis, measure basis, etc., or are broad market measurements sufficient? How do we assure the best measurement methods are applied? Can we revisit methodologies, learn / adapt from other fields, and bring the best defensible – yet practical – methods to bear in the field?

- If detailed measurements are not feasible or deemed unnecessary, can our new approaches:
  1) assure public dollars are being responsibly spent?
  2) apportion dollars and efforts between alternative strategies (if there aren't program-based metrics); and
  3) help us identify the appropriate time for exit strategies (or program revisions)?

These are some of the issues in the background (and foreground) as the issue of measurement and evaluation in energy efficiency is reviewed.

1.3 Research Approach and Sources

This paper represents the results of outreach to more than 100 researchers in the energy evaluation and related fields,[6] detailed interviews and/or survey responses with dozens of professionals in the field, and review of more than 100 papers and reports representing research in the four key topics covered by this paper. Although the topics certainly warrant even more work, budget constraints limited the scope and outreach for the paper. The work does, however, attempt to identify the state of the art and its strengths / weaknesses, potential

[5] It may also suggest that the accuracy on one "tail" of the research may be different than the other tail.
[6] Culled from lists of attendees at major conferences, researchers in the field, and literature searches.

improvements and how they relate to behavioral issues, and recommendations on next steps and next research directions.

1.4 Background and Organization of the Paper

According to the U.S. Environmental Protection Agency's (EPA) "Model Energy Efficiency Program Impact Evaluation Guide" (EPA 2007), the basic steps in conducting an impact evaluation include: "…

- *Setting the evaluation objectives in the context of the program policy objectives.*
- *Selecting an evaluation approach and preparing a program evaluation plan that takes into account the critical evaluation issues.*
- *Implementing the evaluation and determining program impacts, such as energy and demand savings and avoided emissions.*
- *Reporting the evaluation results and, as appropriate, working with program administrators to implement recommendations for current or future program improvements."*[7]

The EPA report also suggests that the three impact evaluation results that are typically reported include: [8]

- *Gross savings (energy or demand) estimate*: The change in energy consumption or demand resulting from …"program-promoted actions (e.g., installing energy-efficient lighting) taken by program participants regardless of the extent or nature of program influence on their actions."
- *Net savings estimate*: The portion of gross savings attributable to the program, extracting the impacts due to other influencers (internal or external).
- *Non-energy benefits (NEB) estimate*: Positive or negative (non-energy) effects from the program's intervention, including examples like "…comfort and productivity improvements, job creation, and increased maintenance costs due to unfamiliarity with new energy-efficient equipment."[9] EPA also particularly cites the NEB of avoided air emissions (from central generation or site-based) as one important reported NEB in impact evaluations.

"Estimates" is a deliberate word in these bullets, as these effects can likely never be measured directly, and EPA notes that

> *"…evaluation results, like any estimate, should be reported as "expected values" with an associated level of uncertainty. Minimizing uncertainty and balancing evaluation costs with the value of the evaluation information are at the heart of the evaluation process."*

In a sense, following the EPA outline, Chapter 2 of this document addresses the estimation of gross savings, Chapter 3 further details issues related to sorting out the "net" attributable impacts from the gross savings, and Chapter 4 discusses NEBs. Chapter 5 covers an additional topic – the persistence or retention of the savings (measure lifetimes).

[7] EPA 2007, "Model Energy Efficiency Program Impact Evaluation Guide Section ES
[8] EPA 2007, "Model Energy Efficiency Program Impact Evaluation Guide Section ES
[9] Op. cit.

2. MEASUREMENT OF GROSS IMPACTS

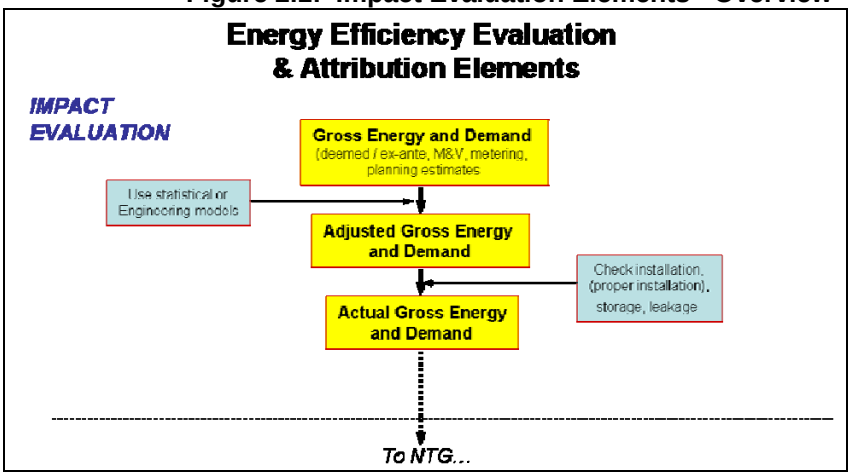
In this section, we discuss issues related to the estimation of gross impacts from program interventions—that is, the share of impact specifically attributable to the intervention – beyond what would have happened without the intervention. The refinements necessary for estimating "causation" or the "net" share of the impact due to the program are addressed in Chapter 3; however, because some topics are not easily divided or always sequential, some of the "attribution" issue is also addressed in this chapter.

2.1 Current Practices and Uses

Measuring and evaluating gross energy savings means determining the impacts—energy savings, demand savings, or both—that directly result from program-promoted actions. In other words, these are the impacts of energy conservation measures (ECMs) promoted by the program that are installed by those directly participating in the program, regardless of the extent or nature of the program's influence on their actions.

Impact Evaluations

Figure 2.1: Impact Evaluation Elements - Overview



Impact evaluations apply at least one of the following five general methods: the first three are discussed in the 2004 California Evaluation Framework (TecMarketWorks, 2004) and apply to traditional rebate or incentives programs. The last two methods may be thought to overlap the first three methods somewhat, but they are most often applied to market transformation and education or social marketing programs.

1. Measurement and Verification (M&V): using metering or estimating key parameters from a sample of participants and applying it to all participants.
2. Deemed Savings: applying "deemed" or agreed-upon savings obtained from other evaluations or manufacturers' data to all program participants.
3. Statistical Analyses: applying statistical regression models to utility billing or metering data of all program participants
4. Market Progress / Market Share: using information from sales, shipments, or other similar data to develop estimates of changes in sales (and implied usage) of program-recognized energy-efficient equipment relative to non-program equipment and estimates the associated energy (and/or demand) savings.
5. Surveys and Self-Reporting: surveying certain populations to gather information regarding knowledge or behavior to estimate the savings-related changes from

behavioral / educational / social marketing programs, perhaps in concert with market progress methods described above. While there can be difficulties linking to direct savings (and some evaluators simply don't try to count or evaluate these programs), experimental design, which includes random assignment to test and control groups of adequate size, can provide estimates.

In some cases these approaches are combined, particularly the deemed savings and M&V approaches (EPA 2007).[10] Each method, which is described in more detail below, varies with the different approaches used in implementation

Impact Method 1: Measurement and Verification (M&V)

The 2002 International Performance Measurement and Verification Protocol (IPMVP) describes the M&V methodology (EVO, 2002). To use this method, one first selects a representative sample of projects within a program, then determines the savings from the selected projects, and then applies this information to the entire population of projects.

M&V covers all field activities dedicated to collecting site information, including: equipment counts, observations of field conditions, building occupant or operator interviews, parameter measurements, and metering and monitoring.

Individual project savings are determined using one or more of the following approaches:

- Engineering calculations with estimated or metered parameters
- Isolated ECM metering
- Whole facility metering
- Calibrated simulation facility modeling

Each of these approaches is discussed briefly below.

- ***Engineering calculations*** may be performed using data from the ECM system's metered or estimated key performance parameters, or from parameters deemed significant for a project's success. For example, a measurement of power draws for a sample of light fixtures may be obtained from inputs to a building study where a large number of light fixtures have been replaced and operating hours remain constant. Similarly, metering an appliance's run time may provide input for an efficient appliance upgrade study. Equipment with constant loads and usage patterns (e.g., automatic outdoor lighting or constant flow industrial motors) only needs short-term metering, while equipment with varying loads (e.g., heating or cooling equipment and variable motors) may require continuous metering. This approach works well for examining simple heating or cooling equipment replacement, lighting retrofits without heating or cooling interactions, and industrial motors with constant power consumption.
- ***Isolated ECM metering*** isolates the ECM system and then measures hourly energy consumption during baseline and reporting periods. Spot or short-term measurements may be sufficient to measure the baseline condition, but the goal is to calculate savings based only on metered results without including stipulations or estimations of major factors. Of the four approaches, this is generally the most expensive, as it may require equipment rewiring plus continuous data collection and analysis for, ideally, a full year

[10] We separately describe "market share"-based work, which some might consider a combined-type approach; however, others would consider it quite separate and not overlapping.

before and a year (or more) after installation. It applies well to ECMs that can be easily isolated, such as replacing a chiller or boiler.

- ***Metering the whole facility*** relies on a macro-analysis and assumes changes in energy use over time can be explained by the change resulting from the ECM, factoring in adjustments for changing conditions (such as weather or operating hours) to make periods comparable. The meters used might be the same as those employed in utility billing, although other meters can also be used. This approach works best for evaluating savings from whole building retrofits. Metered results may then be evaluated using either a simple bill comparison (if the building is not affected by changing weather or other factors, such as occupancy) or multivariate regression analysis. The approach only works well with savings sufficiently large to be distinguished from random or unexplained energy variations normally found through whole-facility metering. In addition, longer periods of "before" and "after" use are needed to reduce the impacts of short-term, unexplained variations. Typically, savings should be more than 10 percent of the baseline energy use, so that they can be separated from the "noise" in baseline data.
- ***Calibrated simulation facility modeling*** involves the computer simulation of the facility's energy consumption based on key building characteristics, such as size, thermal envelope characteristics, energy-consuming equipment, weather, and occupancy data. This approach attempts to reconcile results with utility hourly or monthly utility billing data. Manufacturers' data, spot measurements, or short-term measurements may be collected to characterize both baseline and reporting period conditions and operating schedules. Whole-building models typically require 9 to 12 months of data for proper calibration. Often, this approach is chosen for new construction projects because of the lack of baseline data (other than "standard practice" knowledge or building code requirements). As with metering conducted for a whole facility, savings must be greater than the associated simulation modeling error. Calibrated simulation facility modeling also works well with a high degree of interaction among installed energy systems and when metering individual components proves too difficult or costly.

Impact Method 2: Deemed Savings

Deemed or stipulated savings estimates are used when a project has well-known and documented savings values, and when savings estimates are relatively small. The "deemed" values are based on reliable, traceable, and documented information sources, such as manufacturers' specifications, engineering calculations, or a previous evaluation result. For example, deemed savings could be used for refrigerator replacement, which may be small relative to whole-house energy, as the refrigerator would be consistently operated. This method may also be used when a budget is small and accuracy is not critical. Deemed savings are often estimates based on past evaluations or on building simulation modeling.

A popular source for deemed savings estimates is the Database for Energy-efficient Resources (DEER), which contains information on many energy-efficient technologies and measures (www.deeresources.com). While developed for California climate zones, information on measure costs, measure lifetimes, and non-weather dependent energy savings may be applicable across the country.

The California Energy Commission developed the first version of the database in the early 1980's expressly to compile energy savings and incremental cost data on common energy-efficiency measures using multiple information sources such as a utility's demand-side management (DSM) program filings. The database contained estimated average costs, market

saturation, expected life, annual energy savings, and summer on-peak demand reduction estimates of common DSM measures for both residential and nonresidential applications. Originally, the database was intended for use by DSM planners to estimate measure and program cost-effectiveness for regulatory filings. Another intended use was forecasting DSM program demand reduction and energy savings potential in specific market segments and utility service territories. As updates were performed over time, the information contained in the DEER database on selected baseline and energy-efficient technology applications became extensive.

Consequently, the DEER database has evolved into a source of common savings values (such as deemed savings and both full and incremental measure cost data to improve the consistency of information and assumptions used in energy-efficiency analyses). Since its inception, the DEER has undergone four major updates: the 1994 NEOS Study, the 2001 Xenergy Study, the 2004-05 DEER Update Study and, most recently, the 2006-2008 DEER Update Study finalized in July 2008. Currently, the database contains information on over 250 energy-efficiency measures. Other individual utilities or states may have their own versions of the DEER database, which often use information from DEER and incorporates modifications to reflect varying climate zones and typical measures.

Impact Method 3: Statistical Analyses

Statistically analyzing large sums of data involves using whole-facility utility metering and applying a variety of statistical methods to analyze the resulting data. Depending on the program size, a census or a sample of participants may be used (such as that described in the measurement and verification section, above). Three approaches can be employed for measuring savings through statistical analysis:

- Comparison group or “difference of differences”
- Time series comparison or “billing analysis”
- Combination time series/comparison group

Each of these approaches is described briefly below.

- ***The comparison group (or “difference of differences”)*** approach compares program participants’ energy use after projects are installed with non-participants’ energy use. For example, if an evaluation of a utility’s residential weatherization program showed participants saved 2400 kWh, but nonparticipants also saved energy (let’s say 600 kWh), the difference of these differences was 1 800 kWh, which was the “net” program savings attributable to the program. This approach was commonly used in early evaluations (late 1980’s and early 1990s), but it has since been superseded by other regression-based methods. While, it may be useful for new construction programs where no baseline data exist, the major challenge is finding a comparison group, particularly in areas with a long history of program offerings. Although the approach was accepted as a default by the California Public Utilities Commission in the 1992 measurement protocols, it is no longer accepted in the new (2006) version of the protocols. Ensuring comparison groups are indeed comparable can be difficult since the two groups must be sufficiently similar so the only difference between them is their participation status.
- ***The time series comparison (or billing analysis)*** compares program participants’ energy use before and after project installation. This approach is a way to control for the effects of weather and to produce a statistic that is basically kWh per heating (or cooling) degree day (HDD or CDD). Thus, evaluators can group buildings from different weather

zones to examine other patterns, such as behavioral effects, demographics, etc. Once the weather effects are taken into account for a specific year or time period, evaluators can calculate normalized annual consumption (NAC), which is simply the kWh/HDD number multiplied by the average annual number of heating degree days over, say, the past 30 years. Weather normalization does not apply to non-weather sensitive buildings or extremely complex buildings.

- ***The combination time series/comparison group*** combines the previous two approaches through an analysis that compares post-project consumption to pre-project consumption and compares a participant group to a non-participant group to account for non-project-related changes in energy consumption.

For any of these methods, statistical analysis defines a relationship between a dependent variable and one or more independent variables. The dependent variable (or output) is energy or demand consumption and savings. Although many different models may be devised, the challenge lies in finding the one that best fits the data. Model examples include:

- Normalized annual consumption (NAC)
- Conditional savings analysis (CSA)
- Statistically adjusted engineering models (SAE)
- Analysis of covariance models (ANCOVA)

Each of these approaches is summarized briefly below.

- ***Normalized annual consumption (NAC)*** analyzes monthly energy consumption data by applying statistical analysis software, such as the Princeton Scorekeeping Method (PRISM), SAS, or SPSS. It is most applicable to whole-house retrofit programs, as it does not compute results by individual ECM.
- ***Conditional savings analysis (CSA)*** models the change in consumption by using a regression analysis against the presence or absence of energy-efficiency measures. A value of 1 is assigned if the ECM is installed; a value of 0 is assigned if no ECMs are installed.
- ***Statistically adjusted engineering models (SAE)*** incorporate engineering estimates of savings as dependent variables. For example, a SAE model can use the change in energy as the dependent variable in a regression model against estimated savings for installed ECMs. These estimates may be provided in the design phase or through secondary sources. A value for estimated savings is assigned to the variable if the ECM is installed; a value is 0 assigned if the ECM is not installed.
- ***Analysis of covariance (ANCOVA)*** models are also called “fixed effects” models. This approach allows each participant or non-participant to have a separate estimate of the intercept term. The intercept term represents the base component, which accounts for the individuality of participants. The fixed effect approach also can be used with any of the other models previously described.

Variations in methodology (ranging from model structure to the variable types and content) can be almost infinite in size. Additionally, specific modeling decisions can have a major impact on the results of the analysis. Clear and objective standards for identifying the best model are needed to avoid choosing the model that fits the desired results rather than the model that best

fits the data available. Several approaches for using objective standards to identify the best model exist; however, presentation and discussion are beyond the scope of this paper. (See Parlin 2007, Burnham and Anderson 2001, Kmeta 1980, and McQuarrie and Tsai 1998 for approaches to assessing model fit.)

The primary objective for model selection should be to find the simplest model that adequately fits the data. Models with too few variables may produce biased estimators, whereas models with too many variables can lead to a limited precision of the estimators. While using objective approaches for model selection avoids results-based decisions, no method completely substitutes for the judgment and experience of the analyst. Strict adherence to any set of rules can easily produce results counter to basic common sense and, therefore, all results should be compared to other research in the field and the knowledge of the analyst.

The model selection process involves four steps: (1) defining the candidate models; (2) using diagnostics to assess the appropriateness of the fit; (3) running the models; and (4) comparing and evaluating the results. The first step, which may be the most difficult, is to develop a number of appropriate candidate models given the available data. While it's possible to have numerous candidate models, it may be necessary to make informed judgments to limit the candidate models to a reasonable number, starting with a wide variation in model fits (such as those with or without weather-dependent effects or different error structures). Once certain approaches have been eliminated, fine-tuning can take place.

Once a number of candidate models have been prepared, the analyst can work with variables to find the best-fitting model of that type and then compare and evaluate the results. This is current standard best practice.

Note, however, that there was concern expressed by some researchers that impact evaluations have not become more reliable over time, that statewide evaluations lose too much information, and that analysts need to use care in application of modeling approaches. They argued that impact evaluations have become more and more focused on deemed values, or else on modeling exercises with little locality-based direct measurement instead of focusing on direct measurement combined with some modeling (a preferred approach). (Peach 2009, Blasnik 2009, Gordon 2008). Multiple interviewees noted the need for more specific program-related measuring, metering, and logging data. Others noted that user expertise is a key (and variable) factor (Ogle 2009, Blasnik 2009), and to test this hypothesis, Energy Trust of Oregon is contracting with multiple billing analysis "experts" to analyze the same datasets to compare and contrast results (Gordon 2009).

Impact Method 4: Sales / Market Share

By intervening at a variety of levels (manufacture, distribution, specification, purchase, etc.), market transformation (MT) programs facilitate the adoption of energy-efficiency practices and equipment beyond what would have occurred without the programs. MT programs use many sources – primary and secondary – to gather the data for estimating market shares.

Most MT evaluation projects rely on some variation of the traditional comparison of market shares to measure progress or goals achieved. They tend to use either pre/post program intervention or a comparison (control) group from other states or regions. These comparison groups are selected to be similar (preferably in ALL ways) except for lack of access to the

specific program being evaluated. [11] This comparison group method is one of the most basic methods used to measure impacts from ENERGY STAR™ appliances and lighting measures.

Discussion / Pros and Cons:

Primary data collection can be used to support the MT evaluation; alternatively, depending on the energy-efficient equipment promoted by the program, there are numerous potential sources of secondary data. Primary data is valuable because it can be collected specific to the (utility or other) territory under consideration; however, it is also quite costly to collect. The availability of secondary data sources for key ENERGY STAR™ appliances is summarized in the following table.

Table 2.1: Availability of Data from Sources, by Product Type
(source: adapted from Dimetrosky et al., AESP White Paper, 3/07)

	National Retailer Partner Sales Data	National Manufacturer Partner Shipment Data	Manuf. Shipment Data through Industry Assoc.	Point of Sale Data	Market Studies	18 Seconds . org	Canada (CAMA) (comparison only)
Refrigerators	A	NA	A?	A	NA		A
Clothes Washers	A	NA	A?	A	NA		A
Dishwashers	A	NA	A?	A	NA		A
Room AC	A	NA	A?	A	NA		A
Central AC	NA	A	A	NA	NA		A
Lighting Fixtures	NA	A	NA	NA	NA		A
CFLs	NA	A?	A	A	NA	A	A
Windows	NA	NA	NA	NA	NA		NA
Key: A=available, NA=not available; ?=collected, but not readily available							

In addition, a paper by Dimetrosky et al. (2007) summarizes the applicability issues associated with the specific data sources, summarized in Table 2.2.

Table 2.2: Data Sources and Applicability Issues, updated

Data Source	Applicability Issues
National Retailer Partner Sales Data	Data only account for do-it-yourself retailers Reporting retailers may change year to year Delays in receipt of data Data only available for four product types Data only report ENERGY STAR™ vs. Non-ENERGY STAR™ sales rather than tiers of efficiency levels
National Manufacturer Partner Shipment Data	Significant potential for non-response bias due to lack of enforcement of shipping data requirements Data reported as "shares" without sales figures Data only provided nationally, no regional numbers Manufacturer shipment data not readily available, limiting cross section comparisons
Manufacturer Shipment Data from industry organizations (like Association of Home Appliance Manufacturers – AHAM), NEMA (for lighting products), ARI (for air conditioners)	Shipment data from 12 product categories – refrigerators, dishwashers, clothes washers, room air conditioners, freezers, and cooking equipment Shipment data is available by state for comparison and baseline proxies Most data only available to the manufacturer members that provide the data Geographic breakdowns can err as shipments recorded to regional distributors which may reship to other geographic areas
18seconds.org (a cross sector network championing CFLs)	Point of sale data reported by major retailers and aggregated to the Metropolitan Statistical Area

[11] Propensity scoring is an accepted method in the academic literature for correcting control group issues.

Data Source	Applicability Issues
	Omits data from do-it-yourself stores
Canadian Appliance Manufacturers Association (CAMA)	Members can download shipment data and market penetration reports to report by different efficiency levels Data are from self-reported estimates from the telephone, not real time sales data
Point of sale data from scanner data from ACNielsen and Activant	Missing data from do-it-yourself stores Are reported at broad regional levels Are recorded by model number, not efficiency level
Market studies from third parties	Include useful longitudinal information, but do not specifically pursue energy efficiency issues
Primary sales data collected from retailers	Valuable for program tracking but often lack data from ENERGY STAR™ partners that report sales to EPA
Survey Data	Cost effective. Rely on self-reported data, which is better suited to upstream actors than end-use consumers. Appropriateness varies by type of measure
Price change as a proxy for market penetration changes	This approach is based on basic economic theory: if price margin attributable to energy efficiency features go down, it reflects increases in relative sales of energy efficient models (Skumatz, 2006). Price differentials may be easier and less expensive to obtain than sales differentials. There may be too few applications to date to assess the performance record.
Changes in home saturation levels	For example, in Massachusetts, onsite studies have estimated "socket saturation"—the percentage of sockets that have CFLs in a sample of households (Nexus Market Research, 2005). If these studies were done over time, the researcher could assume that changes from one year to the next would approximate changes in sales.
Tax records	When tax credits are offered for different tiers as, for example, the State of Oregon offering varying credits on energy-efficient clothes washers, tax return data may be harvested. The delay on obtaining these data is over one year, as it is unavailable until after taxes are filed. Possible biases from underreporting or misreporting.

Dimetrosky (2007) reviews various sources of sales and direct market saturation data, and notes strengths and also significant weaknesses associated with currently-available options. Improved reporting of data in these secondary sources should be enforced to better support evaluation work; he also suggests it would also be helpful to expand the equipment types included in data collection and make the efforts to collect data at the state level rather than national level.[12]

These partner and other secondary data sources may help in evaluating residential measures, which are often key parts of residential interventions (including behavioral interventions). However, they are much less useful – at least to date – in measuring progress for commercial and industrial measures and the effects of behavioral and other interventions.

A similar contemporary analysis of the available secondary sources also noted that there were significant problems with virtually all the traditional sales and shipment data sources – but interpreted the situation with a somewhat more pessimistic lens (Skumatz 2005, 2007). Variations in the individual businesses / sites submitting data lead to shifts in market shares, confounding the evaluator's ability to attribute changes to programs. First-hand analysis had indicated that using data from different (but "similar") sources could lead to divergent results. Of greatest concern was that, long term, the cost and reliability issues might never improve because the dealers / manufacturers / retailers did not have business interests in reporting data. Barring much more aggressive enforcement of participation in reporting (and some important businesses would never report), the track record in California and elsewhere indicated the sales / market share data collection would remain expensive (and labor intensive).

Looking for a creative alternative that would not be hamstrung by data issues, the paper argued that basic economics posits a relationship between price and quantity, and that unlike quantity data, price data in many of these markets were readily available (especially for residential equipment). In a series of papers, the concept of reflecting market share improvements through prices was explored and demonstrated. The argument was that, using statistical hedonic pricing models, if the price premium for the "energy efficiency feature" of the appliance decreased, the quantity of the efficient appliance purchased (and consequently the market share) would tend to increase. Comparisons over time or between program and control regions could be used to attribute effects to programs. The approach turned out to provide robust information beyond the original intent, providing information useful for setting product rebate levels (the price differential at which consumers were indifferent between efficient and less efficient models); and information indicating when markets were "mature" or when programs should exit the market. Several papers and reports starting in 1998 illustrated promising findings for about a dozen ENERGY STAR® appliances and measures; however, the approach is new and needs additional demonstration.

Impact Method 5: Surveys

Surveys are employed to gather information for estimated parameters to be used in engineering calculations similar to an M&V approach (Impact Method 1). Surveys are also used to gather information for the Market Progress/Market Share approach (Impact Method 4). Surveys alone have been used to estimate savings-related changes from behavioral / educational / social marketing programs. While there can be difficulties due to inaccuracies in self-reported data, experimental design that includes random assignment to test and control groups of adequate size can provide useable estimates.

[12] The paper also suggests pursuing data from trade associations, and better identifying the reliability associated with the data.

2.2 Overall Findings

Some methods and approaches are better suited to particular ECMs. The most appropriate evaluation method must achieve the ultimate goal: to balance an evaluation's cost with the value of the information received. The metering of parameters or actual end-use consumption before and after the ECM generally costs more than a deemed or statistical analysis because of the time required at an individual customer's site. However, many evaluations have found inaccurate assumptions can best be identified and remedied through expert verification on a project site. Since many ECMs are unique to a particular customer, failure to conduct a site visit and obtain actual measurement and analysis particular to that site can reduce the accuracy of the results.

Variations by Types of Measures, Sectors, and Programs

The sections below describe how evaluation approaches may vary by sector and type of program.

Large Commercial and Industrial (C&I) Programs

Large C&I programs typically involve a high degree of individuality and may include one-of-a kind measures, such as upgrading refrigeration systems, changing entire manufacturing processors, or adding an energy management system with controls unique to that building. For these projects, the evaluation must include an M&V approach highly specific to the ECM installed.

Depending on the evaluation budget, different parameters may be measured or estimated but, in either case, it is important to ensure assumptions (such as operating schedules) remain as accurate as possible. For instance, a constant process load may be assumed to operate 24 hours a day, 365 days per year, but operators may fail to mention that the equipment shuts down for maintenance two weeks per year (Barbieri et al. 2007). Such incomplete information can lead to overestimating savings.

As an example of a C&I program using a combination of methods, NSTAR used engineering calculations and metered parameters (Select Energy Services 2004) to evaluate a program for commercial refrigerators. Meters monitored compressor run times, outside air temperatures, and evaporator fan times. From these data, a regression analysis was performed to simulate the baseline energy consumption, while the reporting period energy consumption was measured by logging equipment run times (which measure instantaneous power draws of the equipment). In this case, results were used to revise savings predictions for future programs.

It is important to verify accuracy of assumptions as well. In one documented evaluation in the Northwest, random verifications of operating parameters found oversimplified modeling procedures, changes in operating use, imperfect estimates of baseline operating parameters, and calculation errors that resulted in a realization rate of about 93% of the savings initially verified (Scott et al. 2005). In some circumstances (about 5% of industrial processes and 30% of commercial measures), savings were deemed due to lack of available data.

Commercial lighting retrofits may significantly affect heating and cooling requirements in commercial buildings; thus, they are best evaluated using a Statistical Analysis method. For best results (and to test various modeling approaches) it is important to gather information on independent variables (such as weather, facility size, and operating schedules). The savings

also need to be large enough (roughly 10% of the total bill) to be differentiated from unexplained energy variations.

If anticipated savings are smaller than 10%, an M&V approach may prove a better approach. Again, the evaluator needs to determine operating hours, as a number of actual evaluations have shown operators often assume lights are off at night when a significant number remain on.

Commercial motor replacements are also best evaluated using an M&V technique, unless changes represent savings greater than 10% (in which case, statistical analysis can be used).[13]

Residential and Small Commercial Programs

Residential weatherization is a program that may have interactive effects and often involves consumer behavioral changes that are difficult to measure directly. A statistical analysis method works well for these programs, or a calibrated simulation facility model may be used.

In a New Hampshire study of a low-income weatherization program, when evaluators compared results using M&V with engineering estimates and billing analysis, they saw a lower realization rate from the billing analysis method (64.5% vs. 86.9%) (Barata 2006). This study found billing analysis better accounted for behavioral changes in energy consumption, but the M&V method (with site verifications) was important for ensuring accurate assumptions regarding installations.

Residential appliance upgrade and compact fluorescent rebate programs, on an individual basis, may have savings too small to quantify using statistical analysis alone, and these measures are too small and varied to warrant the expense of an M&V approach (unless it is performed on a small sample). However, in a relatively unique example of a refrigerator evaluation, the New York Power Authority metered refrigerator usage and then applied the results to a statistical model predicting energy savings as a function of refrigerator label rating (Pratt et al. 1998). The results of the study helped them to divide the metered load into baseline load (dependent on the refrigerator rating), occupant-associated load (dependent on occupant usage), and defrosting load (dependent on whether the refrigerator had manual or automatic defrosting).

The deemed savings approach is increasingly being applied to residential programs, as deemed savings estimates have become more sophisticated. Some utilities that have used an M&V approach in a previous evaluation will apply those results to the deemed savings method for later program evaluations. For compact fluorescent lighting (CFL) programs, this method becomes less accurate over time, as CFL saturation per home increases. Some evaluations are finding that with the CFL-per-home saturation increases, hours of use for incremental lights installed also decrease (CPUC 2009).

New construction programs are unique in that baseline data do not exist. They also tend to include a number of ECMs that interact with each other, such as insulation levels, energy management systems, and efficient fluorescent lighting. An M&V approach using a calibrated simulation facility model is often used in this situation. Utility whole-facility metering results should be calibrated to the model predictions of energy consumption over a 9 to 12 month period following construction to improve accuracy. In California, for example, a residential new construction program was evaluated where the M&V approach of calibrated simulation modeling was compared to end-use metering of different loads in the home. Assuming correct metering results, the findings showed the simulation models over-predicted energy use (Bernier et al. 2007).

[13] Note that several interviewees suggested that measurement in large industrial sites was becoming compromised because of over-surveying (including Sulyma 2009).

Plug loads are a growing area of ECM opportunities as computer and entertainment loads increase in the residential class. As these loads are small relative to a whole facility, an M&V method is most appropriate. Site visits or detailed surveys, however, are required to gather all specific information on plug loads needed to perform an evaluation. One such study, performed for NYSERDA, was a baseline assessment to estimate energy consumption of plug loads in offices, college campuses and schools (Sabo 2007). Equipment surveys and interviews were used to estimate parameters and then engineering estimates applied to calculate the baseline.

Education and Outreach Programs

Program managers, utilities, and consultants have long known that regardless of whether an energy conservation measure is installed or not, the full measure potential cannot be realized without customer buy-in, participation, and knowledge. For instance, even if customers have digital thermostats installed, the thermostats will not save energy if the customers are not taught how to program the thermostats, lower the heating/cooling on hot/cold days, and change setting for home versus not home periods.

Behavioral and education programs tend not to receive energy-saving credit. They are largely viewed as "supporting" or "indirect", and many states are unwilling to attribute savings to these efforts separately – even when conservative cases are presented. They can cause energy saving practices that are either 1) ignored / fall between the cracks or 2) attributed to measures and reflected in their evaluations. In either case, the tendency produces a strong disincentive for program implementers/ utilities to invest in behavioral and education efforts, thereby foregoing energy savings that could probably be achieved cost-effectively (Bensch 2009).

An impact evaluation in California and the Northwest that examined daylighting using photosensors in office buildings yielded a disappointing realization rate when modeled using M&V calibrated simulation modeling. In many cases, phone surveys and site visits found the system had been turned off due to occupant complaints (Heschong Mahone Group 2006).

Education programs can be customized for a variety of target audiences. For low-income weatherization, customers are provided information on achieving savings by turning down thermostats, turning off lights, or lowering set points on water heaters. For commercial and industrial customers, education programs can provide building operators with information on the importance of operating and maintenance procedures or aid building owners through energy audits or building studies.

Savings from such programs can vary greatly and are dependent on whether the customer implements the recommended behavioral changes. The statistical analysis method may be appropriate in these situations, as long as expected savings are sufficiently large. This approach generally involves pre- and post- billing analysis, usually with a control group (another community) or a treatment group that didn't receive the education or program intervention. Often, however, education changes may be too small to apply a statistical analysis method.

Because education programs are often lower cost, following up with surveys to determine if and how the information has been applied may provide sufficient accuracy. In one research report for low-income weatherization, the authors used billing analysis and were able to attribute savings of 0-12% from the education portion of the programs (Drakos et al. 2007).

An analysis of several evaluations using billing analysis and engineering estimates found that impacts from education-only programs ranged from 2.5% to 12.5 % (Drakos et al. 2007). Research on education-induced savings from energy education centers and other similar approaches have also demonstrated impacts (Peters 1999).

In the commercial sector, in an impact evaluation of NEEP's education program to provide training to building operators, the evaluator used surveys requesting that program participants (operations and maintenance personnel) estimate savings themselves (RLW Analytics 2001).

One of the methods sparking most interest for behavior change is "community based social marketing" (CBSM). CBSM is gaining widespread recognition as a model behavioral change program using a framework based on traditional product marketing and sociology to change target audience behavior patterns. This strategy argues that engaging personal commitments, social interaction, pledges, and other personal responsibility elements to achieve behavioral change can be more effective than traditional broad-based, impersonal advertising. CBSM literature indicates that programs based on this approach provide greater participation and behavior change, penetration to previously unconverted participants, and greater retention of the behavioral change.

Marketing and Advertising

The energy sector has used advertising and marketing strategies to change markets and behavior. Measurement of these results is of increasing interest as energy-efficient markets mature. Skumatz summarizes approaches used to measuring effectiveness of advertising and marketing as follows (Skumatz 2000):

- Focus groups and surveys examine success at points in the decision-making process by asking about recall, intention, and actual purchases. They also attempt to track quality of advertising copy and to assess the correlation between intention, reported purchase, and advertising exposures.
- Data tracking agencies track pre-and post campaign purchase data using compilations from electronic scanners used in purchasing
- Randomly assigning special groups of communities to receive different cable feeds that allow inclusion / exclusion of ads from groups within the same community enables the comparison of purchase rates.

In an assessment of several case studies (some within and some outside of the energy-efficiency industry), Megdal (2006) recounts several approaches to evaluating the effectiveness of advertising campaigns. One conclusion from the paper is that research designs using multiple comparisons provide a greater ability to measure effects and the level of effect generated per increment of advertising than a simple pre-post survey design.

Market Transformation Programs

Market transformation programs may offer retailer incentives or widespread marketing to influence the market in promoting and carrying energy-efficient products. For example, several utilities have implemented a program in which they (1) pay retailers incentives to buy down CFL costs and (2) enter into agreements with retail outlets to advertise products directly on behalf of the utility. The idea is the market will transform and, ultimately, utility rebates will be unnecessary to maintain a high saturation for these products.

The evaluation of such programs requires a few more steps than do traditional utility rebate programs. Customers who purchase these lights at retail stores may not be easily identified as

there are no rebate forms containing customer identification. Also, customers may be purchasing these lights, but it remains unknown as whether some or all of the lights are actually installed in their homes. Survey techniques to capture the customer at the point of purchase must be used to gain contact information. Then follow-up surveys or even site visits may be necessary to understand how many of the new lights are being used and in what capacity. These types of programs may be more susceptible to free ridership and spillover into non-target areas (the stores selling the goods are not in the utility service area, neighboring states are selling goods, consumers from out of the target area are making purchases in participating stores, etc.). Thus, special care should be taken to avoid overestimating or underestimating savings. (This is addressed in more detail in Section 3.)

Of the several examples of program evaluations using state or regional comparisons that have been done, Wisconsin Focus on Energy and NYSERDA chose a comparison group of states—based on income and education levels—that do not run local ENERGY STAR™ programs. The change in market share over time for both the program state and the comparison area is then evaluated. A technique used in Massachusetts relies on a regression model that accounts for a more comprehensive list of explanatory variables, including energy prices, climate zone, population center distribution, and precipitation/drought. The regression solves for the incremental market penetration due to the program (Dimetrosky 2007).

Demand Response Programs

Demand response programs apply rate design, incentives, or technology to motivate customers to change their demand in response to utility prices or system conditions. These programs may either be dispatchable (where the utility requests the demand reduction) or non-dispatchable (where groups of customers work together to respond to pricing signals or plan schedules to reduce their peak demand during desirable periods). For such programs, metering measures the reporting period demand; however, a baseline must be estimated.

Approaches to estimating the baseline may include use of a “representative day”—that is, a day where load was not curtailed but had similar conditions to the day load was curtailed—or, using a Statistical Analysis technique, predicting what the baseline would have been had load not been curtailed (Violette et al. 2007). Because the reliability of communications systems used (which can include paging networks, cellular networks, Internet metering, or metering through home phone lines) is an important component of the program and evaluation, it should be verified. ISO-New England (NE) studied the ability of a demand response program to meet its ancillary services requirements. ISO-NE used an Internet-based, real-time metering system to track five-minute load increments assessing actual equipment demand. The study also utilized a regression analysis technique to determine whether the load reduction was significant compared to normal, unexplained load variations (Agnew et al. 2007).

Performance Contracting Programs

Performance contracting programs often utilize an energy service company or engineering firm to guarantee savings that will pay for program costs. The M&V method, which is stipulated in advance between the energy service company and the customer, is best used to evaluate individual savings. A separate program evaluation may involve additional verifications and could employ statistical analyses, measurements of additional parameters, and/or calibrated building simulations. In one example, an evaluation of a federal performance contractor performed its evaluations in three tiers (Schonder et al. 2007). Tier 1 applied an M&V method using engineering calculations based on estimated parameters. Tier 2, a subset of tier 1, involved verifying earlier calculations, focusing on operations and maintenance savings, substituting measured values for some stipulated values and, where possible, using measured values for

key parameters. Tier 3 verified results under real-world conditions for a set of from three to five projects.

Variations by Use/Application

The type of impact evaluation method may vary, depending on how it will be used. Possible uses of an impact evaluation are:

- Informing the utility and regulatory commission on the overall cost-effectiveness of the program and making a decision whether to keep, change, or eliminate the program.
- Using the results as a basis for paying the program participant (such as in bidding programs or large C&I individual programs).
- Using the results as a basis for paying the utility a specific incentive specified by local regulators.
- Using the results as a gauge in assessing the utility's progress towards a measure's achievable or technical potential.
- Using the results as inputs for DSM planning activities.

Use of an impact evaluation that affects payment of either the participant or the utility will demand more rigor than impact evaluations being used for other actions, particularly if savings are large and involve large payments.

When program plans show a program is clearly cost-effective and the uncertainty around savings is small, the deemed savings method is sufficient, particularly for smaller, residential programs. For a program with larger savings estimates, the statistical analysis method is relatively inexpensive and can be applied with a reasonable level of certainty.

When using an impact evaluation to feed inputs for DSM planning, a cost/uncertainty trade-off should be considered. The critical question is, "What is the value of perfect information?" Large C&I programs with a large impact will be worth spending additional money to gain more accurate information.

Another factor to consider is the length of time the evaluation needs to cover. For example, analyzing end-use metered data (as in the M&V isolated ECM metering approach) for 10 years is likely to be a much greater investment than monitoring it for a single year.

Variations by Region of the Country

Evaluation methods and approaches do not vary systematically by regions of the country, but they do vary according to individual utilities and are dependent on the utility's regulatory requirements and budgets. In some cases, utilities perform two types of evaluations and then compare results to determine the performance of alternate approaches. California was the first state to develop specific protocols to define the level of rigor required for different types of programs. These protocols are outlined in the document "2006 California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Energy Professionals" (TecMarketWorks 2006). Other utilities also apply these protocols; however, these protocols require a significant financial commitment to evaluation.

The California budget for evaluation of its2006-2008 programs is 7.6% of program funding ($163 million). In comparison, the average budget outside of California ranges from 1.62% to 3.1% of program funding, which equals from $1.3 million to $3.6 million in total budget —

significantly smaller in both percentage of program funding and in total funding dollars (Schiller 2007).

The Northeast Energy Efficiency Partnership (NEEP) is developing consistent protocols, and it started this process by assessing the evaluation methods of its members (Northeast Energy Efficiency Partnership 2006). NEEP found that evaluation protocols vary significantly across states and utilities in the Northeast. In particular, the study found baseline conditions are not consistently defined, even though similar algorithms are used to calculate gross savings. Also, deemed savings and standard input assumptions vary significantly, as does the level of rigor or sophistication of the modeling.

2.3 Issues/Problems Identified

While the approaches outlined in Section 2.1 all work for evaluating energy-efficiency programs, actual evaluation experience has uncovered issues and problems associated with the specific application of approaches. An issue common to most measures is the lack of methodology for specifically evaluating *peak demand savings*. A nationwide review of the conference proceedings from the International Energy Program Evaluation Conference and the American Council for Energy-efficient Economy Summer Study Conference Proceedings (from 1994 through 2006) focused on finding methodologies specifically for evaluating peak demand savings. This review found a prevalent lack of methodologies for evaluating demand savings (York et al. 2007). Most evaluations applied load factors or load shapes to energy savings to compute demand savings. Other evaluation issues were associated with the type of measure / sector / program or with use / application as described below.

Problems Associated with Type of Measure/Sector/Program

Unique evaluation problems and issues may arise depending on the type of measure, sector, or program.

- *Large C&I programs* almost always use a measurement and verification (M&V) method. The ECMs in these programs may be complex and require complex assumptions and calculations. The key evaluation issue is whether the assumptions are valid. For instance, a plant operator may say a plant operates at a 90% load factor, while metered data might indicate a 50% load factor. Was the plant operator guessing or were the metering results in some way exceptional? Because wrong assumptions create wrong results, metering is preferred for making correct assumptions. However, even metered results should be verified through plant personnel. Further, isolating the parameters to be metered is not always simple and may be an expensive endeavor.
- *Commercial lighting retrofits* face two key issues: (1) how to accurately account for interactive effects with cooling and heating systems; and (2) how to correctly identify operating hours. One evaluation found that even though lights are scheduled to be off at night, a significant portion was actually left on. Installing end-use meters around the lighting system is the preferred approach, but isolating the building's lighting from its other systems may be challenging. Even with metering, interactive effects with the heating and cooling system still need to be assessed.
- *Residential programs* are difficult to measure because behavioral changes (e.g., changes in use patterns, occupancy, and household size) can overshadow savings when using statistical analysis. M&V approaches require surveys or site visits, and residential customers may not be willing to agree to site visits as are commercial or industrial customers. Furthermore, the cost to visit many residential sites may be prohibitive.

- *New construction programs* are time consuming to model using the M&V approach of calibrated simulation. Models must be accurately calibrated and are sometimes difficult to fit to unique building characteristics. For instance, atriums in commercial buildings cannot be modeled easily. The other approach for new construction is the statistical analysis method, which requires both a sufficiently large participant sample and a similar, corresponding non-participant sample. Particularly for commercial new construction, it may be difficult to find enough participants and corresponding non-participants for the sample.
- *Education programs* are difficult to evaluate because of (1) the variability of how information is presented and (2) the variability of how customers follow the steps on which they have been educated. Savings are sufficiently small so that a statistical analysis may not capture the differences. Also, an M&V approach using surveys could generate biased results (e.g., some participants may feel pressure to say they are changing behavior when, in fact, they are not). The major evaluation challenges are:
    - The lag between campaigns hitting the street and evaluation of the program;
    - The self-report (awareness) problem;
    - Getting through the clutter of energy conservation ads and finding a sample that has been exposed to your message;
    - Attribution of the effect from your program's efforts, distinct from the clutter of other (nationwide, local, regional) campaigns, incentives, and messages affecting behavior;
    - Evaluating/measuring the level or degree of the change and sorting out the say/do gap (i.e., identifying appropriate / useful metrics, and verifying the change); and
    - Assessing retention of the change.

  While most of these issues have been examined in various ways (and work remains to develop better approaches), the literature demonstrates that almost no work has been done to examine the retention of the behavioral change (Skumatz et al. 2000). If these campaigns are to be assessed on the same page as widget-based programs-- or power generation—evaluators will need to tackle this issue.

- *Market transformation programs* are complicated to evaluate, requiring information-gathering from distribution channels as well as from consumers. Multiple surveys, site visits, and even metering may be required to evaluate fully the impacts from such programs. A large amount of secondary data is available for the residential sector; however, challenges with inconsistency among the various sources and questions about availability of data create a number of potential landmines. The analyst needs to be fully familiar with the data, its source and limitations when applying it to an evaluation. Further, even primary data collection has its challenges. The typical use of a non-program comparison group may soon be obsolete as they are getting more difficult to find and ensure comparability to the program group. Finally, available approaches to date have generally not been applicable to commercial and industrial markets.

- *Demand response programs* require finding the right model and conditions to portray baseline characteristics accurately, either by averaging specific days before and after

the interruption to be the "representative day" or through regression modeling of the hourly data.

Problems Associated With Use/Application

Unique evaluation problems and issues may arise depending on the use or application of evaluation results:

- *Cost-effectiveness*: When evaluations are used to inform regulatory commissions about program cost-effectiveness, the particular assumptions and parameters used in an evaluation can significantly change the results. Regulatory commissions or advocates may second-guess assumptions without having any real evidence of a better assumption. Having agreed-upon protocols for verifying assumptions in advance can help prevent this problem.
- *Participant Payments*: If evaluation results are used as a basis for paying program participants, the participant may figure out how to game the assumptions to maximize payment. Sometimes program planners plan a second evaluation following the M&V approach to determine participant incentive payments and to ensure evaluations are not skewed by participants' incentive payment methodologies.
- *Utility Incentives*: If evaluation results are used to determine a utility incentive for implementing a program, a program planner may also be able to game the assumptions feeding the M&V plan. Because of this, most utilities hire independent companies to perform impact evaluations.
- *Progress toward Potential*: If evaluation results are used to gauge progress towards achievable or technical potential, changes in evaluation methodology over time could invalidate previous achievement estimates. Comparisons of resource plans and forecasts over time could raise questions about the planning process among interested parties.
- *DSM Planning*: When evaluation results are used as inputs for DSM planning activities, it can be difficult to understand the amount of uncertainty included in the estimates. Through planning techniques, the DSM planner can assist in determining the value of improved accuracy for the planning forecast.

Depending on intended use of the results and the importance of accuracy, additional tools can be applied to assess objectively the appropriate level of rigor and spending. During 2006, the California Public Utilities Commission (CPUC) through an evaluation team of industry experts[14] developed evaluation protocols for evaluating California IOU's 2006-2008 programs. As part of this effort, a subset of this team was asked to design a system for recommending an approach to allocate resources for ex post evaluations across the portfolio's energy-efficiency programs. This report (Hall, Jacobs, and Kromer 2006) also informed the evaluation planning process regarding which programs should be more or less rigorously evaluated, and it was decided that a Monte Carlo simulation of the CPUC's portfolio could systematically track and quantify the hundreds of relevant data gaps. The Monte Carlo results were used to guide the allocation of evaluation resources cost-effectively to the most deserving elements of the portfolio, i.e. those with the greatest risks from uncertainty.

[14] Hall, Nick, Johna Roth, Carmen Best, Sharyn Barata, Pete Jacobs, Ken Keating, Ph.D., Steve Kromer, Lori Megdal, Ph.D., Jane Peters, Ph.D., Richard Ridge, Ph.D., Francis Trottier, and Ed Vine, Ph.D.

Variations by Region of the Country

Issues and problems do not vary as much by regions of the country as by method and approach used (and how these are applied to specific programs). California, the Northwest, Wisconsin, and the Northeast (including New York) generally offer the most programs and conduct the most evaluation work. These regions have been leaders in developing consistent protocols for evaluation.

Overall Findings/Key Issues Identified

Impact Evaluation Approaches

Evaluators use consistent methods and approaches for evaluating programs, but evaluations still vary by assumptions, rigor, and sophistication of the models used, which impact the amount of money spent on evaluation.

Repeatedly across all methods and approaches, the question arises as to whether *assumptions* are valid. From complex industrial process changes to simple residential compact fluorescent (CFL) programs, validating program assumptions remains vital. Many studies find that a site visit is the only way to determine whether tracking systems are working correctly, installations had been installed or dismantled, or assumptions regarding operating hours varied significantly from actual operations. The person performing the site visit must be qualified to know where to look and to identify which assumptions are most important.

The *level of rigor* is also important. How much evaluation is too much? Some evaluation customers have expressed frustration ("evaluation fatigue") at the amount of information requested to feed an evaluation. Evaluators need to understand both how the evaluation will be used and the amount of savings in question, so they can conduct the evaluation with the appropriate level of rigor. Methods for uncertainty analysis (such as using decision analysis, Monte Carlo analysis, and portfolio theory) can help evaluators ascertain the value of additional (or perfect) information to help set rigor levels of evaluation for different programs.

How *sophisticated do models have to be*? This is also tied to questions about how an evaluation will be used and how much to spend on program evaluation. For most utilities, there appears to be no consistent strategy for calibrating and choosing either an engineering simulation model or a statistical model. One evaluator conducting the modeling who tries to develop a model with a good fit will not necessarily choose the same model that another evaluator would choose.

How many consistent evaluation results does it take to switch to a deemed savings method? If California's *Database on Energy Efficiency Resources (DEER)* database were expanded nationwide, more evaluation results could be applied to more programs. Finding consistent savings among program types and across utilities could allow more applications of the deemed savings method and reduce the evaluation costs.

What is the correct method for *evaluating peak demand savings*? Most evaluations focus on accurately measuring program energy savings, and then they calculate demand savings by an adjustment based on estimated load factors or load shapes. As programs become large enough to contribute significantly to a utility's resource portfolio, accurate peak demand savings must be determined. This requires both an understanding of how the local utility system peaks and what customer savings are coincident with that peak.[15]

[15] A reviewer points out that the DEER database does have stipulated peak demand savings for many measures and may provide an example for other states interested in using this information in their technical reference manuals.

2.4 What Has Been Learned: Emerging Approaches and Experience

The evaluation field has made great strides in the last thirty years, and especially the last ten years. The development of the DEER database and key protocols, guidelines, and databases (such as the IPMVP, the 2000 and 2004 California Evaluation Framework documents, the 2006 California Energy Efficiency Evaluation Protocols[16], and the DEER database) have provided critical resources for evaluation professions. Also helpful has been the many evaluation reports and conference proceedings specifying lessons learned when applying these methods and approaches to specific energy-efficiency programs. For instance, only through this evaluation experience have we learned about some of the assumptions most likely to result in misdiagnosed energy savings, such as lighting neglecting interactive effects, incorrect assumptions about operating hours, and so on.

Key Issue 1

Assumptions are key to obtaining the best evaluation results. Assumptions can be improved in several ways:

- *Incorporating site visits into every evaluation*. Even when using the statistical analysis method, site visits made to a sample of participants will yield information about the likelihood of improper installation or systematic issues with ECMs that could impact the evaluation.
- *Applying existing experience to the site visits*, either through using experienced evaluators who have performed other site visits or through the use of a protocol developed specifically for a type of building and ECM, which would identify specific parameters and the appropriate means for verifying these parameters.
- *Creating a nationwide database of ECMs, savings estimates*, and other information regarding specific measures to serve as a resource to others using the results (similar to some of the California technical reference manuals used outside of California).

Key Issue 2

Protocols have proven to be very useful. The evaluation field is expanding, and protocols are a valuable resource for newer, less experienced evaluators. They also support utilities seeking regulatory approval for their evaluation plans.

Current protocols are a start, but further protocols could be developed, such as:

- Detailed measurement protocols for specific measures in specific building types when using the M&V method. These protocols could detail which parameters are most important to measure, and, if measuring is not possible, how to accurately estimate that parameter.
- A specific protocol or software tool to strike a balance between rigor and budget. Ideally, this would apply uncertainty analysis techniques to measure the value of more accurate results.

[16] Some suggested that California has set itself up as the (quite costly) "gold" standard, and the costs are not practical for most other utilities and states (Mulholland 2009).

Key Issue 3

Education and consumer-based social marketing deserve special attention. As a result of this review and considerable other research, we propose several suggestion for analytical methods to use to get past some of the difficulties associated with measuring the impacts of education and social marketing campaigns and for optimizing outreach efforts. Our suggestions include:

- Apply test and control approaches using baseline and appropriate experimental design methods where possible. To provide reliable results, it is important to revisit the issue of RANDOM assignment to groups, which can be politically complicated, but is essential for reliability in evaluation.
- Consider setting up quasi-experimental program designs using different communities as treatment and control groups.
- Consider gathering cross-section information from programs implemented in multiple communities and use regressions to control for differences in programs, demographics, and educational efforts / designs.[17]
- Consider more frequent use of some of advertising techniques (including focus group tests of intentions to purchase) and use these methods to do a preliminary test of campaigns and educational materials for effectiveness.
- Consider evaluating several common types of education/outreach programs (template programs) and apply their results for similar programs in other communities as order-of-magnitude estimates.

The state of the art in education measurement and evaluation is lagging behind the effort and best practices that have been developed in administering outreach/education. Practitioners, researchers, and experts of marketing (especially consumer/community based social marketing, CBSM) have established a proven framework to increase the impacts of education and outreach. However, evaluators of education/outreach program impacts have not developed a parallel framework for measurement. A review of case studies and publications has proven beyond a doubt that outreach and education play a significant role in ECM impacts and effectiveness; in fact, these efforts can increase energy efficiency by as much as 50% over control or test routes. Unfortunately, the data on impacts of outreach/education are sparse and measurement is still in its beginning stage.

Generally, given the complexity of finding control groups and of controlling the recipients of information, it may be worth examining whether detailed evaluation of all education programs is important or cost-effective. If programs are reasonably similar, evaluation may not be needed for each program, especially if the level of evaluation needed is to (1) assure money is being spent responsibly, or (2) provide the level of accuracy needed to guide program decisions or avoid expensive wrong decisions. This second point does not always need precise information. Instead, several common types of programs (templates) could be evaluated, and their results applied in orders of magnitude to other cities. This might be used in developing broad guidelines for expenditures on education versus measures in programs. As a substitute, additional pre-testing of materials for quality and resulting changes in intentions (as they do in advertising) may be a useful and cost-effective evaluation approach for most programs. Focus

[17] This approach proved very successful when applied to outreach for recycling programs. After gathering data on more than 120 recycling education / advertising campaigns, the researchers used regression techniques and were able to develop estimates of marginal impacts from outreach campaigns, different media / outreach methods, etc. (Skumatz, 2000, Skumatz and Green 2002).

group work to control for quality of the program materials may then be the most effective use of the evaluation funds.

The review also made it clear that multivariate techniques have seldom been applied in these energy fields--likely because of the complexity of the behavioral changes involved, the difficulty of separating the effects from hard factors (such as program components) from soft factors (such as outreach methods) and because the field matured to that degree. Using cross-section regression analysis methods to examine the impacts of education/outreach programs may be a fruitful approach.

2.5 Conclusions and Additional Research Needed

Protocols have been proven effective in aiding evaluators in the preparation and implementation of evaluation plans. Additional research is needed in creating protocols to the next level of detail: by specific end-use, program, and building type. Other research needed involves the balance between precision, uncertainty, and costs of evaluation, so that the results could be input to a decision-making protocol or tool used for preparing evaluation plans and budgets.

2.5.1 Conclusions

A number of conclusions can be derived from the research.

- Energy savings is the absence of energy use. This is an elusive substance to measure as it requires measurement of “what did not happen.” The common approach is to compare the before and after states to determine what has changed. This element of change should reflect the elimination of some portion of the prior energy use after implementation of a program.
- For the three different kinds of energy efficiency programs (resource acquisition, education/information, and market transformation), estimation of gross savings starts with an estimate of “participation.” Definition of participation varies by program type from direct installation of a measure, to change in behavior impacting energy, to change in market penetration of an energy saving technology. In all cases, a per unit gross savings estimate is needed.
- In estimating gross program impacts, one can use: secondary data, deemed savings, engineering models, statistical models, or metered data.
- Behavioral programs have the potential to greatly increase the impact of EE programs, especially when CBSM techniques are incorporated into program design. However, measuring and valuing the impacts of these campaigns are difficult and compounded by independent factors.
- Evaluations often utilize the International Performance and Measurement and Verification Protocol (IPMVP) Option A for evaluating savings from measures installed and behavior changes. Option A is appropriate in instances where combined uncertainty from all estimates will not significantly affect overall reported savings and estimates are realistic, achievable, and based on equipment that can produce savings.
- Option A is used where multiple energy conservation measures are installed and savings are expected to be less than 10% of the utility metered consumption. This methodology is less costly than billing analysis, and can be used to control evaluation costs if key parameters used to compute savings are well known. Additionally, use of phone and in-person surveys, plus engineering algorithms, offer reasonable and cost-effective means to estimate savings.

- Billing analysis using weather-normalized consumption data provided by the utility commonly is used to estimate gross savings. Billing analysis requires consistent residency for two or more years, so one year of pre-program data can be compared with one year of post-program data.
- Billing analysis may be used to estimate gross savings of education programs combining low-cost measures and behavior modification. However, as billing data are inherently too "noisy," gross savings less than 10% of pre-consumption levels are hard to detect.
- Programs that are more far reaching (general outreach and education or market transformation) pose a serious challenge to evaluators as participants are not usually known.
- When measure savings are established in more rigorous studies, the use of self-reported data may provide sufficiently reliable estimates of gross savings.
- Self reported data are often augmented with site visits and selected metering (e.g., hours of use).
- Random digit dialing combined with sales data are used to estimate program impact on saturation levels. Change in saturation combined with per unit engineering estimates of savings can be used to estimate gross savings.
- Partner and other secondary data sources may help in evaluating residential measures, but are less useful in measuring progress for C&I measures.
- Per unit savings often use simple engineering algorithms. However, when interaction impacts are significant, simulation models are preferred.
- Evaluation methods/approaches vary by individual utilities and are dependent on the utility's regulatory requirements and budgets. Choosing an evaluation method must achieve the ultimate goal: balance the evaluation's cost with the value of the information received.
- Gross demand savings are estimated using existing load profiles, simulation models or end use metering.
- As the stakes of energy efficiency efforts increase, the need for transparency also increases. Evaluations need to be transparent and results need to be reproducible. This calls for common approaches. Several protocols currently exist in different parts of the country. More effort is needed in creating more common approaches at the national level.
- At a minimum, evaluation methods need to be clearly laid out before any data collection is conducted. When evaluation is concluded, all limitations of methods and results need to be clearly identified.
- All evaluation approaches must include an assessment of the associated uncertainty.
- Having agreed-upon protocols for verifying assumptions in advance can help prevent regulatory commissions or advocates from second-guessing assumptions and parameters used. Results should never dictate methods.
- Through planning techniques, the DSM planner can assist in determining the value of improved accuracy for the planning forecast.

- As widespread education campaigns affecting both target and non-target audiences become more common, finding a baseline to measure against is more difficult - it is hard to uncover a population with a "zero" behavior baseline
- The major evaluation challenges are (1) the lag between campaigns hitting the street and evaluation of the program (2) the self-report awareness problem (3) getting through the clutter of energy conservation ads and finding a sample that has been exposed to your message (4) attribution of the effect from your program's efforts, distinct from the clutter of other (nationwide, local, regional) campaigns, incentives, and messages affecting behavior, (5) evaluating/measuring the level or degree of the change, and (6) identifying appropriate / useful metrics, (7) verifying the change, and (8) assessing retention of the change.

Best Approaches Summary

Protocols and guidelines are already available: the IPMVP, the 2004 California Evaluation Framework, and the 2006 California Energy Efficiency Evaluation Protocols. These are used by evaluators across the nation and provide a basis for the best approaches to evaluation.

2.5.2 Additional Research Needed

Emerging Research Approaches

New technologies and approaches, such as using the Internet for metering devices, developing baseline studies for plug loads, and preparing detailed evaluations of minute-by-minute responses of demand-reduction programs, are some of the more innovative methods for getting better evaluation results.

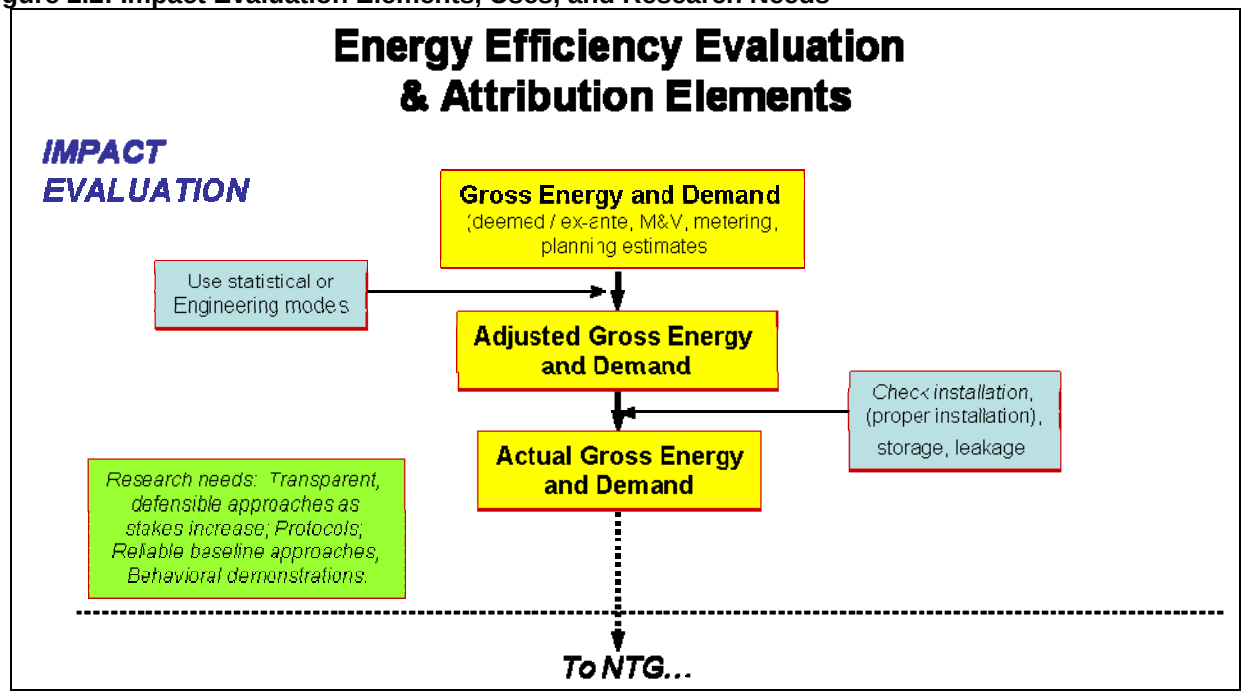
Additional Research/Steps to Address Remaining Issues

The following research items are needed to develop an additional protocol or tool for the trade-off on rigor, budget and uncertainty levels:

- First, research into techniques for evaluating uncertainty (such as decision analysis, Monte Carlo analysis, or portfolio planning) could be useful in helping make trade-offs between budget and rigor.
- Second, research into typical budget levels or ranges needed to attain specific levels of rigor in an evaluation, summarized in terms of per site or per study, depending on the method of evaluation.
- Third, specific algorithms could be developed to balance the rigor level against budget ranges. These algorithms would consider the use or application of the evaluation as well as the type of measure and program used.
- Additional testing of new methods for estimating or reflecting market progress attributable to programs is needed. In measuring market share progress, improvements in reporting for the underlying sales / shipments data sets are important to provide more confidence in evaluation results. More exploration of alternative approaches (price decomposition or others that might be identified) to track their accuracy or consistency compared to traditional approaches could prove valuable.
- A research need common to most measures is a methodology for specifically evaluating peak demand savings.

- Protocols are a valuable resource for newer/less experience evaluators and need to be further developed.
- Research into techniques for evaluating uncertainty could be useful for trade-offs between budget and rigor.
- Research is needed for summarizing budget amounts for specific levels of rigor.
- Research is needed for developing specific algorithms to trade-off the rigor level against budget ranges.
- Consider evaluating several common types of education / outreach programs ("template" programs), and use their results for similar programs in other communities as order-of-magnitude estimates.
- To best measure impacts of education, social marketing, and outreach campaigns, apply test and control, quasi-experimental programs, use regressions to control for differences in programs from multiple communities, use softer advertising techniques, and evaluate several types of programs for a template.
- For education programs, the literature demonstrates the need for work to examine the retention of the behavioral change. There has been little to no published work documenting retention.

Figure 2.2: Impact Evaluation Elements, Uses, and Research Needs

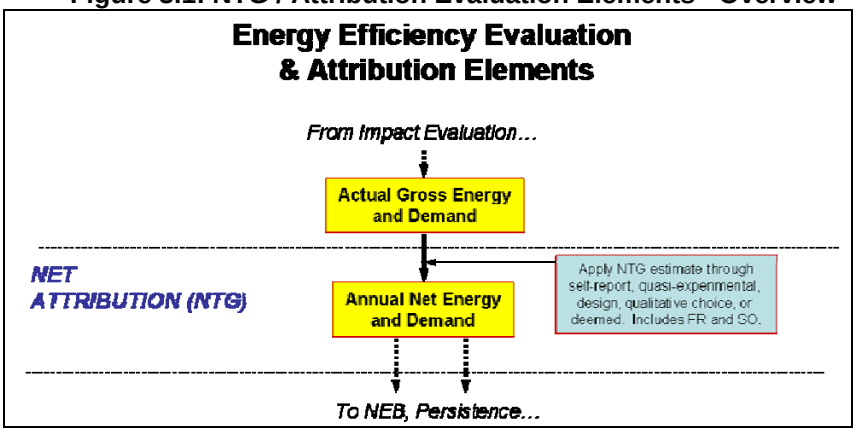


3. ATTRIBUTION / FREE RIDERS / NET TO GROSS

Whichever of the techniques for estimating energy savings is used, estimation of gross effects is only one step in the attribution of "net" effects to specific programs. The "net" effects are a significant element of the assessment of benefits and costs for a program, computations that in some states can determine the start, continuation, or termination of a program's funding.

3.1 Current Practices and Uses

Figure 3.1: NTG / Attribution Evaluation Elements - Overview



Estimating the effects of the program above and beyond what would have happened without the program involves another step – identifying the share of energy-efficient measures installed / purchased that would have been installed / purchased without the program's efforts. Some purchasers would have purchased the measure without the program's incentive or intervention. They are called "free riders" – they received the incentive but didn't need it. Others may hear about the benefits of the energy-efficient equipment and may install it even thought they do not directly receive the program's incentives for those installations. These are called "spillover"[18] – implementers that were not recorded directly in the program's "count" of installations. The combination of the "negative" of free ridership and the "positive" of spillover are computed as a "net to gross" (NTG) ratio, and are applied to the "gross" savings to provide an estimate of attributable "net" savings for the program.[19] The NTG, or its components, have been addressed in four main ways:

- *Deemed (stipulated) NTG*, where some net ratio is assumed (1, 0.8, 0.7, etc.) that is applied to all programs or all programs of specific types. This is generally negotiated between utilities and regulators or assigned by regulators.
    - Advantages: Simple / uniform / eliminates debate; no risk in program design / performance; if less than zero, reflects the likelihood of some free ridership with most programs; inexpensive.
    - Disadvantages: Does not recognize actual differences in performance from different programs / designs / implementations.

[18] There are commonly three types of spillover. Inside project spillover occurs, for example, where refrigerators are rebated, and the person receives / installs that equipment, and then later installs an energy-efficient dishwasher. Outside spillover occurs, for example, when a builder gets rebates on one project, but then starts to install similar efficient measures in other homes even without rebates. Non-participant spillover occurs, for example, when builders hear about energy efficiency and do not participate or receive any rebates, but decides to install efficient equipment to serve his customers or to keep up with other builders, etc. No incentives were provided for these measures. Sometimes, the first two examples are referred to as Participant Spillover and the third example as Non-Participant Spillover.

[19] The literature shows computations of this NTG ratio by adding the factors (1-FR+SO) or by multiplying the factors ((1-FR)*(1+SO)). Both are used in practice.

- *NTG adjusted by models with dynamic baseline*: in this case, a baseline of growth of adoption of efficient measures is developed, and the gross computation of savings is adjusted by the estimate from the baseline for the period.
    - Advantages: Can reflect differences in performance for good / poor designs and implementation.
    - Disadvantages: Complicated to identify appropriate baseline; data intensive; potentially expensive; introduces more risk to program designers related to program performance; may lead to protracted discussions.

- *Paired comparisons NTG:* Saturations (or changes in saturations) of equipment can be compared for the program (or "test") group, vs. a control group. The control group is similar to the test area in all possible ways, but does not offer the program being studied – or those particular customers do not receive the program. Ideally, pre- and post-measurement is conducted in both test and control groups to allow strong "net" comparisons.
    - Advantages: Can reflect differences in performance for good / poor designs / implementation; straightforward and reliable evaluation design.
    - Disadvantages: Control groups can be difficult to obtain[20]; if imperfect control groups are used, statistical corrections may be subject to protracted discussions.

- *Survey-based NTG:* In this approach, a sophisticated battery of questions is asked about whether the participant would have purchased the measures / adopted the behavior without the influence of the program. Those participating despite the program are the free ridership percentage.[21] These are then netted out of the gross savings. Spillover batteries can also be administered to samples of potential spillover groups (participants, non-participants).
    - Advantages: Provides an estimate of free ridership and spillover; can explore causes and rationales.
    - Disadvantages: Responses are self-reported leading to potential bias or recall issues; may be expensive; can be difficult to get good sample of respondents for free ridership[22]; requires well-designed survey instrument which can be long and which affects response rate.

The measurement of spillover is more complex than the measurement of free ridership. Free ridership emanates from the pool of identified program participants; the effects from spillover are not realized from the participating projects and, in many cases, not even the entities that participated. Identifying who to contact to explore the issue of spillover and associated indirect effects can be daunting.

---

[20] Control groups may be local, using the same pool of customers, with random assignment of potential participants into groups that can vs. may not participate. These can be politically difficult (utilities are unwilling to refuse participation to otherwise-eligible customers) or technically difficult (CFLs are on a shelf and only some customers are allowed to purchase – although presumably some scratch cards good only at checkout could be offered). Often, non-local control groups are constructed. That is, another state or county that is demographically or othewise similar to the test area or utility territory is selected, and a sample of those households or businesses is used as the control. It can sometimes be difficult to locate areas that are unambiguously similar to allow control for all relevant influencing factors, and the difficulty is increasing (as fewer areas are program-less). Some statistical techniques (propensity scoring and other corrections) can show promise for helping address this issue to varying degrees of success (Skumatz and Gardner 2006, Skumatz 2002)

[21] Using enhanced batteries of questions allowing for partial spillover, and asking corroborating questions can help improve the reliability of this approach. This is the direction that survey-based or self-report NTG analyses have increasingly conducted over the last 5 years (Skumatz and Violette 2004, Schare and Ellefsen 2007, and others)

[22] This latter group consists mainly of persons that never participated in the particular program and have no compelling reason to reply / respond.

Our interviews and literature review suggest that a number of utilities consider free ridership, but do not include spillover (also called free drivers) in their analyses of program effects. As an example, one major California study addressing net to gross explicitly limits its analysis content to free ridership. This asymmetric approach undervalues energy efficiency.

There is considerable – and growing - controversy regarding the use of net to gross, particularly in regulatory proceedings. NTG ratios can be used to reduce (incorporating free ridership) or potentially expand (if extensive spillover is associated with the program) the amount of savings attributable to a program. The argument is that the program carefully estimates (gross) savings that were delivered, but then the savings (and, directly, the associated financial incentives to the agency delivering the program) are discounted by a free ridership factor measured by potentially less-than-reliable means. This has huge potential financial impacts in some states in which utilities may receive financial awards for running programs and running them well. The controversy arises from the following main issues:

- The potential for error and uncertainty associated with these measurements, because of difficulties in (1) identifying an accurate baseline; (2) identifying / implementing a control group; or (3) relying on self responses to a survey.
- The expense of high quality analysis – with arguments that the money could be better spent on program design, implementation, incentives, etc.
- Baselines and effects are harder and harder to identify and analyze as programs move up stream, involve different levels of vendors and other actors, and lead to changes in baselines up the chain. In addition, program spillover complicates control group assessment.
- The difficulty in separating out the effects and influences of different programs within a marketplace (own utility / agency and outside utility / agency).
- Concerns that using measured NTG or free ridership ratios introduces a great deal (to some, an unacceptable level) of risk into the potential financial performance metrics for the program, which will lead to "same old / same old" programs and reduce innovation in program offerings.[23] In addition, some programs cannot control the amount of competing activity that enters the target areas, limiting the analytical research.

Baselines are a very important part of the problem of measuring net to gross (NTG), free ridership (FR), and spillover factors (SO). Documenting what "would have happened" is the biggest challenge in evaluation (Saxonis 2007). Many interviewees suggested that strong market assessment is needed up-front to provide the maximum amount of baseline information. However, when it comes to the dynamic retail sector, it may be impossible to predict what they would have done without the program (Messenger 2009) – especially if changes occur upstream.[24]

Baselines relate to what would have happened without the program, which is generally understood to mean standard practice. Standard practice might generally be expected to relate to codes and standards. However, in one study (referred to in Mahone 2008), the issue of baseline was found to be quite complex. Mahone (2008) notes that for at least the multifamily

[23] Innovation is valuable, but agencies will not innovate (cannot justify innovating) in programs unless the risk is reasonably predictable. However, on the other side, regulators must assure that the reward structure doesn't encourage ineffective programs and that funding is spent appropriately and prudently.

[24] And some of those upstream changes will spill over to areas that might otherwise be considered potential control areas. If a manufacturer is induced to change the manufacture or mix of product, and they do so for California which is a big enough market to swing production in general, the new product lines will become available in the potential control areas and the (important) market effect is then reduced.

sector, none of the buildings were being built to the level of baseline codes – i.e., they were underperforming, so that the baseline of standard practice was below the baseline of codes. In this case, NTG would be estimated as greater than 1. More research on standard practice in the field would provide a stronger basis for baselines.

3.2 Overall Findings on NTG Results - Consideration and Values

A number of states reportedly use the California Standard practice manual, or large portions of it, for energy savings, free ridership, non-energy benefits, and benefit cost regulatory tests, including Oregon, Washington, Idaho, Montana, Wyoming, Utah[25], Iowa, Kansas, Missouri, New Mexico, and Colorado. (Hedman, 2009). Several studies specifically examined state and utility practices regarding free ridership and net-to-gross. These studies find that utilities treat the issue of NTG differently. In some cases, there is no regulatory agreement on the estimation of NTG, and they historically treat FR only in the calculation of the NTG ratio. The Nevada Power and Sierra Pacific Power collaborative examined FR and spillover in 23 states and/or utilities serving states. They found 15 states (69%) did not use FR (equivalent to defaulting to a FR value of zero) in estimating net savings (Quantec 2008). Other states say NTG is too costly and biased. Massachusetts prefers to have utilities focus on MT programs and correct for factors affecting gross to net savings in program design. California requires deemed FR values in the calculation of the NTG, but excludes spillover. In Iowa, estimating NTG is not a priority - they feel FR is balanced by spillover and make no further efforts.

Data from organizations around the nation found that about half of the studies (49%) assumed or calculated a NTG factor of 1. Two-thirds assigned values between 0.9 and 1.0. In most cases, the NTG was based only on FR or on deemed values. There was little reporting of spillover (Fagan 2008). Minnesota and Wisconsin publicly stated that FR and SO cancel out (Quantec 2008). Iowa said it assumed a NTG value of 1 because measurement of FR and SO was unreliable; when it did measure NTG, it came out near to 1 (Quantec 2008). In Illinois, NTG ratios of 0.8 are assumed for low income, lower for appliances, and they are looking at others (Baker 2008). Washington reportedly doesn't support savings from behavioral changes or NTG / FR / spillover allowances or disallowances (Drakos 2009).

In addition to studies of state and regulatory practices, we were interested in identifying patterns in results for programs and regions. We assembled and reviewed more than 80 evaluation studies from California, New England, and the Midwest that contained that contained estimates of free ridership and / or other elements of net-to-gross. The studies, which covered residential (including low income) and commercial programs, provided estimates for lighting, HVAC, new construction, appliances, motors, and numerous other measures delivered through rebate / incentive, and non-incentive programs. The studies covered programs dating from 1991 to 2008. We examined the studies for patterns in methods between areas of the country, and in free ridership and net-to-gross results by sector, measure, or region. Although the studies were assembled as a convenience sample, not a statistical sample, we found the following general results.

- The vast majority of the studies relied on self-report surveys to generate the results, using variations and enhancements on questions related to likelihoods to participate or purchase without the program's influence. A small percent reported using logit / ranking / discrete choice modeling approaches.

[25] Utah only allows one year of lost revenues in the Rate Impact Test.

- Less than 10% of the studies reported confidence intervals associated with the estimates.
- Studies from the Northeast were more likely to include estimates of spillover than the California analyses.
- There were far fewer estimates of free ridership for kW than for kWh. For indicative purposes, NTG estimates for the energy (kWh) tended to be higher than for kW, but note that kW sample sizes were small.
- *Ex post* NTG estimates for programs generally clustered around 0.7 to 1.0, but dipped as low as 0.3 and as high as 1.3.
- Only a small subset of the programs included free ridership or NTG estimates for gas / therms.
- *Ex post* free ridership estimates clustered around 0.1 to 0.3, but ranged as high as 0.5 and 0.7 for some commercial HVAC / motors programs, and some refrigerator initiatives. Low values were found (0.03) for several low income programs included in the sample.
- Net to gross figures for whole homes and retrofit programs tended to be relatively high – usually in the 0.85 to 0.95 range. However, outliers included whole home programs with NTG figures as low as 0.5 and as high as 1.0.
- Some studies included both *ex ante* and *ex post* NTG figures for the same program. Our review shows that the *ex post* values were generally 10-20% lower than the *ex ante* values. The most obvious exceptions were some cooking measure programs (*ex post* was about half the *ex ante* value), and some refrigerator programs that reported spillover values greater than 0.5.
- Net realization rates were provided for about one-third of the programs, and the values averaged about 0.7 to 1.0. A number of values exceeded 1.0, including examples of commercial HVAC rebate programs (1.07), refrigerator rebate program (1.15), and a few others. Several showed net realization rates between 0.3 and 0.5 including several CFL programs, some refrigerator programs, and some gas cooktop rebate programs, and some EMS initiatives, among others.

Reviewing the results across programs, we found that the high and low performing programs (in terms of NTG, FR, or realization rates), did not include all the programs focused on cooktops, refrigerators, or other specific measures. Instead, some programs including or focused on these measures were outliers and others performed more toward the means. That is, measure-level NTG performance varied, presumably depending on elements of the underlying program design and possibly due to measurement techniques as well. While these findings are useful, additional, and more comprehensive, work of this type is clearly needed before broad conclusions can be drawn.

Table 3.1: NTG Results

	Net To Gross , Free Ridership, Spillover
General results	Most utilities and regulators exclude NTG or assume values that incorporate only FR and range from about 0.7 to 1.0 (*ex ante*). *Ex post* results have been measured for many programs; spillover is measured much less often than free ridership.
Variations by measure type, program type or region	Clear patterns for FR, SO, or NTG results by measures, program types, and regions have not been demonstrated to date. The assumption is that variations in specific program design and measure eligibility definitions are important to results. NTG results in the literature are also affected by whether or not spillover is included in the assessment.
Variations for behavioral vs. measure-based programs	Studies addressing net-to-gross, free ridership, or spillover estimates associated with strictly behavioral programs were not found, and if available, are probably too few in number to lead to overarching conclusions or patterns.

3.3 Issues / Problems Identified - NTG Measurement Approaches and Practice – Emerging Approaches and Experience

Refinements in Standard Practice
Historically, fairly simplistic measurement methods have been used to estimate free ridership. The computations have been based on self-reports, with error coming from faulty recall in the form of bias toward claiming the program was not influential or influential, and with bias from the form of hypothetical questions.

Improvements in the self-report literature have included questions to allow "partial" free ridership. Later, studies combined partial free ridership with a review of "influencing factors" or "corroborating questions" which were used to adjust FR reports based on the combined evidence from the other questions. For example, the questions might ask about the importance of the rebate in decision-making, whether the purchase was moved forward two years or more, whether they were already aware of the measures, and similar questions, and used these responses to validate or adjust responses to direct free ridership responses. (Skumatz, Woods, and Violette 2004). Some consultants have required free riders to meet four criteria – they had to be: aware of the measure before the program, intending to purchase before the program, aware of where to purchase the measure, and willing to pay full price. If the four conditions were met, the household or business was classified as a free rider.

In the Northwest, the Oregon Trust conducts long-term tracking on a number of programs –they assess the market, identify program influencers, and conduct in-depth research in order to determine how much of the gross savings to claim for the programs (Gordon 2009).

But most organizations use simple questions (yes/no), which leads to response bias. MALM (1996) circumvented these difficulties by analyzing revealed choices of high energy heating systems purchases among different clusters of customers and found that 89% of households would have bought EE even without the subsidy. Statistical methods, for example difference of differences, are also used.

Splitting the Credit
One key refinement may be the recognition that we may not be able to attribute "causality" to one program or intervention, but may need to consider splitting the credit. The issue of "chatter in the marketplace" is a concern, but this is also an issue for technology / measure / economic based programs as well as education / outreach programs. However, the industry has been more willing to apply causality to technology measures because we can see something put an implementation or desired decision "over the top" more clearly. It is important to understand what is happening in the market and if a 0/1 litmus test is required for causality, it is unlikely to be "proved" as attributable to a particular program or element (Messenger 2008). Recent attitudinal research from the Energy Center of Wisconsin confirmed that people get energy-saving information from multiple sources including utilities, and programs and elsewhere, concluding that… "it may take a village to raise a behavioral kilowatt-hour sometimes" (Bensch 2009). This may make it hard to attribute the kilowatt-hour to one specific influencer, but that doesn't make the kilowatt-hour less real. The solution may be to acknowledge shares of the kilowatt-hour to multiple contributing factors (for behavioral and technology measures) and share the credit (Bensch 2009). And sharing the credit may be the right answer, as people may only pay attention if it is a 'whole choir singing the "save energy" song' (Bensch 2009). Sulyma (2009) argues that it is more than time to move beyond only "one" plausible explanation for impacts, and that probabilistic methods should be used to address this attribution issue.

Randomized Methods

The issue of a control group or baseline that is reliable is a continual problem. Train (2009) suggests that the best way to address the issue is up-front random assignment, a technique that was in use 15 and more years ago, and that is still, he argues, the best approach – providing there is political will to institute this approach and deny some volunteers the ability to participate in the program. Many interviewees also agreed the historical tools of well-designed randomized control and treatment groups were well-suited to impact evaluation (and attribution) for behavioral programs and would provide results that could be generalized; however, it was suggested that in some cases the tools weren't "blessed" for use, or the evaluators and regulators have not developed the kind of faith in them that they have in other measurement protocols. The use of these approaches with appropriate modeling (including mixed logit, discrete choice, etc.) shows best promise (Ridge 2008, Train 2008, Barnes 2008). There is also concern that these random techniques become more complicated as controlling for the many influences is complex (including spillover), making a battery of questions important to the analysis (Messenger 2008, Cooney 2008, Train 2008). To help control this, evaluations need to dig deeper into understanding the topics influencing consumer decision-making. However, these kinds of tools – well-accepted in other social fields and with history in energy - apply well to energy-based behavioral programs. More evaluations of behavioral programs, and greater widespread cataloguing of the results (along with time), may be necessary to gain greater acceptance by regulators.

The discussion of NTG and free ridership has resulted in a number of papers over the last decade, as discussed below.

Methodologies

Methodological work has been a focus for a number of papers.

- Skumatz, Woods, and Violette (2004) summarized NTG approaches based on interviews with multiple stakeholders, and question batteries that account for partial free ridership and used "corroborating questions" to triangulate responses and confirm self-report responses to work toward more robust NTG estimations.
- Chappel et al. (2005) summarized major free ridership analysis approaches, including billing / econometric approaches, difference of differences, econometric choice, and self-report.
- Saxonis (2007) looked at spillover and free ridership in New York versus other programs, using a multi-question approach. He found wide variation in results, noted that there had not been a focus on NTG findings, and that there had not been enough measurement of NTG. He suggested a need to improve reliability and leverage results to maximize the value of evaluations and increase collaboration.
- Friedman (2007) examined the value and accuracy of NTG estimates and argued that at different stages of the market, the NTG ratio changes depending on the actor / participant and the maturity of the market, not always directly due to, or under the control of, the program. He recommended changing current policies to account for spillover.
- Meissner et al. (2008) examined 12 programs using Monte Carlo and risk analysis work, examining NTG as a source of uncertainty. They concluded that some evaluations should focus on NTG to limit uncertainty of the results.
- Cook (2008) compared the pros and cons for self report, econometric, and market share approaches, noting surveys can have low cost and can be used with any program, while econometric methods need a great deal of data and are expensive.

- Peters and McRae (2008) highlighted the problems with the use of free ridership, suggesting that common estimation methods overestimate the effects, and oversimplify the underlying decision-making (decision-making isn't linear, people change their mind, and cognitive dissonance leads them to misstate the influence of the program in the purchase decision).
- Titus and Michaels (2008) conducted interviews with professionals around the nation. They recommended (1) minimizing the use of deemed values for net impact adjustments, (2) making free rider inclusion mandatory (including partial and full FR) and spillover (participant and non-participant) optional, and (3) encouraging continued creativity in estimation methods (to be revisited in 2-5 years to review progress and develop more standardized methods for New England).
- Peach 2009 argued for a model in which policymakers call for adoption of physical targets, and FR is understood as an overhead in being open for business and that the "S-curve" adoption is a simple reality.
- The meta-analysis / best practices study by Fagan et al. (2008) provides methodology conclusions as well. This study examined residential and commercial free ridership from 52 evaluation studies. They concluded that portfolio level FR values have been relatively constant since 1980 despite widespread changes in equipment markets. They noted that 2004-5 DEER-based average FR values were 0.72, with a low of 0.49 for residential new construction programs and a high of 0.87 for commercial new construction. Average values for 1994 were 0.7, ranging from 0.3 to 0.97, and 1988 figures were 0.80 for commercial audits, 0.60 for commercial incentives, and 0.50 for industrial incentives. They concluded that spillover estimates were too uncertain for use in estimating net benefits. They suggested that the choice of a specific methodology for measuring FR is complex and should consider the policy context, level of market transformation, specific program delivery approach, size of the evaluation budget, and availability of comprehensive and reliable data sources.
- A study by Ridge et al. (2009) examined self report methods for computing NTG and used a more exhaustive approach gathering data from five sources (program files, decision-maker survey, vendor survey, account representative survey, and other information) to develop estimates that they felt were more defensible estimates of program influence and free ridership.
- There has also been significant work to develop standardized approaches and questionnaires for use in California.[26]

Quantitative Studies

Most studies relied on self-report survey techniques to develop estimates, and report the results using various terms that represent similar effects – net to gross and realization rate, among others. Results from a few studies are listed below; however, there are scores of reports in the literature.

Non-residential studies:

- Rufo et al. (2000) examined NTG in the small and medium non-residential sector. For a furnace program,
- Macrae et al. (2005) used on-site audits and telephone approaches to estimate NTG for lighting and mechanical equipment programs and found values of 83%, and realization rates of 78% for kWh and 66% for therms.

[26] For example, the Joint Simple Net of Free Ridership and Participant Spillover Self-Reprot Survey Battery (3/6/08), and the Proposed Net-To-Gross Ratio Estimation Methods for Non-Residential Customers.

- Yogesh et al. (2005) used phone and on-site survey approaches to estimate NTG for a business program, and found realization rates of 87%.
- Ross et al. (2007) estimated commercial HVAC free ridership (self report method) and found free ridership values of 38%-47% (2005, 2006), and spillover of 5% and 2% for the same years.
- Erickson (2008) analyzed commercial / industrial technical assistance, rebate, and audit programs at two utilities. He attempted to explore the carryover effects of programs (one analysis examined current and prior year impacts; the other looked only at current year effects). Examining multi-year impacts, he found FR was 35%, with 8% of the savings due to previous year participation, and 35% spillover. No estimates were presented for the other analysis.
- Torok and Bradley (2009) examined non-residential audit programs and conducted more than 6,000 customer surveys to develop estimates of free ridership scores for each measure. The scores were weighted by energy savings to determine weighted free ridership values for each measure and overall. They found that NTG values for the audit program varied based on small versus large customers, and rebated versus non-rebated measures. They also concluded that the audit program had generally lower free ridership ratios than rebated measures.

Residential Programs:

- Tiedeman et al. (2005) surveyed residences and trade allies and a difference of differences approach to estimate NTG.
- Austin et al. (2005) used telephone and on-site visits to estimate NTG and free ridership for rebates, direct install, and information / training programs and identified NTG values for electricity savings ranging from 75% to 100% depending on measure.
- Dohrman et al. (2007) found NTG ratios for refrigerator programs ranged from 0.35 – 0.53.
- Schare and Ellefsen (2007) used a self-report methodology to estimate free ridership for a loan program and found 33-40% of measures would likely have been installed without the program.
- Bicknell et al. (2008) conducted 900 surveys for a national air conditioning benchmarking study to estimate spillover.
- Hoefgen et al. (2008) examined NTG for clothes washers and CFLs and found the CFL NTG ratio was 2.48-3.28, and the figure for the appliance program was 0.28. The difference was attributed partly to the relative maturity of the appliance program.

Experimental Design – Measurement Options

Reliable measurement methods are available that suit many program types:

- *Require random assignment for participants and non-participants for as many program types as feasible.* The experimental design approach has been well known for decades. The regulators, utilities, or agencies will need to "bite the bullet" in terms of the political fallout from those that want to participate but are put into the "no treatment" bucket. This will require measurement over time, and cookie cutter evaluations may not work for all programs – program specific evaluation design will be needed especially when vendors and upstream agents are involved. This approach may be especially important for

outreach and behavioral programs.[27] Train (2009) suggests pairing this with a discrete choice model[28] to predict behavior. The issue that is problematic for best practices is that participation is usually correlated with unobserved factors. In real practice, it seems there is never a true control group of those not offered a program. There is propensity scoring and other approaches that try to correct for this problem, but there is little substitute for a true random selection experimental design – which then allows transferability of results.

- *Consider survey designs that introduce a real-time data collection element*. There have been several instances in which utilities have introduced NTG-surveys as part of the program participation documents and gather early feedback – near the point of actual decision-making – on the program's influence in adopting the measures (Skumatz 2008). This provides several benefits: increases return rate / sample size (and eliminates the problem of finding participants after they have moved or after years of delay); provides on-going data and allows evaluation at virtually any point after the program is implemented to support on-going refinement of programs; significantly reduces the cost of surveying and evaluation; provides more accurate data if the point of feedback is close to decision-making (recall may be improved); and helps to sort out which programs had what degree of influence. This may be suited to education programs as well as "widget" programs.[29]
- *Consider discrete choice modeling approaches*. These approaches introduce explanatory variables that help to address issues of imperfect control groups, unobserved factors, etc. to allow improved estimates of attributable impacts.

Application and field-assessment of more reliable and robust measurement options is an important issue and bears further research.

Uses of NTG and Its Elements

Reports and experts were concerned that California's methods, results, and applications hold too much sway across the nation – that California is not the rest of the country, and vice versa. There is greater need to recognize degrees of FR and SO, and to capture non-participant spillover, and to recognize that FR may not be a bad thing in a market transformation world (Albert 2009). Spillover (both participant and non-participant types) is recognized especially important for behavioral / education programs as they may have greater potential for SO than

[27] Many advertisers measure the impacts as far as message or adversiting retention, but this should probably not be sufficient for energy efficiency measures because there are so many steps to achieving the ultimate goal of energy savings. They must not only purchase the measure (probably the end of the concern for product advertisers), but must install it, use it properly, and hopefully retain usage. Energy efficiency evaluators may need a higher standard to assure programs are well designed and public dollars are being effectively spent.

[28] A discrete choice model predicts a decision made by an individual (purchase a measure, adopt a behavior, participate in a program) as a function of any number of variables, including demographic, attitudinal, economic, programmatic, and other factors. The model can be used to estimate the total number of eligible households, businesses, etc. that change their behavior in response to a program or action. The model can also be used to derive elasticities, i.e., the percent change in participation or behavior change in response to a given change in any particular (program design, demographic, or other) variable. A discrete choice model, commonly using a logit function, is a mathematical function which predicts an individual's choice based on the utility or relative attractiveness of competing alternatives.

[29] This could be through forms filled out after program delivery, through web surveys, or other approaches with appropriate follow-up to monitor adoption and retention of desired behavioral changes.

other programs. The omission of spillover in California leads to somewhat unbalanced results (Megdal 2008), and that may particularly affect the evaluation results for behavioral programs.

Almost all reports and interviewees agreed that FR assessments should be used in assessing program design and for use in stopping or transforming programs. However, some were concerned about its use as a penalty against utility cost recovery (for instance, Peach 2009); others thought it was very important, and that without a role in cost recovery it would simply be "one of the market impacts" from a program (Mulholland 2009).

NTG, Behavioral Programs, and Applications in Regulatory Tests

Most behavioral and educational programs seem to be treated as indirect programs and not included in regulatory tests. This has a problematic side effect: lack of credits for benefits or savings from these programs probably means we are under-investing in these efforts. In addition, the regulatory aspect of energy efficiency tends to discourage innovation (in order to ensure accountability), locking in place traditional programs / approaches that have a history of passing regulatory tests. As utilities look for best practices in behavior and education, they risk ending up with mediocre homogeneity if all jurisdictions' limit themselves to current best practices without wanting to take risks in experimentation and innovation (Bensch 2009).

In addition, the uses of tests as currently applied are based on geographic boundaries and political jurisdictions designed for a DSM world that does not incorporate the broader effects of climate change. Programs may be needed that do not pass a "local" cost-benefit test – and education / behavioral programs are particularly prone to crossing boundaries (as are the factors influencing these effects) (Bensch 2009).

Many interviewees argued that a modified TRC test was needed, suggesting changes in a number of issues and related policies related to greenhouse gas and NEBs treatment, FR, innovation incentives, and other factors.[30] These issues affect all programs, but current practices can be particularly punishing to behavioral / educational programs (with their strong spillover, hard to measure impacts, and "cross boundary" issues). In addition, arguments were made for regulatory tests at the portfolio level, to address some of these same issues. Some technologies or programs may draw customers in, even if they are not the most cost-effective – and this may be particularly true for behavioral programs. Measure by measure tests can encourage cherry-picking – and the math tends to exclude behavioral and outreach initiatives.

Key Uses for NTG

There are issues with NTG; however, despite these concerns, to quote one prominent researcher in the field, "not measuring is not the answer".[31] Rather, it may be important to consider the uses to which the free ridership, spillover, or net to gross ratios are put. Based on our analyses, however, elements of the NTG measurement are important for the following reasons:

- Free ridership is important to identify superior program design. High free ridership can mean incentives are provided for measures that would already be selected in the marketplace. This can feed back to process evaluation as well as impact assessments. Programs with higher free ridership might benefit (in at least cost-effectiveness terms) from refining outreach, targeting, rebate / intervention levels, or efficiencies of measures.

[30] Interviewees also suggested that that there was inappropriate application of some tests. For example, if energy is the target, the RIM test for residential is not an appropriate test (Sulyma 2009). Interviewees noted that we "do well' for the 1980s using the vanilla cost tests from California or their mild variations, they are not well suited to 2009 and beyond.

[31] Mike Rufo presentation at the 2009 ACEEE Market Transformation Conference, Washington DC.

If programs have high free ridership, or if free ridership has been increasing significantly, it may be time to phase out the program, delete or upgrade measures, or significantly redesign the program.[32]

- Free ridership can help to identify program exit timing. Depending on the program, it may provide a signal that a program may not be needed to induce the desired efficiency behaviors.
- Spillover is a very important metric in assessing the performance of education / outreach / behavioral programs. In fact, spillover is one of the key desired impacts from these types of programs. Ignoring these effects may significantly understate the program's performance in a way that would bias program investment away from education, training, and market-based program interventions basically because some of the impacts are indirect and hard to measure.

There are mixed reports about whether utilities or agencies incorporate a feedback loop from these kinds of evaluation results into program design. Those that reportedly do include NYSERDA, NH, MA (Albert 2009), NEEA (Rasmussen 2009), and BC Hydro (Sulyma 2009), and there may be progress among the California IOUs incorporating some process evaluation findings in association with the 2006-2008 programs. Some respondents commended the power of logic-driven model designs and evaluations (and their applications for linking back), and others felt that honest logic models would incorporate numerous education initiatives at multiple levels in most programs (Albert 2009, Bensch 2009).

Not examining free ridership and spillover *ex post* will make it impossible to distinguish and control for poorly designed / implemented programs, and for programs that may have declining performance over time and may have outlived their usefulness, at least in their current incarnation. In addition, highly successful programs will be overlooked. Some interviewees said 'deemed savings are ridiculous' for this reason. None of these applications necessarily requires precise measurements, although of course, reasonable reliability is needed to provide useful information.[33] To provide the best chance for optimal programs, we need:

- NTG or FR and spillover estimates that are as reliable and precise as needed for the particular use – with greater precision needed for calculation of incentives vs. quasi-quantitative / qualitative[34] uses;
- NTG or FR and spillover estimates that provide replicable results and are based on credible, defensible estimation methods suited to the accuracy needed;
- Methods that provide different levels of accuracy for estimates of NTG, FR and SO at reasonable cost levels;
- Flexibility in the application of NTG, FR, and SO results depending on type of program (whether programs are new / innovative / pilot; "same-old-same-old"; cookie cutter; custom; information-based; etc.); and

[32] On a related note, free ridership is affected by NEBs (see next chapter) and affects program participation. High free ridership may occur because measures have a host of attractive features other than energy saving. Therefore, computing rebates or incentives based on economic payback only may lead to higher rebates than needed to induce adoption of efficient measures or behaviors.

[33] This may mean protocols or minimum best practices are needed. However, it is also important to maintain an environment that allows innovations that develop better measurement practices and innovation.

[34] For example, for determining the direction of program free ridership, the order of magnitude of spillover for program planning uses, and potentially, for screening programs based on spillover (education, training, behavioral, etc.).

- Application of the results in ways that don't discourage the development of new and creative and potentially effective programs, making risk of fiscal investment in programs manageable and reasonably predictable.

A case might be made that the most "accurate" metric is pure *ex post* measurement[35] as these metrics are used in planning and reward purposes. Thus, if the main "rub" arises when NTG elements are part of computations of financial reward or program approval, there are several possible options:

- *Short-term deemed values*: Assigning a deemed value for year 1 or the first two years (to allow for refinement of the program without significant fiscal consequences) and then requiring measurement in year 2 or 3 of a program. A utility / agency can decide whether to drop the program after the first year (or two) if it performs poorly without having to incur the financial penalties. This may help avoid the innovation penalty, and may be suitable for new innovative programs, pilot programs that aren't traditional, etc. True-up at some point is necessary to assure that the field learns about the performance of different program types, and that ineffective programs are not rewarded indefinitely.[36] Deemed spillover values may be especially needed for programs targeted at education.

- *Long-term deemed values*: Allow "deemed" NTG values for well-known program types based on measured NTG from programs around the nation; check the performance perhaps every 3 years; and penalize programs that perform more poorly than the norm,[37] or require program comparisons against "best practices" periodically (every 3 years). Again, periodic true-up is needed. [38] This might be most suitable for cookie cutter / traditional programs.

- *Negotiated options*: For some large, important, or innovative programs, negotiations for a priori values might be used.[39]

Reliable measurement methods are available that suit many program types, but more work remains, including the following.

- The improved NTG, FR and SO methods that have been evolving have shown promise, particularly for the non-financial applications of NTG elements (process, etc.). These include accommodations for partial free ridership, and incorporate adjustments for "corroborating information" (Skumatz and Violette 2004, Cook 2008, KEMA 2008).
- Experimental design including random assignment for participants and non-participants should be used for as many program types as feasible.

---

[35] Assuming, for the purposes of htis point, that "accurate" *ex post* estimates could be developed.

[36] In these cases, if the *ex post* true-up value is different from the *ex ante* deemed value, the *ex post* should be probably be used for all financial reward applications for the years after the first or first two. There should be an investigation into mechanisms that reward innovative programs that perform particularly well to encourage investment in innovation.

[37] In the area, but also comparing to performances around the nation.

[38] In these cases, if the *ex post* true-up value is quite different from the *ex ante* deemed value, a mechansim needs to be developed that will (1) keep risk in investment in programs low, but (2) penalize programs that deliver significantly poorer performance than those found from best practices. However, since a key purpose of this project is to consider the case where market chatter complicates program evaluation, the reward mechanism may need to provide mitigated upside and downside benefits from performance achievements and under performance. Options like this should also be investigated, as well as what programs might constitute "traditional" vs. "innovative" under a moving playing field.

[39] This may cover programs such as those offered to only a very few large businesses (industrial, etc.), for example. This is suggested by the method NYSERDA is implementing for measuring NTG from their custom program that has very few participants (Cook 2008).

- Comprehensive market assessment work provides baselines for purchase decisions and may provide a good source of information for decision-modeling in the absence of programs. Techniques could / should be explored that allow market assessment or saturation surveys that would provide information useful in estimating non-participant spillover. This is important for many training, education, and behavioral programs.
- Consider introducing data collection approaches that introduce a real-time data collection element that piggybacks on program handouts / materials / forms. This assures greater response rates, documents influences closer to the time of decisions, and allows periodic reviews of performance in time to refine programs (Skumatz 2008).
- Use discrete choice modeling approaches more regularly that introduce explanatory variables to help address issues of imperfect control groups, unobserved factors, etc. and allow improved estimates of attributable impacts.

Results on elements of NTG should be accumulated in a database and continuously updated with new research and evaluations to support analysis of overall findings, patterns in results, lessons for successful program designs or elements, and to allow comparisons and tracking.

3.4 Conclusions and Additional Research Needed

Conclusions, recommendations, needed research, and other key issues uncovered in the analysis are detailed below and in Table 3.3.

3.4.1 Conclusions

- The "net" effects are a significant element of the assessment of benefits and costs for a program, computations that in some states can determine the start, continuation, or dissolution of a program's funding.
- Net savings in many states also determine levels of cost recovery, performance incentives, and penalties. As such, estimation of Net to Gross ratios has drawn significant attention in the energy efficiency industry.
- Traditionally estimation of net savings proceeds in a series of steps. Starting with planning gross estimates, evaluators make adjustments based on installation rates, failure rates, baseline assumptions, and possibly leakage. This process produces *adjust gross savings*. Engineering or statistical models are used to make further adjustments producing *verified gross savings*. Finally, through the use of primarily self reported methods, *net energy savings* are computed (see diagram below).[40]
- Net savings' main adjustments include free riders, spillover, and rebound or take-back effects (to a lesser extent).
- Net savings, and its components, have traditionally been addressed through quasi-experimental design, self-reporting through surveys, enhanced self-reporting surveys, qualitative choice models, and straight stipulation (using results from other studies).

[40] Net to Gross diagram inspired by / adapted from training materials prepared by Dr. M. Sami Khawaja.

- Our interviews suggest that a number of utilities consider free ridership (FR), but do not include spillover (also called free drivers) in their analyses of program effects. This asymmetric approach undervalues EE.
- An examination of FR and spillover in 23 states and/or utilities serving states found 15 states (69%) rejected FR in estimating net savings (Quantec 2008). Massachusetts prefers to have utilities focus on market transformation (MT) programs and correct for factors affecting gross to net savings in program design. California requires deemed FR values, but excludes spillover. Estimating NTG is not a priority in Iowa.
- There is little reporting of spillover (SO). Minnesota and Wisconsin publicly stated that FR and SO cancel out (Quantec 2008). Iowa assumes a NTG value of 1,[41] In Illinois, NTG ratios of 0.8 are assumed for low income programs, and are lower for appliances (Baker 2009).
- NTG estimation for upstream and market transformation programs is increasingly more reliant on sales data from program and control areas. Rapid expansion of CFL programs and recent changes in the CFL market have hindered the ability of this approach as a means to provide reliable NTG estimates. Recently, detailed statistical methods involving various geographic areas have been proposed to assess factors contributing to various levels of saturation of CFLs in various service territories. The contributing factors include programmatic components. Data are collected primarily through random digit dialing of customers in various geographic regions.

- There is considerable – and growing - controversy regarding the use of net to gross, particularly in places where incentives, penalties, and attainment of statutory goals are heavily reliant on net savings estimation.
- This heavy reliance on a metric that, in turn, is heavily reliant on subjective methods of estimation is a concern.
- Achievement of national goals of improved energy efficiency and associated climate change implications can be seriously hindered without common protocols for estimation of NTG.
- In regulatory proceedings, the controversy arises from the following main issues: (1) the potential for error and uncertainty associated with these measurements, because of difficulties in (a) identifying an accurate baseline, (b) identifying/implementing a control group, or (c) relying on self responses to a survey; and (2) the high cost.
- Baselines are a very important part of the problem of measuring NTG, FR, and SO factors. Documenting what “would have happened” is the most significant biggest challenge in evaluation, this challenge is on-going and finding a true baseline for outreach/behavior programs will continually be more difficult as EE messages become more prevalent in the US.

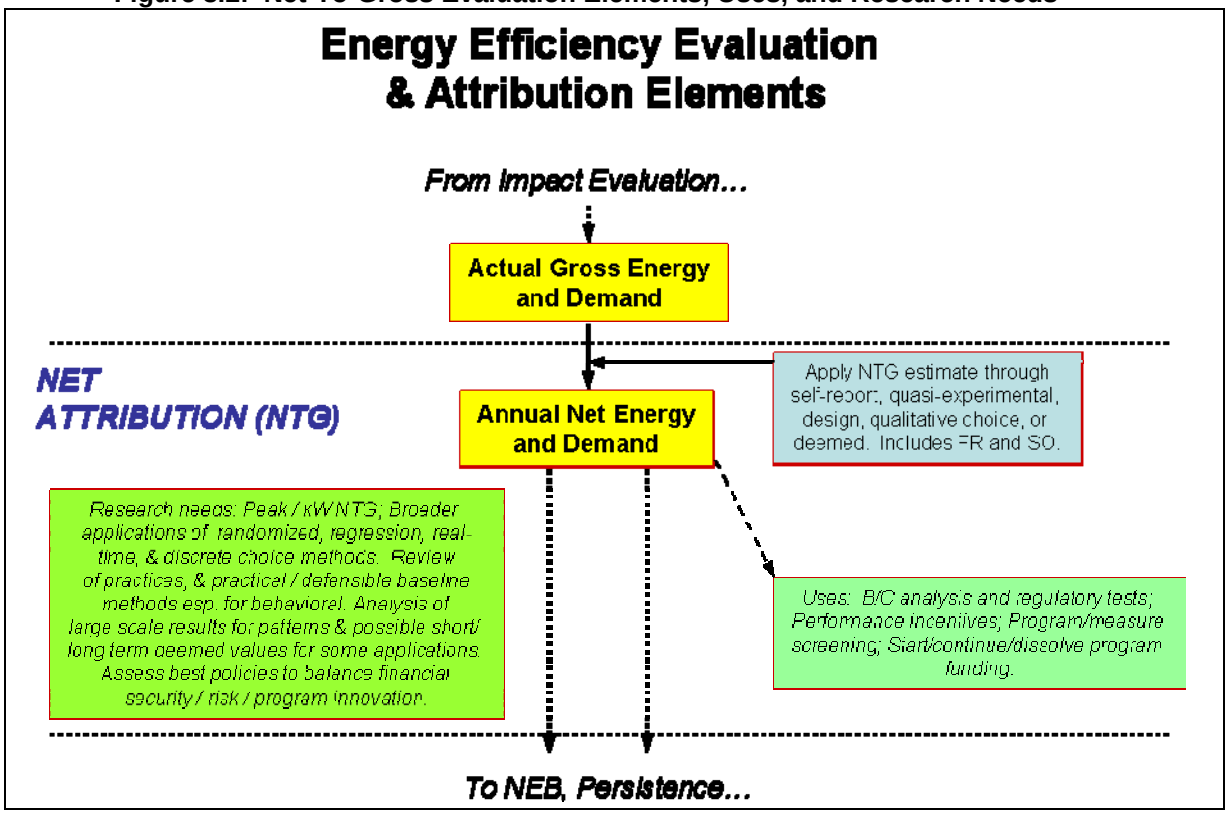
3.4.2 Additional Research Needed

- Detailed regression models from various regions show promise. These models would include explanatory variables describing these various regions as well as the various efforts for upstream programs.

[41] The Iowa study (Quantec 2008).

- Require random assignment for participants and non-participants for as many program types as feasible. This approach may be especially important for outreach and behavioral programs. Train (2009) suggests pairing this with a discrete choice model to predict behavior if they didn't have program available.
- Consider survey designs that introduce a real-time data collection element. This provides several benefits: increases return rate/sample size, provides on-going data significantly reduces the cost; potentially provides more accurate data, and possibly helps sort out which programs had what degree of influence. This may be suited to education programs as well as "widget" programs. This could be through forms filled out after program delivery, through web surveys, or other approaches with appropriate follow-up to monitor adoption and retention of desired behavioral changes.
- Another approach that may work for some programs is to use discrete choice modeling approaches. These methods introduce explanatory variables that help address issues of imperfect control groups, unobserved factors, etc. and support higher quality estimates of attributable impacts.
- Free ridership is important to identify superior program designs or exit timing. Programs with higher free ridership might benefit (in at least cost-effectiveness terms) from tweaking of outreach, targeting, rebate / intervention levels, or refinements in which (efficiencies of) measures are included in the program.
- A paper analyzing which version of computing NTG ((1-FR+SO) vs. (1-FR)*(1+SO)) is more appropriate or justifiable might be useful.
- Tests and comparisons of free ridership and net-to-gross results between standard NTG self report surveys, enhanced approaches that incorporate "corroborating" questions, and data collection approaches that introduce a real-time data collection element (piggybacking on program forms) would be valuable to determine if these enhanced approaches provide improved estimates of NTG.
- More research on baselines and demonstrations of feasible options would be valuable.
- Greater use of comprehensive market assessment work shows promise for providing baselines for purchase decisions, and should be further explored as a standard technique for understanding participant decision-modeling in the absence of programs. In addition, market assessments may provide information critical to the estimation of non-participant spillover, which is a key potential (and desired) outcome for training, education, and behavioral programs.
- A comprehensive analysis of free ridership, spillover, and net-to-gross results to identify whether there are consistent patterns in results for "types" of programs or measures, patterns related to incentive designs, by analysis method, or other lessons that may suggest whether short or long term "deemed" values can be justified for some types of programs. A searchable nationwide repository or database of NTG, FR, and SO results would facilitate this objective.
- When it comes to the dynamic retail sector, it may be impossible to predict what they would have done without the program especially if changes occur upstream, and it needs further research.

Figure 3.2: Net-To-Gross Evaluation Elements, Uses, and Research Needs



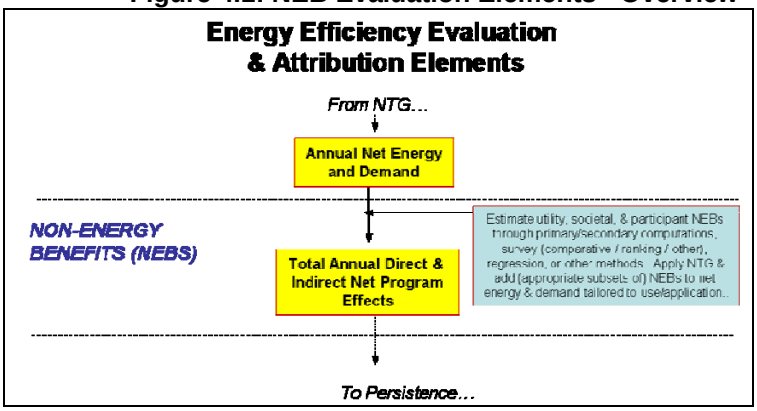
4. NEBS – NON-ENERGY BENEFITS / IMPACTS

Non-energy benefits (NEBs)[42] or non-energy impacts (NEIs) are generally defined as any real or perceived, financial or intangible benefit accrued by an energy efficiency project. They are effects that are omitted from traditional energy program evaluation work, which focuses on impacts on energy savings.

Given their more indirect nature, NEBs are generally relatively hard-to-measure (HTM)[43]. As a consequence, they may also tend to be prone to higher levels of uncertainty than some other measurements associated with energy efficiency programs. The level of effort spent on measuring (or better, *estimating*) these effects should presumably be somewhat proportionate with their potential impact in helping to avoid wrong decisions about programs or EE interventions.

4.1 Background

Figure 4.1: NEB Evaluation Elements - Overview



Strictly speaking, NEBs are "omitted program effects" – impacts attributable to the program, but often ignored in program evaluation work. After years of research, more and more utilities and regulators are considering these effects in program design, benefit/cost analysis and marketing.

Over the last 20 years, a wide range of NEBs has been identified in studies.[44] Early publications focused on enumerating potential categories of benefits or theoretical discussions (e.g., Mills and Rosenfeld 1994, Flanagan 1995), but quantitative work was scarce. The early work in NEBs was applied to low income programs because effects beyond energy savings were commonly included as part of the list of goals for these types of programs.

One difficulty in early studies was that all the benefits were computed using data from secondary sources, which severely limited the array of benefits categories that could be estimated or attributed to the effects of a particular program.

[42] Non-energy benefits (NEB) have been called non-energy benefits, non-energy effects, non-energy impacts, indirect effects, and other terms. The first major term applied to the research was "non-energy benefits" (NEBs). As long as we understand the definition – largely that both positive and negative effects are implied -- the term NEBs will be used in this paper because it assures that the historical literature is not lost, and appropriately retains priority naming rights to the originators of the concept.

[43] Megdal associated this "hard to measure" language with NEBs in several paper (Megdal 1999.).

[44] A detailed literature review covering more than 300 studies is included in TecMarket Works, Skumatz, and Megdal (2001). Versions are included in earlier studies including the following: Skumatz (1997), Skumatz and Dickerson (1998), and Weitzel and Skumatz (2001).

Categorization, Causes, and Uses of NEBs

Starting with work in the mid-1990s, the literature began to explore more consistent measurement methods, and sort these benefits into three "perspectives" based on the beneficiary of the effect – utility / agency; societal; and participant.[45] Each is described in more detail in Table 4.1. In addition, the table presents information on current and potential uses for NEBs.

Table 4.1: Summary of Three Perspectives Accruing Non-Energy Benefits / Effects[46]

	Overall Description	Key "Drivers" / Sources for Effects	Specific Examples	Uses / Applications[47]
Utility / Agency / Ratepayer Effects	These are incremental positive or negative impacts from initiatives that affect ratepayers and utilities and reduce revenue requirements. These effects are generally valued at utility (marginal) costs. These effects vary by type of participant (residential, low income, commercial) by overall energy savings and peak/non-peak timing and other factors.	• Payment / financial burden • Debt collection efforts • Emergencies / insurance • T&D, power quality / reliability • Subsidies / transfers	(Changes in) bad debt written off; Changes in carrying costs on balances; Labor and other changes from changes in bill- and collection-related calls / activities; Changes in shut-offs / reconnects; Changes in line losses from power through lines; Outage frequency / duration	Current: Few. Some used to suggest targeting of bill-payment problem customers. Potential: Regulatory tests (e.g., program administrator cost).
Societal Effects	Incremental non-energy impacts from initiatives that (positively or negatively) affect the greater society or that cannot be attributed directly to utility/ratepayers or participants. These effects are valued as appropriate to the benefit category. These effects vary significantly based on local economy, generation mix, peak / non-peak program effects, and other factors.	• Economic development / job creation multiplier effects • Environmental including emissions • Health • Tax impacts • Water and other resource use • National security	Economic output changes; job creation; changes in greenhouse gas (GHG) emissions; infrastructure savings for energy, water, waste water, etc.; fish and other environmental effects; assessment of energy vulnerability.	Current: A few utilities and agencies use deemed multipliers for GHG emissions or avoided environmental effects. At least one agency (?) presents fraction of environmental and economic benefits as part of "scenarios" for B/C tests and portfolio analysis. Potential: Regulatory tests (e.g., total resource cost (TRC))
Participant / "User" Effects	Incremental non-energy effects from initiatives that benefit or affect the participant users of the energy efficient equipment beyond energy or bill savings. These effects are	• Payments and collection • Education • Building stock • Health • Equipment service / productivity	Change in ability to understand / control energy usage; changes in ability to pay; changes in time spent on bill payment / collections issues; changes in interruptions in service (shutoff, etc.); changes	Current: Program marketing (limited), project screening (limited), scenario analysis (limited); some in modified TRCs when NEBs

[45] Initiated in Skumatz (1997) and repeated in Amann (2006).
[46] Or "beneficiary" categories.
[47] The useage information was augmented by a very useful preliminary paper on NEBs provided by Mallory (2008).

	Overall Description	Key "Drivers" / Sources for Effects	Specific Examples	Uses / Applications[47]
	valued in terms relevant to the participant. These effects vary by user and by program / initiative (e.g., specific measures installed, education / outreach program).	(comfort, maintenance, etc.) • Other utilities / resources (water, etc.)	in other bills (water, etc.); changes in property value; changes in health effects; direct / indirect changes in energy "service" and stream of associated income / utility / satisfaction (productivity, comfort, light quality / quantity, noise, maintenance, lifetime, reliability, etc.), and other (e.g., "green" effects.	readily measurable. Potential: Portfolio development, program refinement, marketing, regulatory tests (e.g., participant cost).[48]

Considerations for Appropriate Attribution of NEB Impacts

While there are measurement issues associated with estimating HTM effects like NEBs, credibility also suggests that a basic methodology be considered in assessing and attributing NEB effects to EE interventions which accounts for the following issues:

- **Redundancy in sources or categories**: Similarly-named benefits can arise in multiple perspectives without being redundant. For example, fewer billing-related calls to a utility save money and time for both the utility and the household making the call. These are distinct impacts.[49] Of course, each needs to be valued in terms appropriate to that beneficiary.

- **"Net" Positive and Negative**: Non-energy benefits or non-energy effects may be positive or negative, and the "net" effects may also be positive or negative. Negative benefits can be interpreted as barriers in some applications (see discussion below).

- **"Net" of standard equipment choices**: When NEBs are applied to energy efficiency programs, it is critical that the impact be measured above and beyond the base of what would happen without the program – specifically, the (presumably, standard efficiency) equipment that would be selected without the program.[50]

- **"Net" of free riders**: To the extent that the interest is in NEBs that are attributable to the program above and beyond what would have happened without the program, the NEBs would have a free ridership (and potentially spillover) factor applied.

- **Minimizing Overlap / Double Counting**: The drivers for NEB effects tend to emanate from a limited number of key impacts associated with energy-efficient equipment.[51]

---

[48] Ibid.

[49] This is not double-counting benefits – rather, it recognizes that some effects have multiple beneficiaries, and each is valued by the appropriate tailored valuation method. The utility benefits would be valued at the utility marginal wage rate for customer service staff, and the household would have the same amount of time valued at the minimum wage, leisure wage, or some other appropriate value. Benefits are recognized and realized by both groups. However, whether either or both of these benefits are included in the ultimate sum total of the NEBs for the research or calculation depends on the purpose of the research. Elements (e.g., utility impacts) may be included in direct program-related benefit cost work, but these computations would likely ignore the participant impacts. Analysts choose the appropriate NEB categories based on the purpose of their research, or the appropriateness to their decision-making objective (Skumatz 1997).

[50] For estimation work based on survey responses, it is important to ask about NEBs not between a household or businesses' OLD equipment (which they may be replacing), but between the standard efficiency NEW equipment that they would otherwise purchase vs. the higher efficiency new equipment that the program promotes (Skumatz 2002).

[51] See the third column of Table 4.1 above for the types of factors to which we refer.

Multiple, closely related benefits and impacts could be measured, but it is likely the individual benefits would be able to be separately measured or valued by the participant, and thus, by the researcher. Too many categories of impacts exacerbate the problem of overlap and double-counting.[52]

4.1 Current Practices, Measurement, and Use

4.1.1 Utility Perspective NEBs – Measurement Methods

The vast majority of initial work on NEBs in the 1990s focused on the utility perspective, particularly addressing topics related to arrearage changes from low income programs.[53] Significant impacts were attributed to the programs. The estimated impacts in this literature ranged from 0% reduction to 90% reduction in arrearage balances. The average value for these studies was 26% reduction, and the median for programs not targeted at customers with bill payment difficulties was 18%. Valued for the utility at carrying charges, these arrearage effects were small for each participant. When compared to the values associated with other benefit categories from the societal and participant perspective[54], the arrearage and debt / financial benefits from programs represented a small fraction of overall NEBs. Limited work may still be proceeding on these impacts on a program-by-program basis (e.g., low income programs), but they are generally fairly program specific, have fairly clear measurement approaches, range within limited bounds, and generally are not making it into the literature.

However, there are a fair number of utility-perspective NEBs that are not being addressed in the literature – probably because they can be difficult to estimate – and some of these may have significant weight and value. Additional research would be beneficial. These NEBs include:

- **Line loss reductions**. These may be very important and valuable and are relatively easily measured.[55] Some utilities have, in the past, used rules of thumb for this loss that are fairly high. If these are reflective of fact, then they represent an additional adder to EE programs that has significant value. For example, transmission line losses may be 2% and distribution losses may be 4.5% for a total of 6.5% (NWPPC 2001). These losses may vary by time of day. Additional research on this point may be valuable in computing a total savings associated with specific EE programs or portfolios.
- **Time of day / capacity impacts / avoided infrastructure**. These are potentially quite large and very important, and are relatively easily measured. However, it may be that the estimates associated with demand response programs may currently be considered direct

[52] As discussed above, early work on NEBs focused on developing laundry lists of possible impacts / sources. Care is needed in defining the specific NEBs measured within these categories to minimize overlap and double-counting. For instance, owners may have difficulty separating out labor changes from maintenance benefits and might assign a value to each and possibly double-count at least a portion of the value. However, pulling back and focusing on non-overlapping "drivers" (as listed in the table above) as sources may help alleviate some of this issue. In addition, asking about total values can be used to "normalize" individual categories of impact, reducing the potential overestimate of impacts from a pure "bottom up" valuation approach (see Skumatz and Gardner (2002)).

[53] A previous literature review by the author (Skumatz 2000, later inlcluded in TecMarket, SERA, and Megdal 2001) reviewed about 30 arrearage studies.

[54] See Skumatz 2001 and TecMarket Works, Skumatz, and Megdal 2001.

[55] Certainly there are engineering factors available, and factors like average utility line length per customer or similar numbers can be used. The next level of sophistication could be peak vs. non-peak, and ultimately hourly dispatch estimates. See the parallel discussion in the section on societal impacts from GHG emissions that is in the next section of this report. Again, the degree of sophistication (and related cost) needed depends on the how the results will be used.

energy impacts, rather than NEBs. There are effects associated with a wide array of programs, and these indirect benefits are valuable in reducing costs associated with building capacity that can be avoided from well-designed or specifically-targeted EE programs.

- **Safety and Health-related impacts**. Utilities may save significant funds in insurance and liability costs from safety-related effects that result from audits and inspections associated with many EE programs. Potential health effects may also be reduced by EE program efforts.
- **Other**: To the extent that the utility can avoid other future risks or liability claims due to the efforts of EE programs or to the avoidance of generation, these are beneficial to the utility and its ratepayers at large in terms of reduced revenue requirements. These effects have not been studied.[56]

In addition, the utility NEBs have tended to focus on entirely on the "energy" and have not considered the extra NEBs associated with the "peak" or "demand" impacts from energy efficiency programs. There are several utility perspective NEBs in which this could be quite important and valuable. Capacity impacts or avoided infrastructure NEBs would potentially be much higher for programs with peak impacts, and line loss values presumably differ by time of day and by season (temperature). How important these factors are has not yet been explored.

4.1.2 Societal Perspective NEBs – Measurement Methods

There has been real progress in this area of NEBs research, the impacts appear to be significant, and measurement of some of these impacts (from both measure-based and behavioral programs) has interest outside the traditional evaluation literature and applications (e.g., climate change, stimulus remedies).

Climate Change

Energy efficiency strategies can provide environmental benefits to the region and to society, particularly due to their role as a pollution abatement strategy. Early studies evaluated the benefits in terms of helping to meet Clean Air Act goals, reducing acid rain, and a variety of other environmental benefits and their associated health effects.[57] More recent work focuses on quantifying the impacts in terms of metric tons of carbon equivalent (MTCE) or metric tons of carbon dioxide equivalent (MTCO2E). These "stand in" for the array of emissions, and,

[56] Many of these effects may be parallel or related to the effects listed under the societal perspective. To the extent public health suffers from generation or EE programs or other activities, the utility may end up paying a judgement some day. That would represent a utility NEB (positive or negative) and benefit (or harm) the ratepayers. It is nearly impossible to judge the sources of those risks *a priori*, but as standards of business ethics and practices change, liabilities change. Could printers know their inks would later contaminate sites and cause Superfund cleanups? Careful study of possible sources of these kinds of risks may have merit.

[57] Literature summary is based on Skumatz, 1997; Skumatz 2000; and TecMarket Works, Skumatz Economic Research Associates, and Megdal and Associates, 2001). In particular, a` number of these concepts are addressed in early work including Ottinger et al. (1990), Consumer Energy Council of America Research Foundation (1993), and Galvin, Enbridge, and Woolf (1999). Brown et al. (1993) developed quantitative estimates of these benefits relative to the low income weatherization assistance program. Brown attributed a net present value of $172/household (1989 dollars, discounted at 4.7 percent over 20 years). The Northwest Power Planning Council (NWPPC) (Harris 1996) provided policy guidance to utilities in the area regarding valuing the benefits from conservation relative to new power. The NWPPC historically assigned a 15 percent "adder" for environmental benefits associated with conservation programs, applied to the avoided costs of the program.

depending on the monetization factor selected, can represent the value of the associated harmful effects from the emissions.

With the signing of the "Proposed Endangerment and Cause or Contribute Findings for Greenhouse Gases…" by the US EPA on April 24th 2009, the EPA has officially stated that "the case for finding that greenhouse gases in the atmosphere endanger public health and welfare is compelling and, indeed overwhelming." The ruling proposes that the six major greenhouse gasses be covered under the Clean Air Act, giving the federal government the authority to regulate the emissions of these gasses due to their imminent threat to human health, the environment, and the US national security and well-being.[58] This provides a strong basis for considering at least some non-energy benefits in program design and planning, and for the measurement of at least some non-energy benefits in regulatory arenas. The potential for cap-and-trade credits, the enormous stimulus package (much of which is directed toward the environment and especially to energy efficiency and job creation), and new attitudes in Capitol Hill bolster the need for the measurement of key societal non-energy benefits in association with energy efficiency programs.

Displaced GHG Emissions through Energy Efficiency – Simple vs. Complex Measurement Approaches

Measuring these impacts can be fairly simple, or can be intricate, depending on the degree of complexity and tailoring selected. Direct stack measurement or plant by plant analysis can be extremely expensive. However, secondary information is available on several critical components that can be used to derive estimates fairly directly. For example, factors for the air emissions per kWh from a variety of fuel sources and/or age of generation plant by type are available from a number of studies. This allows tailoring of the estimates of emissions avoided from a program by selecting the fuel mix for power avoided during peak times, or a different fuel mix based on power avoided off peak / base load and multiplying by the appropriate number of kilowatt hours saved.[59] Many studies list multiple pollutants or GHG constituents, many of which can be valued based on calculated risk, regulatory values, or cap and trade values.

Using adopted or average numbers is an important simplification. Without this step, the valuation of environmental benefits becomes extremely complicated. The value of an additional ton of constituents of ozone would be dramatically different depending on air shed, time of day, number of persons in and near the air shed, quality of air starting out, and numerous other factors, resulting in prohibitively expensive research on a utility-by-utility or state-by-state basis.

The literature has focused on three strategies for estimating the emissions impacts:

- **System Average:** the least expensive method, and as with many other least expensive methods, the least reliable. Under this approach, a system wide grid average is used for the local, regional, or national grid, and emissions factors per MWh are estimated. This may be the lowest cost approach, however, it allows for the greatest level of uncertainty in emission impacts. It also masks potentially important differences between peak / off-peak programs.

[58] The original Supreme Court case overturning a lower court ruling stating that the EPA could not regulate GHGs (*Massachusetts v. EPA*) was based on vehicle emissions; however, the EPA proposal is expected to have large reaching implications going well beyond vehicle emissions.
[59] See Woolf (1999).

- **Margin Operations:** used to look at the potentially displaced emissions for on-peak and off-peak hours, different seasons, and shoulder months.[60] This method uses different emissions factors for off- and on-peak hours, and considers that EE impacts will most significantly affect the marginal energy producers, or the plants that come on last at high demand periods. These plants – and associated emission factors - may vary depending on the season.

- **Hourly Dispatch:** the most detailed and most accurate method for calculating GHG emissions displaced. At the same time, it is the most expensive analysis to complete. In this method, evaluators look at the individual plants and calculate emission for each plant for each hour. Determining the displaced emissions requires complex modeling of energy reduction over the entire grid and may include such calculations as the displaced emissions of building a new plant now, compared to in the future, when the plants may be more efficient.

While none of the three methods has yet distinguished itself head and shoulders above the rest as the accepted measurement method, evaluators seem to agree that the second two methods are preferred. The first is too simplistic for most uses, and the second requires only marginally more information for a far more robust and refined outcome.

Issues Complicating Use of GHG Emissions Avoided from EE/RE in Cap and Trade and Other Applications

Typically, in energy, it is not necessary to consider the locality or the specific source of the energy savings reductions within a utility territory. Evaluators are able to report the net impacts overall, regardless of exactly where the specified energy savings are originating. However, when it comes to GHG reductions, the exact source of the associated reductions becomes integral. If the reductions occur in an area with noted smog problems, it could influence the evaluation of – and particularly the valuation and the associated harm with - the displaced emissions.[61]

If we want to value the emissions in the market in preparation for cap and trade, auctions, other trading arrangements, or for verifying credits for GHG emissions, there are three key problems that must be "solved" or resolved for improving the credibility of energy savings computations and associated emissions.[62]

- **Additionality**: Additionality has been reported as one of the main potential stumbling blocks in attributing GHG emission reductions. Parallel to free-ridership, in GHG measurement, additionality refers to emission reductions that are attributed to a program beyond those which would have occurred without the program's presence. This issue may become more prevalent as regulators begin to think about cap-and-trade programs and start to set limits on emissions. If a utility is mandated to reduce emissions below

[60] The State of Wisconsin's Focus on Energy program's "middle ground" is a good, and well documented, example of this approach (Sumi 2009).

[61] On a health basis, the local air shed is critical. However, the industry currently seems to be treating a MTCE as a MTCE rather than associating specific values with health benefits. As the market matures, or as auctions arise, this may or may not change.

[62] The problems associated with these topics are addresssed in many papers. See, for instance, Price et.al. (2004), Dickerson and McCormick (2005), Schiller, Vine, and Prindle (2005), Sumi, Bloch, and Erickson (2005), Sumi, Ward, and Hall (2007), Nemtzow and Siddiqui (2008), Sumi and Ward (2008), and others. Solutions have rarely been discussed in the papers. [DOE's NAPEE (National Action Plan for Energy Efficiency) Guidelines address this, according to Ed Vine, Lawrence Berkeley Laboratories]

level x, and an EE program reduces emissions to that level, the question of double counting and who gets to count and own the displaced emissions becomes important.

- **Program vs. project**: The issue of whether to measure a *program* or a *project* has also been cited in much of the literature regarding GHG attribution. Generally, a single *project* such as an office audit and retrofit will not result in large avoided emissions, and the evaluation may be costly. Looking at an entire group of similar projects, or completing a *program* evaluation using a sample of projects, may be more cost effective and result in higher quantifiable emissions reductions.

- **Error, Uncertainty, and Risk**: Estimates of energy savings associated with energy efficiency and renewables strategies will have a component of error. These errors may be lower with renewables, as the comparison is "no plant". Energy efficiency represents a more complicated situation as the savings estimates are affected by baseline estimates, potential behavioral influences, etc, and in this case, uncertainty is a relevant term to use. Uncertainty estimates might be discussed in terms of confidence intervals around savings estimates, or as a subjective assessment based on the risk to the trading program associated with over- or under-estimated savings. Others recommend Schiller et al. (2005) recommend *"… uncertainty levels be defined to be within certain confidence limits at the program or portfolio level. The confidence limits can be used to discount, if applicable, the allowances from an energy efficiency project. The optimum level of M&V varies by project and program and is that which finds the proper balance between uncertainty and cost – too much of either can result in an unsuccessful trading program."*

Note that while nearly a dozen papers in the field list and define these issues,[63] none have been in a position to resolve the issues described. This will largely have to await international discussion.

In the meantime, for the purposes of the estimation of NEBs for program and planning uses – NOT for carbon trading – the peak / non-peak and hourly dispatch models provide suitable methods, and there are reasonably-reliable models for use in developing the desired estimates.

In most cases, periodically updated "deemed" factors (potentially ranges) for each generation fuel, and potentially categories of vintage of plant[64] will provide a suitable method to estimate emissions. Applying these deemed values to programs would require assigning the program shares of "peak" vs. "non-peak" generation fuel mixes by utility or territory. For most program evaluation decision-making and uses, this level of detail will suffice, and it is not clear the payback from more enhanced modeling is needed and that it would balance the time and effort spent debating derivations, factors, and models. Based on preliminary research, where variations in emissions impacts on the order of 7% or 14% or less do not affect the direction of the findings (Sumi et al. 2009), the enhanced modeling is not needed. For high value applications, more enhanced (hourly dispatch) modeling may be justified.

In summary, GHG impacts have typically been treated as a qualitative, not quantitative effect, but the computations can be valuable for cost-benefit analysis and in cap and trade programs

[63] For example: Price et al (2004); Dickerson and McCormick (2005); Schiller, Vine, and Prindle (2005); Sumi, Bloch, and Erickson (2005); Sumi, Ward, and Hall (2007); Nemtzow and Siddiqui (2008); Sumi and Ward (2008); and papers by Vine et.al. (2003), Vine and Sathaye (1997), Vine and Sathaye (1999), and Vinee and Sathayee (2000), among others.
[64] Or where available, actual emissions.

(Sumi and Ward 2008). Sumi, Bloch, and Erickson (2005) also suggest that the GHG results are appropriate to use for cost-benefit assessment of other programs, and that they provide an avenue to balance long- and short-term goals of a project. Nemtzow and Siddiqui (2008) suggest that one additional benefit from EE is the deferral of the need for new generation, and that newer generation sources may be even more efficient (and better for emissions) than current sources / options. Stolarski et al. (2008) suggest the development of a revised regulatory strategy that recognizes environmental benefits. Raynolds (2004) concludes that "…keeping climate change 'in the closet' is shortchanging the political debate and perpetuating the misconception that aggressive policies to reduce greenhouse gas emissions in this country would be 'too expensive' and a drag on economic growth." He notes that although financial savings and localized ancillary benefits of EE are more appealing to top-level decisionmakers than GHG reductions, he recommends that advocates embrace GHG emissions in the myriad relevant arenas as a compelling basis for new and more aggressive energy efficiency policies and programs.

Work by Vine (Vine et.al. 2003, among others), looks at the protocols available and discusses what may be sufficient to support international GHG trading scenarios. It describes the International Performance Measure and Verification Protocol (MVP) (and its options), and EPA's Conservation Verification protocols, and also provides a number of recommendations about key "next steps" for research for this area of work to be taken seriously for carbon trading. The recommendations include: develop a Best Practices template or library of research; independent review of policies on discounting savings to limit abuse; definition of roles and responsibilities of third party verifiers to assure transparency; review of MVP and standardize methods to the extent practicable and appropriate; examine the expenses of evaluation versus the resulting budget impacts on project size to assess the proper balance; explore the treatment of uncertainty; and work to apply similar rules for all climate mitigation projects to assure all projects are treated on a level playing field. Finally, the study acknowleges that, in addition to direct energy savings from programs, indirect effects, and the issues of net-to-gross and market transformation are important to assess in association with meaures and services.

Economic Development

Job creation and economic development benefits accrue as secondary benefits from energy efficiency programs. These benefits include increased employment, earnings, and generated tax revenues; increased economic output; and decreased unemployment payments. Energy efficiency is a key job creation engine, and a short- and long-term driver for the economy. This has been reflected nationally through the Administration's American Recovery and Reinvestment Act (ARRA, or commonly known as the "stimulus package"[65]).

A flurry of early work on this topic in the mid-1990s showed strong economic impacts associated with energy efficiency programs.[66] Recent work in the field relies largely on input-output models

[65] The language for the $3.2 billion for the Energy Efficiency and Conservation Block Grant (EECBG) Program, authorized in Ttile V, Subtitle E of the Energy Independence and Security (EISA) Act of 2007, and signed into Public Law (PL 110-140) on December 19, 2007 specifically states that the Act works to reduce reliance on petroleum through increases in energy efficiency. [Jobs is key! Why quote petroleum? Use a reference to jobs.[

[66] A summary of early work (pre-2001) in this field was included in Skumatz (2001), reproduced in TecMarket Works, Skumatz Economic Research Associates, and Megdal and Associates (2001). It summarized work by Pigg and Dalhoff (1994), Dalhoff (1996), Brown et al. (1993), and Harris (1996). The results found high variation between the results; the literature at the time was not as developed as it is now. Later work (Skumatz 2001) noted that some of the early estimates were overstated because they did not provide "net" estimates – netting out the job and economic effects associated with the activities upon which the money would otherwise have been spent (e.g., electricity generation, consumer price index (CPI), or other bundles). This oversight has been corrected in nearly all later work.

– most commonly and cost-effectively using credible, vetted models available from third-party vendors that support estimation to the county, state, or national level.[67] The estimation work requires running a "base" and "scenario" case, specifying the industries in which dollars will be spent incorporating the energy efficiency program, and comparing the results to the base case. For the base case, the literature tends toward two main schools of thought: (1) the money is transferred from electricity generation expenditures into the EE program industries; and (2) the "base case" industry mix mimics the consumer price index market basket, because the EE funds come from public goods charges.

These estimated economic effects may be positive or negative, although energy efficiency programs are generally more labor intensive than electricity generation. Exceptions to the case of a positive economic impact might include the following:

- Cases in which the program's measures are manufactured outside the territory being considered, but electricity generation happens locally. This would be similar if renewables components are built overseas.
- Load shifting programs, where the same energy and equipment is generated and used, but used at different times. These might be estimated as zero economic impact, or one might add the labor associated with the labor intensity for the CPI market basket of goods that the consumers might purchase with any associated bill savings.
- Behavioral programs encouraging lower usage, without changing measures, with the tradeoffs similar to the previous case.

This measurement approach has become fairly common and can be applied fairly easily to a wide variety of programs in energy efficiency and renewables. Furthermore, there exist a limited number of widely available, accepted models. Assuming underlying modeling assumptions are documented and defensible, the results are relatively easily replicated and compared. Thus, estimation of these results is fairly reliable and consistent.

A review of recent literature finds more than a half-dozen studies published since 2000 that focus on estimating economic development impacts.[68] The studies illustrate several key points:

- There is a considerable range in the estimated multiplier results; however, given that impacts vary by program and territory, some variation is to be expected. More work is needed to compare and verify results, and identify and confirm logical patterns in results.
- All energy savings and all programs are definitely not equal when economic impacts are taken into account.
- Economic impacts need to be estimated separately for each program (type) and locality. Economic impacts are local, and "deemed" values are unlikely to be well suited to estimating program impacts.[69]

[67] Some projects with higher funding levels are developing more locally-tailored models that may address specific sub-areas or provide more granularity at the industry level].

[68] See Geller, Bernow, and Dougherty (2000), Skumatz (2000), TecMarket Works, SERA, and Megdal (2001), Mulholland, Laitner, and Dietsch (2004), Josephson et al. (2004), Imbierowicz and Skumatz (2004), and Imbierowicz, Skumatz and Gardner (2006). Results range from a multiplier of 3.54 for national expenditures on EE (Mulholland, Laitner, and Dietsch 2004) to a multiplier of 0.25 for appliance replacement programs (Imbierowicz et al. 2006). In Oregon, one MW saved increases output by $2.2 million [Add Reference].

[69] However, it is possible that regulatory agencies may want to designate acceptable third-party models in order to reduce arguments. There is a considerable range in the estimated multiplier results; however, given that impacts vary by program and territory, some variation is to be expected. More work is needed to compare and verify results, and identify and confirm logical patterns in results.

Theoretically, however, modeling procedures are fairly simple, and credible models are available. This is an area in which impacts could be measured and included / analyzed fairly readily and with a fair degree of confidence, and the metrics could be used to:

- Select (or craft) measures, programs, or portfolios with greatest impact on the local or larger economy;[70]
- Provide credible estimates of auxiliary benefits associated with programs, that may (or may not, from a policy point of view) be included in benefit-cost tests for program planning and selection.

Other Societal Benefits

- **Health and Safety (H&S)**: Not much has been published on health-related NEBs since 2001.[71] Risks from weatherization and other "building tightening" programs include risk from carbon monoxide exposure. Brown (1996) provides some early assumptions and computations of the associated risk. However, the only work measuring incidences related to safety impacts is Blasnik (see Blasnik 1996). One of the most interesting studies on this topic is Fisk (2000 and others). His study contains results that have implications for the societal and the business / household / participant perspectives. He notes that IAQ and the indoor environment affect the prevalence of common health effects, and examines impacts on costs of the illness directly, as well as on employee leave and productivity issues. He develops dollar values for the national productivity gains from improved IAQ[72], considering impacts from communicable illnesses, sick building syndrome, and direct impacts on human performance (including impacts from thermal environment, lighting, and IAQ), He suggests that key measures that might trigger these improvements include: lighting, air economizers, heat recovery, nighttime pre-cooling, operable windows (vs. fixed), insulation, and thermal windows.

  Other than these works, health and other risks associated with other indoor air constituents have not been well researched, and there is a lack of literature assessing dollar impacts of health care (or incremental changes) for changes in chronic or other illnesses associated with energy equipment or indoor air quality (IAQ). This is a potentially important topic, but the research is expensive, generally requiring detailed data on program measures or interventions with health-related effects, and detailed data on pre-post or test/control groups. However, even with these data, it is difficult to make generalizations about health effects associated with programs because of the variety of measures (and behaviors) and the strong potential for interrelated and compounding effects. These effects make energy savings estimation and modeling work difficult. The

[70] And in the short run, identify progarms that may be best suited to "stimulus package funds".

[71] A brief summary of Health and Safety (H&S) literature review is based on Skumatz (2001), TecMarket Works, Skumatz, and Megdal (2001), Brown (1996), and Blasnik (1996). Although early work exained cost per "crisis", number of avoided crises, and other metrics, the traditional interpretation understates health benefits from programs; for example, it does not incorporate the benefits of reduced illnesses, hospitalization, lost income, and quality of life issues related to weatherization programs. Negative impacts may also arise from the program. For example, indoor air quality issues may develop, and it would be most appropriate to consider and compute the net benefits associated with these impacts. Generally, the steps involved would be to develop: (1) the estimated likelihood of crises in eligible households, coupled with an assumption that all carbon monoxide risk, for example, for these households would be eliminated, and (2) the value of the crisis avoided. If this method were to be applied in developing estimates of health and safety effects from energy efficiency programs, it would be appropriate to assume that somewhat less than 100% of the health and safety incidents would be avoided.

[72] Potential annual savings and productivity gains of $6-14 billon from reduced respiratory disease, $1-4 billion from reduced allergies and asthma; $10-30 billion from reduced sick building syndrome symptoms, and $20-$160 billion from direct improvements in worker performance that are unrelated to health (Fisk, 2000).

challenge of taking impacts from individual measures and trying to add them up to provide credible estimates of health effects is daunting unless it is conducted on a program-by-program test/control basis, or the impacts are provided as a "bounding value"[73] rather than an estimate. Taking the leap from these (personal) impacts to the societal impacts of these illnesses on hospital infrastructure needs and insurance rates (the societal reflection of these impacts) is important, but even more problematic and complex. Some effects are reflected in insurance tables – like fire deaths and property damage – and to the extent that these effects can be traced to program measures, credible (partial) H&S estimates can be developed. However, asthma and other chronic diseases may be exacerbated (or improved) by EE design and measures, and these effects may well be very important. At this time, the estimation work needed to monetize these effects does not exist. Given concerns from builders, architects and engineers, and occupants about sick buildings, asthma, and other issues, it is likely valuable to conduct research to estimate the level of these risks sooner rather than later. If large, it should be addressed and mitigated; if small, that fact can be widely disseminated in marketing materials to alleviate fears about EE measures.

- **Water**: NEB impacts on water saved have been analyzed at a household or business participant level (especially in association with clothes washer programs), and estimates of water saved per measure installed are available. Behavioral impacts will have an effect on these estimates and provide interesting programmatic opportunities; but from an accounting and water savings point of view, the estimate of these impacts is almost trivial. The infrastructure impacts related to the deferral of new plant or treatment facilities or other societal impacts have <u>not</u> been studied. In many areas of the country, especially California, water is a precious resource, and the development of new supply is costly. To the extent that energy efficiency programs include measures that save energy for hot water and secondarily save water, society benefits. The volume of avoided water and waste water use (which are easily estimated from program records) can be valued at the avoided water cost or cost of the next water supply source where that information is available. Deferring development of a dam or next water source has potentially very significant societal benefits to communities in investment, access to capital, and helping keeping rates low.

- **Infrastructure, National Security, and Other Societal Benefits**: Little to no work has been conducted on the value of using US-based fuel sources (avoiding disruptions from import restrictions, etc.), for example, or on the value of using EE programs to defer construction of plants until "cleaner" fuels will be available. A preliminary scoping should be conducted to identify at least the bounds for these types of valuations.[74]

[73] The difficulty arises from the fact that the health impacts associated with an EE measure may have immediate effects, chronic impacts, and effects on prescription drugs, hospitalizations, doctor visits, and myriad other elements.

[74] In this case, the most appropriate methodology is probably one akin to that used by Gary S. Becker in his famous study examining the economics of going to church (Becker, 1962, 1976). He breaks the question into two parts – assessing the value (cost) of going to "hell", and the likelihood that there may be hell and that going to church may allow avoidance of the "hell" outcome. He reasons that if hell is as bad as they say (say, infinitely bad), and if there is even a tiny chance that it exists, then it is worth going to church. Similarly here, we would need to asses the costs of a war (or "significant import restriction event" in government-speak) and run scenarios assessing the risks of that outcome.

4.1.3 Participant Perspective NEBs and Measurement Methods

Aside from economic and environmental benefit computations, the greatest activity in the NEBs field has been in the area of estimating the benefits to participants – beyond energy use and energy bill decreases – from the adoption of energy-efficient technologies.

The most comprehensive assessment of the status of measurement of NEB categories is work completed in California for the Low Income Public Purpose Test (LIPPT). This study (TecMarket Works, SERA, and Megdal, 2001) details the computation of more than 11 participant NEB categories.[75] They were: water / sewer savings; shutoffs; reconnects; calls to the utility; property value benefits; fires; indoor air quality; moving/mobility; illnesses and associated economic impacts; transaction costs; "soft" benefits (combining a variety of factors like comfort, etc.); and hardship effects.[76] Using the model and computation recommendations, several other studies (Schweitzer and Tonn 2002, New England studies, and others) have used this model to tailor estimates for other utilities or regions, and pieces of the research have been used since for computing NEBs by California Utilities. Not a great deal of attention has been paid to updating any of the NEB areas other than the extensive work in developing better estimates of the "soft" impacts (see next section). An updated review of this work by the author (Skumatz and Khawaja, 2009) identifies a number of NEBs categories from among this list that bear review and improvement. Many of these estimation improvements will be addressed as part of a "next phase" LIPPT project in California.[77]

- Utility perspective: updates to address kW and peak/off-peak NEB impacts; line losses; health and safety; and capacity building/ deferral values.
- Societal perspective: health and safety; tax credit considerations; national security; and neighborhood preservation.
- Participant perspective: non-energy operating costs; financial computations for maintenance and lifetime effects; fires / safety methodology; mobility, hardship / family stability, and others.

More than 45 studies on NEBs for participants have been included in the major energy journals since 2001.[78] The studies address one or several of the following topics:

- Methods for estimating specific (or groups of) participant NEBs
- Participant NEB estimation results for specific programs
- Recommendations for additional research on participant NEBs
- Recommendations for appropriate uses for participant NEBs

NEBs Measurement

Well-researched measurement work on NEBs, based on detailed literature research and work in contingent valuation, scaling techniques, revealed and stated preference and other methods

[75] The study also included extensive treatment of the derivation and estimation of a series of utility and societal NEBs as well.
[76] Extensive work since has adapted these methods to non-residential programs as well, adapting to estimating productivity,sick days, and many other effects. See Pearson et.al. (2002) and many others in the literature cited in the bibliography.
[77] This next phase work is being conducted currently by Skumatz Economic Research Associates, with subcontractor Cadmus Group. The client is Sempra Utilities for the California IOUs and CPUC.
[78] Skumatz conducted a thorough review of more than 350 studies related to NEBs for a project in 2000/2001 (see Skumatz 2000 and TecMarket Works, Skumatz, and Megdal 2001). The findings and conclusions from that research are still relevant and are discussed in this research paper.

were pioneered in the late 1990s.[79] Granted, NEBs are, almost by definition, Hard to Measure (HTM)[80]; however, not measuring the effects means that decisions about programs are likely to be suboptimal because they ignore key effects. Running scenario analysis around ranges or order of magnitude values would be preferable to excluding the impacts altogether. Thus, approximate estimates provide value; the improving sophistication of measurement methods implies that these approximations are getting better and better.[81]

By far, the greatest controversies related to participant NEBs arise from two issues:

- Measurement / computation approach, and associated confidence in the results, and
- Appropriate uses of the estimated NEBs.

Measurement / Computation approaches for Participant NEBs.

The major approaches to measuring participant NEBs that have been used (or proposed) at the individual household or business level are briefly outlined below:

- Computational approaches, using primary or secondary data, and computational or statistical approaches; and
- Survey-based estimate approaches, including stated preference surveys and revealed preference approaches. These include: Willingness to pay (WTP) / willingness to accept (WTA) / contingent valuation (CV); comparative or relative valuations; discrete choice and ordered logit approaches, and other revealed preference and stated preference approaches.[82]

Direct computation approaches have obvious benefits. Unfortunately, an extensive array of less tangible but potentially important benefits that have been repeatedly listed as important in the literature cannot be estimated directly, including comfort and aesthetics. Thus, relying on computational methods is not sufficient in deriving overall estimates of participant-perspective NEBs. A variety of survey-based valuation methods have been used by economists, social scientists, and researchers in the environmental and advertising fields to develop estimates of the monetary value of externalities and intangible goods. Each method has been derived from a review and the application of well researched academic literature. Methods with particular applicability to energy are discussed below and in Table 4.2 (see Skumatz and Gardner 2006), including direct computation, stated preference survey, [83] and other approaches. We categorize them into 7 different types and 11 methods that have been applied to NEBs to some degree.

[79] Measurement methods have been discussed in detail in previous papers including Skumatz 2002 and Skumatz and Gardner 2006. Choice models have also been applied in several projects (Skumatz and Gardner 2004; NYSEDA 2007), with enouraging results.
[80] A term used frequently by Megdal in her literature (TecmarketWorks, et.al. 2001).
[81] Skumatz argues that zero is probably one of the few values known to be incorrect for NEB values.
[82] Analysis of these approaches is provided in Skumatz (2002) and Skumatz and Gardner (2006).
[83] Since 1994, the standard prelimiary steps in conducting these sureys has been to first ask an open-ended question about what NEBs may have been recognized by the respondent, then whether or not individual NEBs are positive or negative, before proceeding with more complex questions about valuations.

Table 4.2: Participant NEB Computation Approaches Proposed and Used to Date[84]

Category	Description	Specific estimation approaches	Strengths	Weaknesses
A. Computational approach / Primary Estimation:	Some categories of NEBs can be estimated fairly directly. For example, lost work time can be calculated using pre-post office records and wage rates[85] or other monetary values for time.[86] Similarly, water/sewer savings can be calculated using data on actual water and sewer rates.	1. Primary computation	• Strong, reliable, defensible results well executed	• Expensive • Lacks large sample sizes, so applicability and statistical properties are weak • Generally only used for limited number of NEB categories • Self-selection participation bias
B. Computation using Secondary Data Estimates:	Secondary data from various sources are combined to develop a credible estimate of program impacts. For instance, if secondary data are available on risk of fires from particular measures, and the value of each average fire in terms of loss of property and life is available from, for instance, insurance companies, then these values can be multiplied times the number of measures installed to develop a total estimated value of risk from fires (or health and safety).	2. Computation from secondary sources	• Strong, reliable, defensible results • Adaptable to scenario analysis	• As strong as the secondary sources • May only be applicable to a subset of very quantitative NEB categories
C. Computation / estimation using Regression Approaches:	In some cases, statistical and regression approaches have been used to develop estimates of productivity or other effects that can be affected by confounding factors (Okura et al. 2000). These have been applied to several very important NEBs related to daylighting: specifically, sales benefits in retail outlets and test performance improvements in schools.	3. Regression approach	• Strong performance, with statistical reliability associated with results • Can be used with important quantitative NEBs	• Expensive, labor and skill-intensive • Data collection difficult • Can only be used to estimate limited set of NEBs
D. Survey methods – Simple Contingent Valuation and Willingness to Pay (WTP) / Willingness to Accept (WTA) surveys.	Contingent valuation surveys are widely used in the environmental and natural resources fields to estimate the value of intangible or hard-to-measure impacts including recreation, environmental and other effects. The contingent valuation (CV) method of non-energy benefits valuation, in its most basic form, entails simply asking respondents to estimate the value of the benefits that they experienced in dollar terms (willingness to pay (WTP)/ willingness to accept (WTA) are common approaches). An advantage of WTP surveys is that they provide specific dollar values for benefits that can be compared to each other and to the value given for the comprehensive set of program benefits. Disadvantages include the difficulty that many respondents have in answering the questions, the volatility of the responses,	Methods include: 4. Open-ended contingent valuation WTP / WTA questions, [89] 5. Discrete contingent valuation questions,[90] 6. Double-bounded and one-and-one-half bounded question formats,[91]	• Common in literature • Clear in application • Relatively inexpensive	• Difficult for respondents to understand and answer • Volatile responses • Literature cites weaknesses with open-ended responses relative to bounded options

[84] Adapted from Skumatz and Gardner (2006).
[85] As noted in Skumatz and Gardner (2006), there are weaknesses from some of the direct computation methods as well. Direct computations are only available for an almost certainly non-random list of participants, and would likely be biased upward because only those businesses expecting large impacts would be likely to measure them.
[86] Some businesses may have conducted research of this type. However, estimates tend to be limited in nature, covering only the odd business or covering only one measure or a key benefit, limiting the size of the sample (and thus the error band estimation), as well as the coverage of NEBs.

Category	Description	Specific estimation approaches	Strengths	Weaknesses
	and significant variations in responses based on socioeconomic, demographic and attitudinal variables.[87][88] Enhancements over open-ended WTP or WTA options have been used in multiple NEB studies with varied levels of success.			
E. Survey methods – Relative scaling methods	In this approach, respondents are asked to state how much more valuable (specific or total) NEBs are relative to a base. That base may be a dollar amount, or another factor known to the respondents. Initial work focused on asking percentages higher / lower for valuations. After an extensive review of the academic literature, the use of simpler word-based comparisons (much more, etc.) could be justified and adapted, and was tested extensively.[92] The nomenclature in the academic literature for this approach is "labeled magnitude scaling" (LMS).[93]	In summary, the categories of these methods include: 7. Relative scaling in percentage terms; 8. Relative scaling in verbal terms (LMS)	• Well demonstrated in academic literature • Easy for respondents to answer / understandable • Less volatility than WTP / WTA / CV approaches • Inexpensive	• Requires good choice of enumerative / comparison factor. • LMS requires quantitative translation from verbal several responses

[87] Responses to open-ended contingent valuation questions are more prone to bias (Arrow et al. 1993), and the experience of the authors has been that such responses vary more than those provided by any of the other valuation techniques discussed in this paper (Skumatz 2002; Skumatz and Gardner 2006). Arrow et al. (1993) list the following criticisms of the contingent valuation (CV) method for environmental valuation: (1) CV can produce results that appear to be inconsistent with assumptions of rational choice; (2) responses can seem implausibly large when considering multiple programs; (3) relatively few previous applications of the CV method have reminded respondents of relevant budget constraints; (4) it can be difficult to provide adequate background information on the programs and assume it is absorbed by respondents; (5) it can be difficult to determine the "extent of market" in generating aggregate CV estimates, and (6) CV respondents may be expressing the "warm glow" of giving, rather than actual willingness to pay for the program in question.

[88] Skumatz and Gardner (2006) discuss these aprpoaches in great detail as they apply to NEBs; a summary of key issues follows. Despite the well-known limitations of direct or open-ended CV questions, there are certain situations in which they can be of use in the measurement of NEBs. However, while open-ended WTP can sometimes be useful in generating a baseline, to provide more consistent and credible survey information, several variations on WTP/CV approaches can be used: (1) Discrete CV questions, in which respondents are asked to give a binary "yes/no" response regarding whether they would be willing to pay a given amount for a specified good (e.g., the non-energy benefits that they experienced). This is the CV question format recommended by the 1993 NOAA panel on contingent valuation (Arrow et al. 1993); (2) Double-bounded or one-and-one-half bounded question formats, in which respondents are asked (a) to give a yes/no response to a first value, then give a follow up response to a second value, which is higher or lower depending on the response to the first question, or (b) told that the true value of the goods in question are thought to exist within a certain range, and asked to give a yes/no response to a random value, then asked to give a second response to a lower or higher value depending on the first response, unless the first response was a no to the lowest value or a yes to the highest value - these variations may increase the quality of the willingness to pay estimates obtained from CV questions (see Cooper, Hanemann and Signorello (2002) for a discussion)' (3) Ranking cards to estimate WTP (also called ordered logit) - the survey instrument used in this approach differs and asks respondents to rank several hypothetical scenarios in which the amount of non-energy benefits, other characteristics of the program, and a numeraire are varied at random, and a rank-order logit model is then used to estimate the parameters on the utility function. The advantage to the rank-order approach is that it neither asks respondents to provide percentage or dollar estimates of the value of the non-energy benefits that they experienced nor does it ask them, hypothetically, whether predetermined values would be acceptable in exchange for those benefits. An additional advantage of this approach is that the information obtained is very robust, and the models can often be estimated with relatively small sample sizes (Weitzel and Skumatz 2001).

[89] Used by multiple researchers.

[90] Used by multiple researchers.

[91] Used in Skumatz and Gardner 2006 and other work by the authors.

[92] The LMS was applied in Skumatz 2001. Multipliers to allow transtiion between words and values are presented in the literature; however, Skumatz used surveys from more than 500 respondents to confirm and refine these values for use in NEBs. The values from the academic literature were generally confirmed.

[93] The relative scaling method of non-energy benefits valuation is a stated preferences approach in which survey respondents are asked to express the value of the non-energy benefits that they experienced relative to a well-understood numeraire, such as the energy savings due to the energy-efficiency measures installed through the program, program costs, or potentially any of a host of outside / non-program factors. There are several variations on the basic approach. In the direct scaling variant, respondents are asked to estimate their non-energy benefits (both positive and negative) as a percentage of their cost savings on energy. In the Labeled Magnitude Scaling (LMS) variant, respondents are asked to rate their non-energy benefits as being more valuable, less valuable or as valuable as the numeraire (e.g., their energy savings). Responses are then scaled using multipliers derived from academic sources modified by extensive empirical work from energy surveys. The relative scaling method has several advantages for use in survey research. First, program participants often find it difficult to express non-energy benefits, which are intertwined with more directly energy-related aspects of the efficiency measures that they receive, in absolute levels. However, as participants in energy efficiency programs, they are often well-attuned to changes in

Category	Description	Specific estimation approaches	Strengths	Weaknesses
			• Can gain responses from large sample of customers, improving statistical properties	
F. Ranking-based survey approaches	These surveys ask respondents to rank NEBs or measures with alternative sets of NEBs on a two-way comparison basis (for example Analytic Hierarchy Process, AHP) or more numerous options in rank order (usually ordered logit or similar approaches). To make the estimates most robust with the least cards or questions, careful statistical design is needed (for example orthogonal models like latin squares). These approaches use information from the rankings to compute values and preferences. (Skumatz and Gardner 2004, Khawaja 2009, Wobus et.al. 2007)	9. AHP 10. Ranking and ordered logit approaches[94] [95]	• Robust estimates with good statistical properties are derived using this method • Requires less "monetizing" of NEBs by respondents • Strong academic grounding	• Complex question and experimental design • Can require complicated comparisons by respondents • Slower than other responses. • More difficult than some other approaches for analyzing multiple NEBs, measures.
G. Other survey-based approaches - Hedonic regression:	Most of the other methods presented have been the stated preference variety used for non-market (including environmental) goods; they require program participants to directly disclose, in one way or another, their preferences for non-energy benefits. Many non-energy benefits, however, are market goods. They are purchased by consumers, bundled with the energy-efficiency appliances that produce them, and hedonic regression approaches are suitable for these applications, decomposing price of a good as a function of its characteristics (Griliches 1961, Shelper 2001). With some variations, hedonic methods have been applied to NEBs.[96] [97]	11. Hedonic decomposition	• Well demonstrated in academic literature • Provides strong statistical and explanatory power / causal factors	• Expensive, labor and skill-intensive • Data collection complicated • Can only be used to estimate limited set of (quantitative) NEBs

household or business energy costs, and therefore fully cognizant of the value of reduced energy use. Expressing the value of non-energy benefits relative to more obvious energy savings is a natural comparison that most respondents can easily make (Skumatz and Gardner 2006). Skumatz used this approach for NEB use and aplied it in studies of residential appliance and low-income weatherization programs (Skumatz and Dickerson 1998; Skumatz, Dickerson and Coates 2000) and has since applied it in studies of ENERGY STAR home performance, new homes, and appliance programs (Fuchs, Skumatz and Ellefsen 2004). In these studies, respondents found the relative scaling questions much easier to answer than WTP questions and the responses were more consistent than those from WTP surveys.

[94] Linked with statistical modeling approaches.

[95] See Skumatz and Gardner 2004, Khawaja (2009) and Wobus, et.al. 2007.

[96] Because many of the characteristics of goods that give rise to non-energy benefits are abstract and subjective (e.g., light quality), the traditional hedonic regression approach may be difficult to apply. However, using the more restrictive definition of non-energy benefits, a hedonic approach to the estimation of the non-energy benefits that arise due to increased levels of energy efficiency technology is possible and has been used. Caroll (2005) discusses a similar approach, suggesting statistical analysis of revealed preferences. Revealed preference models using a combination of program data, and survey results can be used to derive estimates of NEB value. The models are used to determine how reported intent translates into action, incorporating information on, for example, the cost of the installed measures, the NEBs reported by participants, and the value of those NEBs as determined through a CV survey to derive estimates of the actual costs participants paid for the energy and NEBs associated with common projects or measures (Carroll 2005). One drawback of this approach is the time and expense associated with data collection and analysis. Skumatz and Gardner 2004 used the hedonic regressions approach to associate NEBs with specific meaures in a bundeled measure program.

[97] This technique may not be as robust as the stated preference approaches discussed above in that it is not capable of estimating subjective types of non-energy benefits because the more subjective characteristics of energy-using measures (aesthetics, contribution to household comfort and aesthetics, impact on health, etc.) are not available on a product-by-product basis, and are difficult to distill into readily interpretable units. This limitation notwithstanding, the hedonic regression approach non-energy benefits valuation uses data that are (a) readily available for most energy-consuming measures and (b) less susceptible to bias than direct estimates obtained from surveys. Of course, the hedonic regression approach also assumes that the characteristics of a good are the only significant determinants of its price – an assumption which may or may not be reasonable depending on the good under investigation (Skumatz and Gardner 2006).

Category	Description	Specific estimation approaches	Strengths	Weaknesses
H. Other survey approaches - Reported Motivations and Factor-Importance Judgments.	Customer-reported motivations for pursuing home performance projects and the relative weighting of those motivations can also be used to determine the value of the energy and non-energy benefits resulting from the project. Lutzenhiser asked customers in a California project about their motivations for buying comprehensive home performance retrofits. They reported multiple motivations among six categories (in order of frequency): specific system/building concern; environmental health and energy costs (tied); comfort; resource conservation; and other (Lutzenhiser Associates 2004).	13. Reported Motivations	• Strong performance analytically, statistically • Easy for respondents to answer • Handles quantitative and qualitative, hard and "soft" NEBs	• Expensive, labor and skill-intensive • Data collection complicated

Data Collection: Studies have used a variety of methods for collecting data to support the estimation of participant NEBs, including phone, mail, web, on-site interview and email approaches, as well as detailed on-site data collection using program and business records, etc. Of course, each of these data collection methods has the usual pros and cons (relative cost, speed, length / complexity tradeoffs, etc.). However, when it comes to survey-based NEBs, phone and web approaches provide important advantages; [98] interview and on-site data collection work best for ranking and regression-based options.

Comparison of Performance of Participant NEB Approaches

Advantages and disadvantages of these various approaches have been addressed in the literature and are summarized in Table 4.2. To date, only a few studies have directly compared NEB results arising from multiple measurement methods.[99] These studies compared performance based on several assessment criteria, including: credible methods / demonstrated in literature; ease of response by respondent / comprehension of the question by respondents[100] ; reliability of the results[101]; volatility of results within studies and in comparison to others; accuracy, consistent results; cost; and computational clarity. Various combinations of the studies allowed comparisons between labeled magnitude scaling (LMS), comparative percentage, willingness to accept (WTA), and willingness to pay (WTP) results, and ranking methods. Generally, the comparative research that examined quantitative and qualitative features associated with the NEB measurement methods found that:

- WTP and WTA results (from Group D in Table 4.2) showed much higher variation in results than other approaches (particularly Group E), and were confusing to respondents (resulting in missing observations). Comparative responses (Group E) were generally consistent across programs, and very quick for respondents to answer, supporting reasonable data collection from hundreds of respondents, which improves statistical properties. The verbal comparisons (LMS) (Method 9) were quicker for respondents than Method 8 (percentage), and the factors derived from the comparison of percentage vs. LMS categories were reported to be very consistent with the values reported in the academic literature.
- All methods involving WTP, WTA, and comparative valuation approaches (within Groups D and E) supported practical computation of NEBs for more than one NEB category.
- Ranking methods (Method D, number 7) provided for slower data collection than other methods, with more missing data. The questions were more difficult to construct, and only a few comparisons could be asked, limiting the number of NEBs that could be estimated. The results were more conservative (lower) than those derived using the comparative (LMS and percentage) methods.
- The hedonic method (Group G, number 11) was flexible, and the results were consistent in direction and size with *a priori* theory.

---

[98] These include easy skip patterns (to help shorten potentially lengthy and confusing batteries of questions) and the ability to provide greater explanations if the concepts are unclear to respondents. As costs decrease, larger samples can be accommodated, supporting better statistical properties, so this is also an advantage.

[99] Skumatz 1999, Skumatz 2002, Skumatz and Gardner 2006, Wobus, et.al, 2007..

[100] Assumed to be at least somewhat related to or reflecting reliability of individual responses – less "guessing" involved (Skumatz 2002).

[101] Given the types of categories of benefits being measured, "accuracy" is difficult to assert or verify. The literature that has addressed this issue tends to relate it to the next criteria, consistency of results (across similar programs, or for the same program at different times).

These preliminary results are useful as others explore these and other analysis methods. Several methods show strong results, balancing consistency, speed / efficiency / cost, and flexibility, and allow affordable estimation of multiple categories of NEBs (for example, LMS, percentage approaches). If only one important NEB is necessary to measure, the regression-based techniques may be well-suited to the purpose. However, more work needs to be done to cross-reference and cross-check the performance and especially the consistency of the results from the various methods. Only when considerable cross-checking is provided, along with demonstrated statistical properties, will confidence build for the computation of participant NEBs – especially the "softer", but still important benefits like comfort, and other NEBs. It is recommended that additional estimation work should proceed, employing multiple measures within one study to allow cross-checking and verification. Given that the literature has touted the importance of these benefits for two decades, developing credible measurement methods is important.

4.1.4 Current and Suggested Uses of NEBs

There seems to be no shortage of informal uses or potential applications of NEBs, or reluctance for application of NEBs to formal uses like regulatory benefit-cost and regulatory test applications. Introduction into more formal applications will depend on developing estimates that withstand scrutiny from a range of audiences.

The most commonly suggested current and potential uses of NEBs – which vary for utility, participant, and societal perspectives – are categorized in Table 4.3. Enhancements on these uses are described below.

Table 4.3: Summary of Current Uses for NEB Values
(Updated from BC Hydro 2008)

	Utility NEBs	Participant NEBs	Societal NEBs
Marketing & targeting		Yes	Suitable
Program refinement	Yes	Yes	Yes
B/C analysis for customers		Yes	Suitable
Portfolio development	Yes	Yes	Yes
B/C tests	Potential	Potential	Potential (high)

NEBs provide useful information for program marketing and targeting, program refinement, and many other applications. The benefits from these qualitative and informal / informational applications have been fairly non-controversial. A discussion of the more controversial topic of how NEBs may (or may not) be adopted into program level screening and related applications is included in the next section. NEB values have been used in the following ways:

- **Program marketing / targeting**: Participant NEBs perform a function parallel to market research in product sales. NEB research uncovers those non-energy aspects of EE programs and measures that appeal to businesses and households that may be the target of the programs, and in particular to those potential participants that are not already "sold" on energy efficiency features alone.[102] NEBs can also be used to identify high impact measures and high impact target participants for programs, optimizing impact vs. cost.

[102] For example, Procter and Gamble doesn't market Tide(tm) laundry soap by advertising that consumers should buy it "because it is our highest profit item," which may in fact be the reason they want to sell it. Similarly, just because energy programs want to push certain models becuase they are energy efficient doesn't mean that it is the feature on which the equipment must be marketed. Likely, those customers that care about efficiency are already sold on those models. Appealing to the next level of participants requires marketing on features that reach

- **Program refinement**: NEBs provide feedback akin to that provided by process evaluations. Negative NEBs reflect important program barriers that can be addressed. Differences in perception of NEBs by different actors in the supply chain[103] identify information, training, or other needs at various intervention points. A detailed NEB analysis can provide information for refining the level or design of the rebate or intervention level (Stoecklein and Skumatz 2006).
- **Portfolio development**: NEB analysis allows the design of portfolios that maximize societal, utility, and / or participant benefits (or targeted NEB elements) given a fixed budget. Tradeoffs can be made between programs and measures to optimize a portfolio toward an array of financial and non-financial objectives, and provide a fuller assessment of portfolio impacts.
- **Benefit/Cost (B/C) analysis for customers**: Businesses and households select equipment (and behaviors) based on an internal assessment of the benefits and costs of an array of financial and non-financial considerations and features associated with that measure or behavior. NEBs provide a mechanism for identifying and providing a financial proxy for many of these "other" features. NEBs are a key component to understanding the participant's B/C analysis and their underlying program and participation decision-making. NEBs provide information to refine the program and support the refinement of incentives to make the B/C ratio favorable to program objectives.[104]

It is the area of B/C tests and program-level (and portfolio-level) screening that leads to the greatest controversy in NEBs. This topic is discussed in more detail below.

Alternatives for NEBs in Program-level Screening

Including NEBs in applications with significant financial implications like program screening is hampered by concerns about the reliability of estimates of NEBs. Although estimates of environmental effects are becoming more refined, that was not always the case. There have always been concerns about valuations of indirect benefits like comfort, aesthetics, and other "soft" benefits, or complex benefits like productivity, etc. For that reason, some agencies have defined subsets of NEBs that they consider "readily measured,"[105] and subsets of these are sometimes included in program screening or other applications. Examples of some of these "readily measured" benefits follow:

- Maintenance, GHG, equipment life, reduced waste generation or product losses, improvements in equipment productivity, increased floor space (BC Hydro 2008).[106]

---

them. For Tide(tm), that may be clean smelling clothes; for energy-efficient appliances, that may be lower water or soap use (clothes washers), less maintenance (CFLs) or other feaures identied, ranked, and valued by NEB research. It is easier to "sell" on these other features as well – it breaks through more easily in the cluttered marketplace. These features have been used to "sell" programs in the US and internationally, for example, New Zealand (Stoecklein and Skumatz 2006).

[103] These differences are termed "disconnects" (Skumatz 2004). In research for Focus on Energy (Skumatz and Schare 2002), the authors point out that architecture and engineering firms may be specifying and recommending fewer EE measures than owners would be willing to invest in, and that it may be leading to under-investment in EE in new construction.

[104] An example from a boiler program illustrates this concept (Skumatz, unpublished). Rebate levels were established to provide a customer B/C ratio that would favor the highest efficiency model. However, customers were purchasing a somewhat lower efficiency model more frequently than desired. The NEB analysis demonstrated that one of the highest value features of the other model was its small carbon footprint, and the footprint value outweighed the difference in incentive levels. To modify behaviors, the incentives needed to be adjusted. The utility made the simplifying error of assessing customer B/C in terms of energy costs vs. purchase cost alone, rather than the greater bundle of features. NEBs provide proxies for those underlying values.

[105] This section relies heavily on a very nice and concise analysis of NEBs prepared by BC Hydro (2008).

[106] BC Hydro considers the following not readily measurable: sales, property value, satisfaction, worker / student productivity, health and safety, comfort, noise, aesthetics, convenience, pride / prestige, and sense of environmental responsibility.

- Carbon value on societal test, and present value of deferred plant extension, water / sewer savings. Other specific measures benefits (e.g., lower soap use for laundry) (Gordon, 2008).[107]
- Others defined them in less specific terms, such as: reliable and with real economic value (Massachusetts); maintenance and equipment replacement (Vermont); measurable with current market values (Colorado) (BC Hydro 2008).

As an early approach, some other utilities incorporated percentage "adders" meant to reflect the presence of NEBs, but remaining non-specific about their sources and variations in values that may accrue to different types of programs.

Utilities have used, and proposed, a number of alternatives for including NEBs in program-level screening.

1. **Adder**: Use an adder to reflect all NEBs. An adder is included in cost-effectiveness analysis to represent a range of non-energy benefits. In the absence of a transparent link between the adder and specific NEBs, and to be conservative, the adder could be in the range of 10-15% of participant's energy bill savings (BC Hydro and New Hampshire are currently doing this).
2. **Readily measurable NEBs only**: Options are described above. Vermont, Massachusetts, Colorado and Oregon are currently doing this.
3. **All NEBs** - Readily measurable and best estimates of a selection of HTM NEBs (including subjective NEBs), relevant to the cost test or application. One must ensure that double counting does not occur. There are no current examples.[108]
4. **Hybrid** – readily measurable NEBs and an adder for HTM NEBs: includes readily measurable NEBs and a conservative adder for HTM NEBs. One must ensure that double counting does not occur. There are no examples.

In a recent analysis, BC Hydro (2008) examined the alternatives based on how they met three objectives: maximize DSM opportunities, minimize regulatory risk, and minimize evaluation resources. The summary of this evaluation is provided in Table 4.4.

Table 4.4: NEB Alternatives in Evaluation and Cost Tests (from BC Hydro 2008)

		Alternatives			
Objective	Criteria	Adder	Readily Measurable	All NEBs	Hybrid
Maximize DSM Opportunities	Range of NEBs included	Small range of NEBs included	Moderate range	Wide range	Wide range
Minimize Regulatory Risk	Robustness of NEB valuation + Jurisdictional support	Low regulatory risk	Med regulatory risk	High regulatory risk	Med-high regulatory risk
Minimize Evaluation Resources	Evaluation simplicity	Minimal evaluation resources	Med evaluation resources)	High evaluation resources	Med evaluation resources

[107] Gordon (2008).

[108] Considered in Caliornia as part of the Low Income Public Purpose Test (LIPPT) analysis (TecmarketWorks, SERA, and Megdal, 2001); also, a version of this has been used in New York by NYSERDA and has included alternative subsets of NEBs in various scenarios of the cost test that were presented to the regulator (e.g., non-energy benefits excluding macroeconomic benefits; and another scenario adding economic effects. In other incarnations, percentages of NEBs were included in the Benefit/Cost ratio analyses.). (NYSERDA 2005)

BC Hydro's analysis of the options probably represents the thoughts of many utilities considering next steps with NEBs. They note that

> *"…including HTM NEBs in the Total Resource Cost (TRC) test has the highest regulatory risk, due to concerns about the robustness in valuation methods and the fact that no other jurisdictions were found to include these NEBs in their program screening. And while the adder option has the lowest regulatory risk, it ranks the lowest in terms of maximizing DSM opportunities as it does not allow benefits over the "adder" amount to be considered in the TRC.*
>
> *Compared to the other alternatives evaluated, incorporating readily measurable NEBs in the TRC allows the most NEBs to be considered in the cost-benefit analysis while having moderate regulatory risk. Incorporating readily measurable NEBs can be done with relatively robust valuation methods and is an approach taken in a number other jurisdictions. Further, this alternative can be implemented in the near term and requires only moderate evaluation resources.*
>
> *However, including only readily measurable NEBs could limit the benefits for commercial and residential programs which are more likely to have "hard to measure" NEBs. The hybrid option would allow more NEBs to be included by using an adder to capture "hard to measure" benefits, but suffers in terms of increased regulatory risk (no jurisdictions found to use this approach). … In any of these alternatives, the same methods and effort should be employed to establish any non-energy costs." (BC Hydro 2008)*

The crux of the issue is the confidence in the estimates of HTM NEBs.

BC Hydro summarizes the continuum of NEBs use in program screening options (conservative to more aggressive), with examples of utilities that employ the metric. This information is included in Table 4.5.

Table 4.5: Approaches / Treatment of NEBs (updated from BC Hydro 2008)

NEBs Approach (Conservative to Aggressive)	Program Screen	Examples
Program marketing only - conservative	TRC	Ontario, Manitoba, Quebec
Scenario Analysis	TRC	New York (variety of NEBs included for scenario; programs must pass without NEBs)
Project screen	TRC	Wisconsin (participant-valued NEBs only)
Program screen – readily measurable	Modified TRC PPT	Massachusetts (NEBs must be "reliable and with real economic value"), California (only for low-income); Vermont (maintenance, equipment replacement); Colorado (measurable with current market values), New Hampshire (adder of 15%); BC Hydro; Oregon (especially for commercial and industrial)
Program screen – broader NEBs - aggressive	Modified TRC PPT	None found[109]

[109] Briefly considered / analyzed in 2001 for Low Income Public Purpose Test (LIPPT) for California, but no progress was made. Currnently, the California Public Utilities Commission Is considering modifications to the TRC to incorporate some NEBs as a cost offset. In addition, the State is issuing a Request for Proposals for another round of research on whether NEBs belong in tests for low income programs.

Additional detail and updated information on the approaches taken by a number of different states – including approaches used for low income programs - is provided in Table 4.6 below (Skumatz and Khawaja 2009, Collins 2009).

Table 4.6; Treatment of NEBs in a Sample of States

State / Region	Are NEBs Examined / How	Are NEBs "Officially Used?
California	The State hired a consultant to construct a low income program NEB model in 2001, which computed about 30 utility, societal, and participant NEBs, some of which were incorporated into low income program analyses for the utilities. An updated approach (model or other) is currently being developed to 1) update / tailor assumptions and inputs, 2) add more NEBs and update measurement approaches, 3) transform the model to a measure, rather than program basis, and 4) better coordinate with the other processes and steps for submitting program benefit cost results for program screening and the needed scenarios, etc.	The State investigated formal inclusion of participant-side NEBs in tests of Low income programs (2001), and is currently reinvestigating that issue to some degree. There have also been specific discussions with the regulators about indirect ways to incorporate NEBs into the current benefit-cost test model (Knight 2008).
ID, OR, UT, WA, WY - PacifiCorp	They do not quantify NEBs, except limited arrearage analyses. Some evaluation work – potentially including NEBs – are conducted if the program is performing poorly to see if NEBs can help improve the cost-effectiveness.	They use an environmental "adder" of 10% of the benefits for low income cost-effectiveness if the regulators allow (as they do – or did – in Washington, see below)
NY	Detailed evaluation of NEBs is conducted for many or all of the programs in their residential, commercial, industrial portfolio. They estimate a variety of utility, participant, and societal NEBs. For participant NEBs, they generally use the survey method developed in the literature,[110] For societal figures (emissions and jobs) they use specialized regional models developed by a consulting firm. For utility benefits they generally rely on defaults and proxy values from the literature, adjusted for New York, and do not generally conduct arrearage or similar studies.	NEBs such as comfort, safety, air quality, productivity, etc are included in regulatory cost-effectiveness evaluations for low income. For other programs, they have presented information to the regulators that include NEBs, and regulators are shown the benefit cost results including zero NEBs, participant NEBs, and participant plus economic NEBs, for example – a scenario approach. The NEB results are also used for analyzing marketing and outreach, but this is not a regulatory requirement.
Vermont	A calculation of NEBS associated with Vermont's weatherization program was conducted in 1999, (adapting numbers developed for a California program), and the numbers were updated for the 2007 report. This report used a combination of program, secondary, and literature-based inputs. Currently, this is the only efficiency program in Vermont that quantifies NEBs.	NEBs such as reduced air emissions, property value increases, tax benefits, health improvements and employment impacts are incorporated into formal cost-benefit analysis for the low income program, which is required by the state legislature. The analysis is also used for marketing and outreach.
Pacific Northwest; (from BPA, Energy Trust, and NEEA)	Calculations are measure specific (for BPA), not program specific, and in the residential sector cover lighting, appliances, HVAC, etc. The "Regional Technical Forum" has established a protocol to evaluate the air emissions associated with specific measures (CFLs, appliances, windows, HVAC, etc.), and BPA is developing a method to evaluate the jobs and emissions impacts of energy efficiency projects funded by the Recovery Act. BPA would like to do	The work is being used in regulatory cost-effectiveness analysis. TRC calculations include the value of air emissions reductions. BPA will only fund cost-effective measures with at BC ratio of 1 or greater. Energy Trust / NEEA report that they include the "readily measured" NEBs in the cost-effectiveness reporting.

[110] They generally rely on the comparative measurement methods, and for some, they also incorporate conjoint methods. Each method was discussed in the seminar presented to Xcel at the beginning of this project. The measurement approach / process was initiated / set up by SERA.

State / Region	Are NEBs Examined / How	Are NEBs "Officially Used?
	whole house or program level analyses, but the current model is not designed for this. Energy Trust / NEEA consider "readily measured" NEBs associated with programs (for example, water savings for washer programs, etc.) They are measured using "direct-type methods. "Speculative' or "soft" metrics like comfort, etc. are not measured.	
Montana	The Montana Public Service Commission does not require non-energy benefits to be reported and none of the regulated utilities have done so. A possible exception is for the weatherization program where some non-energy benefits may have been reported for federal requirements. No NEBs are reported for the weatherization program. None of MO PSC's regulated utilities have reported NEBs for economic evaluations.	NEBs do not need to be reported for regulatory evaluations.
WA – Puget Sound Energy	PSE used to quantify some non energy benefits (environmental, comfort, and quality of life indicators), but doesn't currently do so. Usually relied on Regional Technical Forum values and on occasion used participant surveys and data to quantify benefits. No reports are available demonstrating past methodologies. Currently no NEBs are quantified, but since it is believed that significant NEBs are associated with the low-income weatherization program, a B/C ratio of .67 is allowed (a TRC test ratio of 1 is usually required).	NEBs were, but are no longer, used for internal and regulatory cost-effectiveness test. No NEBs are required to be reported for regulatory purposes, but lower B/C ratios are allowed for low-income weatherization programs because NEBs are assumed to be associated with those programs.
MA	The current TRC model does include NEBs, but the methodology and source data used to quantify NEBs is unclear for some of the values. The inputs are derived from various reports and existing literature, but there are concerns about the accuracy, and updates are planned. NSTAR plans to update them, and part of NSTAR's recently filed 3-year plan includes an evaluation of NEBs.	The benefit cost model used for regulatory cost-effectiveness evaluations has NEBs build in for reduced costs to utility (arrearages, termination, collections), and participant benefits (mobility, comfort, etc.).
Arizona	The average air emission (SOx and NOx) per kWh produced by a given utility is used to generate values of emissions reductions. Some utilities are beginning to incorporate the value of carbon reductions as well. Broader NEBs are not currently considered or assessed.	The Arizona Corporation Commission does not require NEBs to be included in cost-effectiveness evaluations, but will allow utilities to report air emissions reductions if presented to them
Arkansas	The Arkansas Public Service Commission efficiency programs are just getting underway. The pilot projects have not required any cost-benefit analysis, but the comprehensive programs will need to demonstrate cost-effective energy and capacity savings. No NEBs will be required to be reported, but the PSC would consider them (if presented).	NEBs do not need to be reported for regulatory evaluations.
Georgia	The Georgia Public Service Commission does allow evaluation of externalities. None of the regulated utilities have reported any NEBs as part of regulatory cost-effectiveness evaluations.	NEBs do not need to be reported for regulatory evaluations
South Carolina	Neither the South Carolina Code of Laws nor the Public Service Commission of South Carolina requires utilities to consider the non-energy benefits of energy efficiency in the utilities' economic analyses. The	NEBs do not need to be reported for regulatory evaluations.

State / Region	Are NEBs Examined / How	Are NEBs "Officially Used?
	Commission would consider such a proposal if presented by one of the regulated utilities.	
Wisconsin	They have included NEB quantification in a number of program evaluations (including participant NEBs), particularly in the low income / weatherization side.	Broad NEBs are not officially incorporated into regulatory cost-effectiveness.

Opportunities for including NEBs in benefit-cost tests are illustrated in the summary of benefit-cost tests used in various locations around North America (Table 4.7). Note that the last several rows include the potential to include subsets of NEBs – should more confidence be gained in the estimates of HTM NEBs. However, in the near term, estimates of the societal NEBs that have achieved a higher degree of measurement confidence (economic, emissions) can be included in the program screening and benefit-cost test analyses.

Table 4.7: Summary of Benefit-Cost Tests (adapted and updated from Amann 2006)

Test	Benefits	Costs	States Using Currently
Utility Cost (or Program Administrator Test)	• Avoided supply costs for transmission, distribution, and generation (TD&G) • Avoided gas and water supply costs	• Program administration • Participant incentives • Increased supply cost	CA, CT, HI, IA, IL, IN, MI, MN, MO, NY, OR, RI, TX, VA, WA, BPA
Ratepayer Impact Measure (RIM) (or No Loser's Test)	Same as above plus • increased revenue	Same as above plus • Decreased revenue	AR, CO, FL, GA, HI, IA, IN, MI, MN, NC, ND, NV, SC, VA, WI
Participant cost	• Utility bill reductions • Participant incentives	• Participant direct costs	AR, CA, FL, HI, IA, IN, MI, MN, NY, VA
Total Resource Cost (TRC)	• Avoided supply costs for TD&G • Avoided gas and water supply costs • Utility bill reductions	• Program administration • Participant incentives • Participant direct costs • Increases supply costs • Decreased revenue	AR, CA, CT, CO, GA, HI, IA, ID, IN, MA, ME, MI, MO, MT, NH, NJ, NV, NY, RI, SC, UT, VA, WA
Societal	Same as above plus • Externality benefits (reduced pollution, improved reliability, etc.)	Same as above	AZ, IA, ME, MN, MO, MT, NJ, OR, VT, WI
Public Purpose (includes NEBs)	Same as above plus • Participant incentives • Quantifiable participant NEBs	Same as above	CA, KY, WI (low income)
Total Market Effects (TMET) (includes NEBs)	Same as above plus • Additional participant NEBs (for program and spillover participants) plus • Broader macroeconomic effects	Same as above	For evaluation purposes only
Program Efficiency (PET) (includes NEBs)	Same as above	Same as above • Excluding participant direct costs	For evaluation purposes only
Initial BCA (Simple BC) (includes NEBs)	Same as Public Purpose Test plus • Participant direct costs (as negative benefit) [111]	Same as above	For evaluation purposes only

[111] Similar to the option proposed by Knight (2008).

A Total Market Effects (TMET) approach would provide the most complete feedback on program impacts, benefits, and costs, and the most comprehensive assessment of the expenditure of public goods dollars. However, to move to a full effects test (like the TMET, or a broadened version of the TRC) will take additional research on participant benefit measurement methods.

4.2 Overall Findings and Variations by Measures and Regions

4.2.1 Utility Perspective NEBs

Early attention for NEBs focused on ***utility perspective*** impacts, particularly in the areas of (carrying costs on) arrearages and on reduced low income subsidies. However, utility NEBs have been relatively ignored for the last decade, largely owing to:

- The trivial relative and absolute size of non-energy benefits for the utility. Impacts from arrearages are well under a dollar per participant households in most cases (for low income), and the size of even fairly aggressive NEB estimates for the utility sector rarely represents more than 10% of the total NEBs estimates for programs.[112] However, identifying real impacts from line loss (rather than ad hoc multipliers), and particularly system capacity avoidance may have much more significant impacts, increasing the relative and absolute size of the NEBs in this category.
- The relatively direct and non-controversial methods that can be used to measure most direct energy impacts.

These categories of NEBs have not been the subject of significant discussion and have been barely addressed in the conference literature since the late-1990s. As mentioned above, there are potentially valuable impacts associated with measures and behavioral / education-based, programs that are not being addressed, including

- Reduced loss through T&D line
- System capacity avoidance
- Safety, insurance, and risk / liability impacts

These impacts bear further study, as the size is not known, yet there are reasonable methods that can be devised to measure each one (TecMarket Works, Skumatz Economic Research Associates, and Megdal 2001), and elements are appropriate for inclusion in various benefit-cost tests.

Table 4.8: Patterns in Utility NEBs by Program Type and Region

	Utility NEBs
General results	Small – less than 10% of total NEBs in most cases.
Variations by Program type	The effects have historically been larger for low income programs because the potential impact from arrearages and the impact of rate subsidy reductions are larger. Some have found that programs that target high arrearage customers have particularly larger impacts from utility NEBs. Few other impacts have been examined in great detail. If capacity impacts

[112] In a review of the relative sizes of NEBs from an array of residential programs (including low income weatherization), Pearson and Skumatz (2002) found that these impacts rarely exceeded 10% of the size of all NEBs.

	Utility NEBs
	are examined and valued, it is likely peak programs will begin to have much more influential effects on Utility NEBs. To the extent line losses are higher or lower proportionally in peak vs. non-peak times; similar patterns will emerge if these values are incorporated.
Variations for behavioral vs. measure-based programs	No specific work has been conducted on this topic. Potential impacts are large because behavioral programs can be designed to shift use from peak times, which may have disproportionately large impacts on capacity savings, once those NEBs are computed.
Variations by sector	Low income programs bring more Utility NEBs for arrearage reduction and reduced rate subsidies.
Variations by region of the country	Climate zones could affect these NEBs because of the effect of harsh winter climates (and high summer conditioning) on bills and arrearages, including for low income households. No specific patterns have been uncovered.

4.2.2 Societal Perspective NEBs

Two key categories – GHG emissions and economic impacts / job creation – have been the subjects of considerable work over the last decade. Both tailored and third-party models have been developed that make the estimation of these impacts feasible on a wide-scale basis. The published work indicates that including these two benefits can dramatically change benefit-cost ratios; for example, Sumi et al. (2003) found that the B/C ratio changed from 3 to 5.7 for one program when these impacts were included.

The remaining issues associated with **GHG estimation** work include:

- **Level of detail**: Estimation work can be conducted using the average system generation mix, or enhanced to include variations for peak vs. non-peak periods; or further enhanced to reflect hourly dispatch. Each provides greater refinement in the results, but also adds to the cost. The literature is fairly clearly leaning away from the first approach, but either of the last two is considered reasonable and practical approximations, depending on the accuracy needed for the application.[113]
- **Methodological issues**: Before these types of results can be used for cap and trade purposes, three key methodological issues must be solved: additionality, program vs. project approach, and errors / uncertainty / risk (described above). For most other applications (B/C tests, program impact comparisons, etc.), reasonable assumptions can be made, and the modeling work is sophisticated enough to provide defensible estimates of these (increasingly) important – and historically under-examined – effects.

In the area of **jobs and economic impacts**, the main remaining issue relates to whether the "base" case that is the comparison against the "program" case is one in which the program funds would otherwise have been spent on electricity generation industries, or if the funds should be considered to have been derived from the public goods charge, and if returned to customers they would spend the funds on a market basket of goods similar to the consumer price index basket. Including both cases may be the most appropriate short-term scenario, especially since the modeling work is fairly straightforward and non-controversial.

[113] Sumi et al. (2009) seems to be one of the few studies that has compared methods. The study seemed to indicate that the comparing the two methods ("margin" vs. Hourly dispatch) led to a change in the estimate of emissions of between 0 and 14% depending on scenario.

There remain a few other societal NEBs that have not been studied very much, but may bear additional research. The most important include:

- The societal impacts of the capacity avoidance issues discussed under the utility perspective
- Deferred or avoided water infrastructure and investment
- Infrastructure effects and impacts / risk from national security / important restrictions
- Health impacts from indoor and outdoor air quality and other pollutants related to generation and energy efficiency in terms health care and quality of life burdens
- Neighborhood improvements / preservation

In terms of priority, the first four are probably the most important. Measurement methods are most challenging for the national security and health impacts metrics, and relatively straightforward for the first two.

There are significant variations in the results by both program type and region of the country; however, more papers have been published on methods than results (and only a few contain the actual results), so quantitative comparisons are limited. The results are summarized in Table 4.9.

Table 4.9: Patterns in Emissions and Job Impact NEBs by Type of Program and Region

	GHG Emissions	Economic Impacts
General results	The impacts are estimated to be very significant.	Range from multiplier of 3.54 for national expenditures on EE (Mulholland, Laitner, and Dietsch 2004) to multipliers of 0.25 for appliance replacement programs (Imbierowicz et al. 2006). In Oregon, one MW saved increases output by $2.2 million.
Variations by Program type	The effects vary significantly with program type to the extent that different programs deliver savings at different types of day / days of week / months of year. Emissions vary with the generation profile for the time the savings are delivered. Emissions reduction during peak hours is often smaller than for baseload reductions (baseload plants are less expensive but put off more GHG). However, see notes regarding region of country below. Thus, air conditioner programs will have different GHG emission profiles than lighting retrofits.	Dramatic impacts depending on program type because it affects different underlying industries affected by the program's specific measures and make-up (e.g., labor intensity). One study found multipliers from 30% to more than 200% for weatherization compared to. Appliance replacement programs[114] (Imbierowicz et al 2006). The study found that appliance replacement programs do not provide much of a multiplier effect even when national scope is considered, largely because appliances are mostly manufactured overseas.
Variations for behavioral vs. measure-based programs	No specific work has been conducted on this topic. Given the driver is environmental savings, the patterns would likely mimic variations by peak / off peak.	No specific work has been conducted on this topic. Given the driver is industrial classification, if the behavioral program does not involve investment in new measures (merely changes in use of existing measures), then the impacts would largely be the transfer of jobs from generation to the market basket of goods, and the market basket of goods is more labor intensive than generation.
Variations by sector	No additional variations than by program type or region as listed elsewhere.	No additional variations than by program type or region as listed elsewhere.

[114] The study found economic output multipliers assocated with weatherization program expenditures are considerably higher locally (more labor intensive) than those associated with appliance replacement programs (46% vs. 25% for WI, 49% vs. 34% for CA, and 106% vs. 25% US) (Imbierowicz, Skumatz, and Gardner 2006).

	GHG Emissions	Economic Impacts
Variations by region of the country	Significant variations by region of the country because the driver is electricity generation mix (at peak and off-peak). Where there is more hydro, emissions are lower, etc.	Variations are significant because the industry mix varies across the nation. The one study examining this impact[115] found that multiplier impacts for both weatherization and appliance replacement programs were always lower in Wisconsin than in California or nationwide (about 10% to 50% lower depending on program type). The study found slightly larger multipliers for California programs (likely due to broader industry mix), and largest when nationwide scope is considered.

4.2.3 Participant Perspective:

There has been considerable activity in the area of participant NEBs, with more than 40 published conference papers over the last few years. The results routinely find the following:

- Participant NEBs are large – commonly equaling or exceeding the value of the energy savings emanating from the program. This is especially true for whole house / whole building programs, new construction, and similar programs in both the residential and non-residential sectors.
- Although the ranking and relative sizes of individual NEB components vary by program and region, the most important NEBs on the residential side tend to be: comfort, 'doing good' for the environment, operations and maintenance / lifetime, and aesthetic effects.
- On the non-residential side, the most valued NEBs tend to relate to: comfort, operations and maintenance / lifetime, equipment performance, 'doing good' for the environment, and labor / productivity issues.
- Negative NEBs – reflecting barriers – have also been measured. On the non-residential side, maintenance is the most common concern; on the residential side, maintenance and aesthetics are the most common concerns.

More detailed information and patterns by type of program and region of the country are provided below.

Table 4.10: Variations in Participant NEBs by Program Type and Region

	Participant NEBs
General results	Large – often equal to the value of the energy savings, depending on program (see below). There are patterns in leading NEBs as listed above.
Variations by Program type	Participant NEBs are higher for whole building programs than individual measure programs. This seems largely related to the inclusion of measures that affect comfort (HVAC, windows, design features).
Variations for behavioral vs. measure-based programs	The literature shows that NEB analyses have been applied to behavioral /outreach programs including retro-commissioning real-time pricing, High Performance design training, ENERGY STAR® programs, low income weatherization with education components, and others. Each shows significant participant NEBs – much of which may be associated with the measures. However, low income participants credit greater understanding of energy use as a high NEB, and “doing good for the environment” scores high for other programs. The commissioning

[115] Imbierowicz, Skumatz, and Gardner (2006)

	Participant NEBs
	project gave high value to greater understanding of systems.
Variations by sector	High value residential side NEBs tend to be: comfort, 'doing good' for the environment, operations and maintenance / lifetime, and aesthetic effects. On the non-residential side, the most valued NEBs tend to relate to: comfort, operations and maintenance / lifetime, equipment performance, 'doing good' for the environment, and labor / productivity issues. Low income programs tend to have higher NEB values associated with feature like "improved understanding of equipment energy use", control over bills, and similar. Negative NEBs – reflecting barriers – have also been measured. On the non-residential side, maintenance is the most common concern; on the residential side, maintenance and aesthetic are the most common concerns.
Variations by region of the country	Climate zones are influential in the value of NEBs because much of the high-value benefits come from comfort (affected by harsh winter climates and high summer conditioning). This single factor is often 15% or more of all participant NEBs. One study found that the highest valued source of NEBs was the insulation work (related to comfort).[116] No specific patterns have been uncovered.

Upstream and Non-participant NEB Results – and the Relation to NTG and Process Evaluation

We add one additional level of complexity to the discussion of participant-perspective NEBs. There are different levels of "participants". Given the increasing prevalence of upstream programs[117], interviews have been conducted with vendors, builders, architects and engineers, realtors, manufacturers, residential / commercial occupants and owners, and others with roles in EE interventions. Interviews have also been conducted with non-participants to assess the impacts.

One key issue discovered is the convergence or divergence of NEB perception among participants and non-participants and between different actors involved in a program. If participants and non-participants have different valuations (and concerns), education may be useful in brining the market forward – and reflects market progress. Similarly, if vendors are skeptical on NEBs that seem to bring customers into the program or to invest in measures, additional training may be warranted. If the market players fear maintenance problems, that fear may affect decision making, and indicates a need for additional research on whether there are real maintenance barriers. NEBs are also indicators of the features that "sell" programs, and help convince participation. These types of findings are valuable for market players in marketing, program refinement and the other applications listed previously.

However, there are two key conclusions to take away from this discussion:

- NEB research relates closely to process evaluation, and can augment the understanding of many process questions like barriers, decision-making, and potentially when programs should exit the market.
- NEB research relates closely to net-to-gross attribution work. NEB factors heavily affect the decision about whether to purchase an energy efficiency measure or participate in energy efficiency programs. Research on these factors can potentially present strong corroborating information for self-report free ridership AND spillover.

NEB research provides quantitative information that provides value in both these applications.

[116] Skumatz and Gardner 2004.

[117] Programs designed to influence the manufacturers or vendors of energy equipment.

NEBs and Behavioral Programs

Behavioral measure and education-based changes in practices are full of NEBs – probably more so than more technology-focused energy-saving measures. These do not get quantified, and if they do get quantified, it is to "sell" the change but not as part of an energy impact evaluation (Bensch 2009). NEBs are an indicator of what motivates people (Peters 2008), and Fagan's research (2008) and other work indicates it an important decision-influencer in the industrial sector. The literature recognizes that these effects exist, but alternative quantification methods are more / less accepted (Albert 2009). Although these effects have been well-studied by a few researchers, an educated policy / regulatory community that grasps the values of NEBs and use them in setting policy goals and valuing programs is lacking (Peach 2009).[118] Most jurisdictions do not permit use of evaluator or program knowledge of NEBs (Peach 2009), and NEBs are acknowledged as hard to measure; however, some argue that methods similar to those used to track energy savings can track NEBs (Mulholland 2009). The treatment of NEBs – and their capture in evaluations – should be more standardized (Mulholland 2009), and this will help gain acceptance of NEBs and the full value of programs in program planning, assessment, and regulatory tests. The potential for behavioral programs to affect demand responses / peak usage is key, and NEBs reflect this effect. Without consideration of NEBs in regulatory tests and other cost-effectiveness work, important effects of behavioral programs are omitted, and relative performance of these programs is understated.

4.3 Issues / Problems Identified

Utility NEBs:

Utility NEBs were the first drivers of NEB research; however, they have not been the subject of much real research over the last decade. There are, however, potentially valuable impacts associated with both measure and behavioral / education-based programs that are not being addressed, including

- Reduced loss through transmission and distribution lines
- System capacity avoidance
- Safety, insurance, and risk / liability impacts

These impacts bear further study, as the size is not known, there are reasonable methods that can be devised to measure each one[119], and elements are appropriate for inclusion in various benefit-cost tests.

[118] Peach (2009) suggests it is particularly important to incorporate global warming costs / damage per unit of carbon or carbon equivalent in program evaluations and regulatory tests and that a modified TRC that takes account if these effects – in an aggressive way (based on the laws of physics, and on financial investment or essentially the price of coal) – is critical. The outcomes of the evaluation work is that we run Plan "B" programs, when we need to be running more radical Plan "C" programs that reduce 70-80% of the energy savings in each sector. He also suggested we need to stop discounting the "out year" savings or reverse the sign on effects associate with global warming.

[119] See TecMarket Works, Skumatz Economic Research Associates, and Megdal 2001.

Societal NEBs:

There has been considerable progress in the estimation of two key societal NEBs in the last decade:

- GHG emissions, and
- economic impacts / job creation.

Both have emerged with reasonable agreement on practical and defensible measurement approaches – and both show significant impacts beyond energy savings that can be clearly associated with EE programs (on the order of halving payback in some cases).

There are only a few policy issues remaining. For **GHG emissions** work they include:

- **Level of detail**: Estimation work can be conducted using the average system generation mix, or enhanced to include variations for peak vs. non-peak periods; or further enhanced to reflect hourly dispatch. Each provides greater refinement in the results, but also adds to the cost. The literature is clearly leaning away from the first approach, but either of the last two is considered reasonable and practical approximations, depending on the accuracy needed for the application.[120]
- **Methodological issues**: Before these types of results can be used for cap and trade purposes, three key methodological issues must be solved: additionality, program vs. project approach, and errors / uncertainty / risk. However, for most other applications (B/C tests and program impact comparisons), reasonable assumptions can be made and the modeling work is sophisticated enough to provide defensible estimates of these (increasingly) important – and historically under-examined – effects.

In the area of **jobs and economic impacts**, the main remaining issue relates to whether the "base" case that is the comparison against the "program" case is one in which the program funds are otherwise spent on a market basket of CPI goods (reflecting the source of funds as public goods charges) or electricity generation (reflecting the direct replacement of this source). Including both cases may be the most appropriate short term scenario, especially since the modeling work is fairly straightforward and non-controversial.

To broaden acceptance, regulators may want to "approve" a few of the leading third-party models as acceptable for regulatory applications, and this may be the case for both GHG and economic estimation work. Inclusion of these NEBs in regulatory tests and program screening applications seems quite justifiable – they should not be considered overly hard to measure anymore.

There remain a few other societal NEBs that have not been studied very much, but may bear additional research. The most important include:

- The societal impacts of the capacity avoidance issues
- Deferred or avoided water infrastructure and investment
- Infrastructure effects and impacts / risk from national security / important restrictions
- Health impacts from indoor and outdoor air quality and other pollutants related to generation and energy efficiency in terms health care and quality of life burdens

[120] Sumi et al. (2009) provides a comparison, but few other studies have reviewed quantitative differences.

- Neighborhood improvements / preservation

Participant NEBs:

The impacts or values associated with participant NEBs have been recognized as important and large by implementers, researchers, and regulatory staff. Many NEB evaluations have shown that difficulty arises in the measurement of HTM elements of participant NEBs. Participant NEBs are useful in:

- Marketing & targeting
- Program refinement
- B/C internal customer
- Portfolio development
- Possibly program screening (for at least a partial list of NEBs) if the measurement hurdles can be overcome

Overriding best methods principles have been developed (assessing "net" impacts, avoiding overlap, including open-ended options, etc.). As for the detail of the NEBs that require participant input, about one dozen different methods have been used to measure participant NEBs, each deriving from the academic literature. Limited work has been conducted examining the performance patterns and reliability / consistency of the measurement methods, but "accuracy" is difficult to measure because the impacts being assessed include "comfort", and other "soft" effects. Ranking and survey methods are showing promise for estimating HTM effects, and results are showing patterns, so if reliability can be improved, simplified models or deemed values for major program types may ultimately be feasible.

Given the feedback potential from NEBs to program design and implementation, as well as in informing decision-making and attribution, NEB elements should be incorporated into either standard process evaluation work or into NTG / attribution surveys, or stand alone NEB work should feed results to the process and NTG analyses.

The most important recommendations regarding participant NEBs are:

- Additional research on developing and proving reliable measurement methods for the array of key HTM NEBs is needed. If methods cannot be developed, then multipliers that provide a fairly confident estimate of effects – or at least a portion of the effects – are needed. Large sample sizes should be used to allow strong comparisons.
- More studies comparing the results of participant NEB measurement methods within the same program and evaluation are needed. This will help flush out the most reliable, practical, and accurate HTM measurement approaches.
- Tests of methods for measuring "soft" benefits should include some NEB categories that are more quantitative and can be measured and verified. This may provide some benchmarking assistance to assess the performance for those NEB categories that cannot be "verified" (e.g. comfort, aesthetics, etc.).
- Additional work on evaluating individual important NEBs (such as the impact of daylighting on retail sales and student performance scores) should be conducted. In addition to the "usual suspects" of categories, examples of NEBs suspected of being strong include: reduced absenteeism due to sick building syndrome, decreased measurable indoor air

quality problems (mold, mildew, and noxious emissions), improved staff / tenant retention, operations and maintenance, and risk reductions for owners (Birr and Singer 2008). These will likely have traction for a number of building decision-makers with a natural skepticism toward NEBs and their measurement.

- Work on evaluating the transferability of results from one utility or one program to another would help leverage the work that has been (and will be) conducted.
- Work with policy makers to identify key NEBs, and acceptable measurement methods so appropriate subsets of participant NEBs can ultimately be incorporated into benefit-cost analysis and program screening protocols.
- In the short run, while measurement methods are being assessed, it may be appropriate to identify some ranges for "multipliers" as proxies for HTM NEBs (the most relevant subset) to be incorporated into program screening and B/C analysis applications at utilities and regulatory agencies. As reliable estimates evolve for more and more NEB categories, they can be removed from the multiplier and added as a "readily measured" option to allow for improved program decision-making.

Cross-cutting Recommendations:
Prioritizing additional research is a bit of a chicken and egg issue. It may not be worth the time to assess additional measurement methods unless they will be put to highly valued or important uses; however, they will not be put to these uses unless reliable and robust valuation approaches are identified *and trusted*.

There are, however, strong arguments for considering NEBs in some regulatory tests (for example, the TRC), at least on a theoretical basis (Skumatz and Khawaja 2009). For low income programs, the principal goals for the programs often relate directly to NEBs. In addition, the measurement of NEBs has matured, and there are applications for NEB values. Incorporating direct and improved economic and GHG NEBs can be useful in program screening and B/C metrics, as can incorporation of readily-measured NEBs. Getting to the "next step", in which proxy values might enter into the conversation, computations, and decision-making, would involve developing acceptable multipliers for the "other" HTM (not "readily measured" NEBs). Using these elements, "hybrid" NEB values could be developed for use in screening and B/C analyses. Research on patterns and values of NEBs across utilities and programs could help identify whether defensible and acceptable shortcuts or deemed values for subsets of important NEBs could be found. Also, as mentioned earlier in the text, there are individual NEB categories in need of further exploration.

Finally, the value of NEBs as input to process evaluation[121] and NTG computations should be further explored and potentially made part of the standard procedure for these evaluation types.

4.4 What Has Been Learned: Emerging Approaches and Experience

A great deal has been learned in NEBs in the last decade:

[121] NEBs provide an improved assessment of barriers and an indicator of potential gaps in the understanding of EE equipment for each actor in the supply chain (Skumatz and Stoecklein 2007).

- After years of just being listed and hypothesized, the literature has focused on developing estimation methods and has suggested that NEBs represent significant value – to society, participants, and to some degree, to utilities or agencies offering the programs.
- Utility NEBs are not substantial, but mainly because NEB categories with significant potential have not been investigated.
- Significant progress has been made in the area of estimating economic impacts from EE initiatives. Widely vetted third-party models seem to provide a good balance between ease and replicability. One issue that arises is that the models generally allow selection of impacts at the national, state, or county level. If a utility or agency's territory differs from these lines, some interpolation may be needed. In some cases, internal models have been developed to conduct the estimation work. This may or may not be necessary, but if the results are to be used for regulatory purposes, they probably need to be made publicly available to allow vetting.
- Significant progress has also been made in the area of estimating GHG emissions effects. Simple and complex approaches have been used, using varying degrees of complexity in generation mix and the associated emissions. The literature is moving away from the most simple methods (system-wide average) toward variations based on at least peak/non-peak generation mix, or hourly dispatch permutations. Where local plant emissions data are available, that may be a useful tailoring of the results.
- A great deal of activity has also focused around developing defensible methods for estimating participant-perspective NEBs, including indirect and "soft" benefits. Variations representing nearly a dozen methods have been used. Many have represented promising approaches, depending on the types of NEBs and the level of detail. Promising approaches include comparative methods, ranking methods, and regression / statistical methods. Willingness to pay / accept methods perform poorly. More work is needed in this area.

With exceptions, utilities and regulators generally have not incorporated NEBs into the regulatory or program approval process. This may be partly due to the relative new-ness of quantitative information, a lack of comfort with the estimation of important, but "soft," NEBs, or concerns about a reliance on self-report survey methods. New directions will likely include the following to advance the field:

- multiplicative adders to represent some or all of NEBs
- inclusion of "readily measured" NEBs
- incorporation of "readily measured" subsets of NEBs, or
- consideration of hybrid approaches including readily measured and some multiplier values

4.5 Conclusions and Additional Research Needed

4.5.1 Conclusions

Conclusions, recommendations, needed research, and other key issues uncovered in the analysis are detailed below.

Overall

- Non-energy benefits (NEB) are indirect and hard to measure (HTM). Consequently, they may also tend to be prone to higher levels of uncertainty than some other measurements associated with energy efficiency programs. The level of efforts spent on estimating these effects should presumably be somewhat proportionate with their potential impact in helping avoid wrong decisions about programs or EE interventions
- Best practices for measurement should be used to assure that the NEB estimates represent attributable impacts - similar to other direct impacts, this involves an assessment of "what would have happened absent the program intervention."
- NEB should be incorporated as a standard tool in evaluation (and potentially protocols).
- NEBs research has covered both widget- and behavioral programs. Those have included: weatherization with education, Energy Star appliance and marketing programs, Energy Star homes and home performance, builder training, real time pricing, and commissioning programs. Results have shown substantial NEB values, and these programs showed no particular measurement difficulties. Best practices recommendations apply to both types of programs.

Utility NEBS

- Utility NEBs were the first drivers of research. Further research has lagged primarily due to benefits being a small fraction of overall NEBs (perhaps 10% of all NEBs).
- Several key utility NEBs deserve additional study as their value is likely important and estimation of the effects will help reduce the undervaluing of EE programs. Research should be undertaken to devise and develop estimates of the following NEBs: reducing loss though T&D lines, system capacity/avoidance, and safety, insurance and risk liabilities. There are reasonable methods that can be devised to measure each one, and elements are appropriate for inclusion in various benefit-cost tests.
- Behavioral programs also generate utility NEBs, and those "driven" by energy savings are computed in a manner parallel to widget programs.

Societal

- There are three main levels of sophistication or complexity in measuring GHG reductions associated with EE measures including: System Average (using grid average plant and fuel mix to estimate emissions per MWh) – the least expensive and least reliable method. Peak / off-peak, or Margin Operations - refines estimates by using uses peak/off peak, or seasonal variations in generation mix. Hourly Dispatch - calculates emissions for each plant for each hour, which requires complex modeling of energy reduction over the entire grid and may include such calculations as the displaced emissions of building a new plant now, compared to in the future, when the plants may be more efficient.
- There are three major issues that need to be addressed and resolved before the environmental NEB results can be used for cap and trade applications. (1) Additionality: Parallel to free ridership, in GHG measurement, additionality refers to emission

reductions that are attributed to a program beyond those which would have occurred without the program's presence. (2) Program vs. project evaluation: The issue of whether to measure a *program* or a *project* has also been cited in much of the literature regarding GHG attribution. Generally, a single *project* such as an office audit and retrofit will not result in large avoided emissions and the evaluation may be costly. Looking at an entire group of similar projects, or completing a *program* evaluation using a sample of projects, may be more cost effective and result in higher quantifiable emissions reductions. (3) Error, Uncertainty, and Risk: Estimates of energy savings associated with energy efficiency and renewables strategies will have a component of error. These errors may be lower with renewables, as the comparison is "no plant." Energy efficiency represents a more complicated situation as the savings estimates are affected by baseline estimates, potential behavioral influences, etc, and in this case, uncertainty is a relevant term to use.

- When determining savings estimates, margin operations should be the minimum required analysis, and the hourly dispatch modeling may be justified for some programs, depending on the application, and evaluation budgets and goals.
- Periodically updated "deemed" factors (potentially ranges) for each generation fuel, and potentially categories of vintage of plant will provide a suitable method to estimate emissions. Applying these deemed values to programs would require assigning the program shares of "peak" vs. "non-peak" generation fuel mixes by utility or territory. For most program evaluation decision-making and uses, this level of detail will suffice, and it is not clear the payback from more enhanced modeling is needed and that it would balance the time and effort spent debating derivations, factors, and models. Based on preliminary research, where variations in emissions impacts on the order of 7% or 14% or less do not affect the direction of the findings, the enhanced modeling is not needed. For high value applications, more enhanced (hourly dispatch) modeling may be justified
- Recent economic impact research has relied largely on available input-output models – most commonly and cost-effectively using credible, vetted models available from third-party vendors. These models can support estimation to the county, state, or national level. The estimation work requires running a "base" and "scenario" case.
- Economic output multipliers associated with weatherization program expenditures are considerably higher locally (more labor intensive) than those associated with appliance replacement programs. Comparing state impacts found slightly larger multipliers for California programs (likely due to broader industry mix). In addition, appliance replacement programs do not provide much of a multiplier effect when national scope is considered, largely because appliances are mostly manufactured overseas.
- The main remaining issue relates to whether the "base" case that is the comparison against the "program" case is one in which the funds are otherwise spent on a market basket of CPI goods (reflecting the source of funds as public goods charges) or electricity generation (reflecting the direct replacement of this source). Including both cases may be the most appropriate short term scenario, especially since the modeling work is fairly straightforward and non-controversial.
- To broaden acceptance, regulators may want to "approve" a few of the leading third party models as acceptable for regulatory applications, and this may be the case for both GHG and economic estimation work. Inclusion of these NEBs in regulatory tests and program screening applications seems quite justifiable – they should not be considered overly hard to measure.

Participant

- This area has received considerable attention in the literature (more than 40 published conference papers over the last few years).
- Participant NEBs are large – commonly equaling or exceeding the value of the energy savings emanating from the program. This is especially true for whole house/whole building programs, new construction and similar programs in both the residential and non-residential sectors.
- Not measuring the effects means that decisions about programs are likely to be suboptimal. Running scenario analysis around ranges or order of magnitude values would be preferable to excluding the impacts altogether. Thus, approximate estimates provide value; the improving sophistication of measurement methods implies that these approximations are getting better and better.

4.5.2 Additional Research Needed

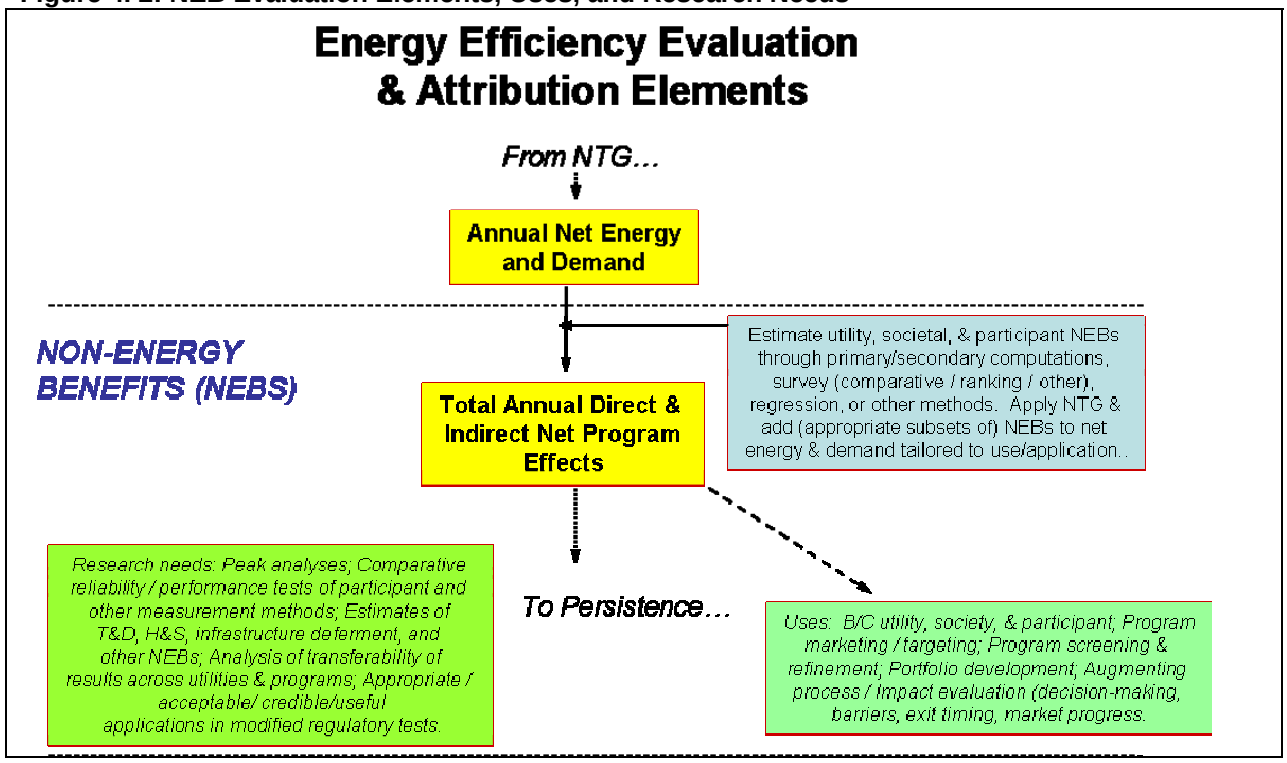
- Research is needed on several societal NEBs including:
  - Impacts of the capacity avoidance,
  - Infrastructure effects and impacts/risk from national security/important restrictions,
  - Health impacts from indoor and outdoor air quality and other pollutants related to generation (and on the other side, building tightening) in terms health care and quality of life burdens, and
  - Neighborhood improvements/preservation.

- NEBs, or variations in NEBs values, associated with peaking or demand elements of energy efficiency programs have not been much explored. Of particular interest (or value / cost) may be the impacts for avoided infrastructure or capacity for the utility (and potentially societal) perspective, as well as changes in line loss impacts.

- One of the most important needs in this area is additional estimation work testing multiple measurement methodologies within one study to allow cross-checking and verification to identify the best performing methods and increase confidence in NEB estimation – especially, but not exclusively, for participant perspective NEBs.
- Additional work showing results for individual important NEBs (such as the work on retail sales from day lighting, or student performance scores) should be conducted, assessing impacts for other quantitative impacts. In addition to the “usual suspects” of categories, examples of NEBs suspected of being substantial include: reduced absenteeism (due to sick building syndrome), decreased measurable IAQ problems (mold, mildew, and noxious emissions), improved staff/tenant retention, reduced operations and maintenance (O&M) costs, and risk reductions for owners.
- Changing protocols or procedures to Incorporate elements of NEBs analyses as a more standard part of process evaluation can bring value in understanding participation decision-making, program influences, barriers, education needs, and potentially provide indications of when programs should exit a market (or be modified).

- Work evaluating the transferability of results from one utility or one program to another would help leverage the work that has been (and will be) conducted.
- Work with policy makers to identify key NEBs, and acceptable measurement methods so appropriate subsets of participant NEBs can ultimately be incorporated into benefit cost and program screening protocols would be valuable. Whether this manifests as multipliers or more directly measured NEBs, including, rather than omitting NEB effects will more accurately reflect the impacts and the costs and benefits derived from energy efficiency programs and, as a consequence, improve decision-making.

Figure 4. 2: NEB Evaluation Elements, Uses, and Research Needs

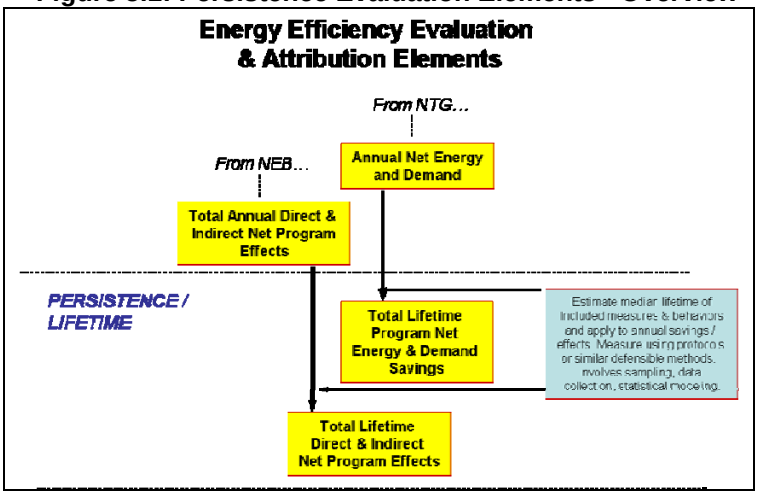


5. PERSISTENCE/ RETENTION / MEASURE LIFETIMES / EULS

Retention studies, also known as persistence or measure life studies, are a critical and highly useful component of energy efficiency research. There have been established protocols for EUL studies in California (California EM&V Protocols), and these protocols have guided the basics of the approach and the timing of effective useful life (EUL) studies for programs in the State. Despite some variations in data collection and treatment methodologies employed, the fundamental purpose of measure retention studies is to estimate the amount of time that a measure will be in place, presumably delivering energy efficiency benefits. The measure life provides a limit for the number of years that a program's annual savings will last. Early programs used figures related to laboratory lifetimes. Studies in the early 1990s demonstrated that a combination of factors affect the years over which a measure delivers savings, and it is not well-estimated using laboratory lifetimes. In the commercial sector, business turnover has a strong effect,[122] as well as changes in styles, perceived functionality, and the ability of maintenance staff to keep advanced equipment functioning. Parallel effects affect *in-situ* retention of equipment in households.

5.1 Current Practices and Uses

Figure 5.1: Persistence Evaluation Elements - Overview



The overall approach taken by most measure retention studies in the energy efficiency (EE) field is to estimate the median EUL of the measure in question. The EUL is usually defined as the median number of years[123] that a measure is likely to remain in-place and operable.[124] This amount of time is often calculated by estimating the amount of time until half of the units are no longer in-place and operable. The key data needed to derive these estimates are straightforward: installation location, measure(s) installed, date installed, and the date that the measure became inoperable or was removed. From these data, a basic measure life study can be conducted.[125]

[122] This is a factor that varies dramatically across the non-residential sector. Restaurants may turn over in 6 months or lighting styles; décor changes every couple of years; schools tend to stay schools and keep measures in place until well past their optimal functioning lifetime and dramatically past their optimal economic lifetime!

[123] Or other time interval, as appropriate.

[124] "In-place and operable" is at least the most common definition of measure survival. Depending on the specific measure under inquiry, alternative formulations of the definition may be more appropriate.

[125] Enhanced data can improve the estimates; these issues are discussed later in the paper.

While this task may seem straightforward at first glance, there are often considerable complications involved with obtaining EUL estimates. Measures often last for a long time, making it impractical to simply wait until half of the units fail in order to determine the median survival time. Measure lives are also frequently interrupted prematurely by the owners or employees of the residence or business in which the measure was installed. Obtaining unbiased EUL estimates, therefore, can require statistical analysis to (1) control for exogenous factors that might affect measure lifetime and (2) predict measure lifetimes based on empirical data. Furthermore, applications for this work require information on the projected results fairly early into the lifetime of much of the equipment installed as part of various programs, when a set of measures is young and only a relatively small portion of the installations may have failed. For example, protocols that were in place for many years in California required periodic verification of EULs when measures had been installed for fewer than five years. While important, this poses a particular challenge, as EUL estimates are based on failures, and few measures projected to last 20 years or more would be expected to fail under that schedule. Developing unbiased estimates of EULs under circumstances of limited data early in measure lifetimes is particularly challenging. Another disconcerting result from conducting EUL evaluations is the wait involved in the studies (4^{th}, 5^{th}, 9^{th} year studies, etc.). However, waiting a sufficient amount of time to conduct a rigorous statistical study will ensure that the technologies may undermine the usefulness of the study for current and future programs, because technology changes quickly.

Best Practices Summary

The authors evaluated more than 120 reports and studies addressing EUL methods, research, and primary studies estimating EULs covering a diverse collection of energy efficiency measures.[126] We compared the different data collection, treatment, and analysis techniques on the basis of their effectiveness in obtaining meaningful results, their ability to produce reasonable EUL estimates, the degree to which they produced statistical models that fit the data, and the defensibility of the conclusions drawn from them. The review of a large number of studies provided an opportunity to view the range of practices used for small and large, and simple and complex measures over a period of nearly ten years. We found a few problems that arose repeatedly:

- **Sampling-Based Issues**: We found that many studies had difficulties with the initial data set, which often is not designed with evaluation in mind. Depending on the type of program and its delivery method, data sets might not contain complete information on all of the three key pieces of information needed for retention studies (participant, measure(s) installed, installation date), or lack full (or updated) contact information. This was more problematic for programs delivered indirectly through commercial channels than programs installed by utilities and their directly-hired contractors. As time goes on (before or between retention studies), the problems in conducting a high quality retention study are exacerbated. In addition, some data sets were based on warranty or registration cards, which may represent a biased population. Finally, if multiple measures are under investigation, sometimes a measure-based population list may be preferable to a location-based sample because sampling might lead to bias from measure clustering. Problems with the initial database were not uncommon, and with poor starting records, the measure life study is probably fatally flawed.

[126] Building on the work conducted by Skumatz et al (2002, 2004, 2005)..

- **Data Collection Issues**: One of the most important factors in evaluating EULs has to do with the data collection method. The cost of on-site data collection is high, but for phone surveys to be practical, the measures must be unique and memorable. Phone surveys work for household furnaces, refrigerators, or water heaters; they don't work for CFLs that may have been installed at different times and are not unique, clearly identified measures. Potential removals need to be memorable, and the measures need to be clearly identifiable. Trained staff is needed to recognize the equipment as the model or type installed, and to probe for the best estimate of failure dates where respondents may not recall exact dates. Failure dates are the most critical data needed to support the estimation work.

- **Analysis Issues**: The most common problems identified at this stage were: (1) insufficient sample or insufficient failures to allow reasonable convergence for the estimate – and the only solution is a larger sample or delaying the analysis until several years later when presumably more failures will occur; (2) weak model selection, with the researchers testing only one specification of the measure survival curves[127]; and (3) neglecting to compare the results against results from previous years for the same program, or with the results from similar programs in the region or elsewhere. If the results do not jibe with existing studies, they may bear a second look to explain any potential inconsistencies.[128]

Table 5.1 presents a set of best practices for measure lifetime and retention studies derived from this research.

Table 5.1: Summary of Best Practices[129] (adapted from Skumatz 2005)

Best Practices
Sampling: 1. Obtain a strong and unbiased population source list from which to conduct a draw a sample. Strong data sets include, at a minimum, data on contact information for the site, measure(s) installed, and date(s) installed. The analysis is enhanced if the installation location within the property is also noted, and IF stickers can be affixed to measures when they are installed, the follow-up process is more reliable.[130] 2. If the number of measure installations is small, conduct a census. Otherwise, use a probability sample. Stratify the sample based on important population characteristics, such as climate zone and energy demand. Consider establishing a panel survey that is revisited every several years as it helps "bracket" the removal date if a date cannot be recalled.[131] 3. If possible, use a measure-based sample, rather than a site-based sample. Data Collection: 1. If phone interviews are conducted, use call management. Schedule phone calls in advance, use at least 3-5 callbacks, and leave sufficient time between callbacks. Assure the measures are unique and/or memorable before selecting phone interviews as the data collection method. 2. Pretest survey instruments for each measure under investigation. 3. Ask about conditions that might affect the operations of the measures (for instance, occupancy during the day, seasonal occupancy issues, etc.). 4. Try to get the most accurate information about measure-failure dates and explore causes / reasons.

[127] Most common statistical programs applied to this work have options available to apply exponential, log-logistic, log-normal, Weibull, and gamma distributions.

[128] This literature review and comparison issue was incorporated into the recommendations for the updated California protocols, as part of the "basic" rigor level of study.

[129] Table based on Skumatz (2005)

[130] Where the measure type permits, a sticker that suggests the household or business call a particular number at the utility if / when the measure is removed provides excellent data on interim failure dates.

[131] The study knows that the removal date must be after the previous panel survey, which is more than is known if a cold call to a new sample point can't recall the removal date.

Best Practices
5. Conduct follow-up interviews at time intervals appropriate to the measures under investigation. 6. Use trained and supervised auditors. 7. If on-site inspections are used: physically verify the status of each measure and affix identifying tags to measures and create a map of the measures sampled. 8. Use standard data management practices, such as double entry of data to reduce errors, and follow up calls regarding questionable responses. Analysis / Modeling: 1. Test for outliers (either visually or with a formal procedure) and remove obvious outliers. 2. Compare different models and model specifications with respect to their congruence with theory, implications for results, and results from formal tests. 3. Include influential variables as regressors to control for exogenous factors. 4. If failure dependency is suspected to be an issue, estimate models for dependent and independent failures. 5. If the sample does not accurately reflect the measure population, weight the data using the most appropriate means, and report both weighted and unweighted results. 6. If the sampling strategy resulted in clustering, use common standard error adjustments to compensate for xxx. 7. Compare results to previous studies and discuss differences, considerations.[132] 8. Clearly document the study and methods, alternatives considered, rationale, and discuss in context of results from other similar studies.

Remaining Useful Lifetimes / RULs

Some programs are designed to intervene at the time measures are being replaced, and the years and savings values to be assigned for the lifetime of the savings are fairly unambiguous. However, some programs may be geared toward replacing existing (lower efficiency) equipment with energy-efficient equipment before the old equipment ceases to function or before it would otherwise be replaced. These issues arise most often in early-replacement programs, relating to the assumption of whether the savings should be calculated as the difference between energy use for the old measure replaced and the new EE measure (we'll call this "enhanced delta"), or whether the appropriate savings computation is between the new standard measures available on the market compared to the EE measure induced by the program (we'll call this "standard delta"). The question arises, then, whether programs that lead to early replacement should be able to take credit for extra savings – "enhanced delta" during early replacement period– its "remaining useful lifetime" (RUL) before its EUL, and standard delta for the period after the normal measure lifetime for the old equipment.[133] That is, assuming data on age of the existing equipment can be gathered, should the program be able to count higher savings for the RUL period?

We conducted interviews with utilities and professionals across the nation on practices regarding RULs. Comments ranged from "we don't use these at all" to "they're used constantly", depending on the region / utility called. Many of the interviewees agreed that RULs were a concept that had some potential application in the situation of early removal of equipment. The theory is that different savings estimates should be used for the two periods of time – enhanced delta for the period between program replacement and when the measure would have been replaced without the program; and standard delta through the remainder of the

[132] See Skumatz et al. (2005).

[133] The total lifetime would still be the EUL. Presumably the RUL period would be assigned based on the difference between the age of the existing equipment and the EUL for the equipment. This would seem more consistent with EUL and savings computation practice, rather than conducting interviews asking when the equipment might have otherwise been replaced for each individual installation. Certainly, there will be those early replacements that are installations that tend to hang onto equipment longer than others, but on average, this may be the most computationally elegant approach, should it be used.

measure's EUL period. However, every respondent noted the difficulty of measuring the period in time for the early replacement – and noted it was (1) do-able, but (2) required separate questions for the program research. None believed it was appropriate to ONLY assign to the program the enhanced delta for the period in which the decision was moved forward, a possibility that had come up in some discussions.[134]

There were only a few primary studies of RULs. One was a Wisconsin study that examined a program that accelerated residential central air conditioner replacement. They gathered the age of the equipment that was being pulled out (using model numbers) and used the lifetimes associated with that equipment to calculate a mortality table (which properly takes into account the fact that if you've lived to 90 you stand a better chance of living to 100). These data were used to document the savings stream. A utility in the Northeast is undertaking a survey approach to examine this issue for a few programs.[135] A study in New York (Gowans 2005) and a study in California (Peterson 2005)[136] also developed approaches to associated RULs with specific programs and estimated impacts on program-associated energy savings.

Theoretically, it seems that using mortality computations on verified equipment age (or survey information where model information is not available) is a valid approach to the issue of RULs. *Ad hoc* assumptions (e.g., assuming that 1/3 of lifetimes remain) are not appropriate, especially given the much extended replacement intervals in schools, for example, compared to other business types. The two-part savings calculation is likely theoretically appropriate (it is shortchanging a measure if ONLY the early adoption portion is counted). However, the issue is not only one of when equipment would have died – it also involves a subjective decision by the business or homeowner. The estimate of years until the owner "would have replaced" the measure may be even less certain than the self-report information difficulties that some are having with responses used in free ridership and NTG computations. However, given that early replacement programs can have an impact on getting inefficient equipment out of the marketplace, it probably bears a few years of study to see whether reliable techniques can be used to generate responses – especially as the research will only require gathering a few more pieces of information at the point of program implementation / installation / replacement.[137] [138] [139] Research is clearly needed to identify best procedures for identifying the "hypothetical" expected removal date for early removals.

---

[134] The issue of RULs may also apply to behavioral programs, so if the issue is solved for measure-based programs, the same policy may apply. Consider the following hypothetical. If codes and standards were going to be implemented in a future year that would mandate some behavior (e.g., you may no longer leave outdoor lights on all night – they must be on a timer), and if a program moved that behavior-related impact forward, it is possible that a parallel situation with the measure-based program arises.

[135] Three other interviewees told about related issues. One belatedly found enhanced deltas were recorded for all participants for a program that needed later adjustment; another stated they had issues with first year savings being used throughout the life of the measures (they believed decay functions should be used); and another found out that the auditors were assigning all remaining years of early replacement to the first year – leading to a much over-estimated value for savings. These remain cautionary tales in looking at savings, early replacement, and savings computations and recording.

[136] The California study (on the residential program AC Energy Hog) based its analysis on assembled grouped data from the Residential Appliance Saturation Survey of 2002, linking stock counts and estimated hazards by age-of-appliance ranges, and Weibull specifications for the resulting survival function. The memo does not suggest the differences in estimated energy savings. The New York memo discusses the topic of RULs from a policy point of view.

[137] Specifically, equipment model and age, and survey questions on when they would have replaced and age of equipment that can't be traced through models.

[138] One utility interviewed uses the entire savings – old measure to new measure (enhanced delta)– throughout the lifetime of the measures, and they assume that a (majority) share of those installing the new measures (e.g., CFLs) will replace with CFLs again, so their savings go out beyond the initial lifetime. The utility notes that if something like this is not assumed, you should probably be readjusting your demand forecasts.

[139] The case of the very effective "cash for clunkers" early automobile replacement programs may be worth examining. Whether the early replacement period was assigned higher emissions savings than the later periods may suggest a precedent for the policy issue in energy.

However, the other part of the equation is the savings to be assumed for the period AFTER the equipment would otherwise have been removed. It may be:

- "standard" efficiency at the point of early replacement
- "standard" efficiency at the future date when removal would have occurred
- codes and standards level now or at that future date
- standard practice now or at the future date
- some other baseline

Identifying the standard efficiency at that future date is far from straightforward. The recommendation would be "standard practice", but practical methods to estimate a useful proxy for that metric would be needed to estimate the most accurate savings. The most practical alternative in the meantime (current "standard") would deliver an overestimate of savings. This is neither optimal nor conservative, and a more conservative alternative would be to minimize criticism and skepticism when EE savings are compared to generation alternatives. Research into other alternatives (adoption curves, incorporation of known standards upgrades, etc.) would be beneficial to see if any are applicable to this question.

This concept carries over directly to education and behavioral programs. Bringing forward in time a behavior that would tend to be generally expected (or mandated) in the future has near-term value. With more and more "green" education coming through a variety of mass media channels, greener behaviors are likely to become (more) standard, including energy saving behaviors. The measurement issues are even more complicated, but just as necessary to examine, if these programs are to become an increasing share of portfolios. Current behaviors are not going to stay the baseline forever, especially with increased mass media attention on green behaviors and potential for GHG emission reduction mandates, etc. Theoretically, this will tend to decrease savings associated with programs; however, the net result will depend on whether behavioral persistence (EUL) is longer than the baseline new behavior adoption.

Needless to say, no work has been conducted to date on this topic. Again, some kind of adoption curves may serve as a proxy; but research is needed. Many questions arise, such as: what would be assumed for timing? What would be assumed for the ultimate efficiency of the behaviors? How many different behaviors? Policy-wise, early adoption of new behaviors is an appropriate concept. However, measurement is a significant issue.

Technical Degradation / TDFs

We also explored the topic of Technical Degradation Factors (TDF), factors that are addressed in the California EM&V protocols. Another factor affecting how much savings is being delivered from program-related installations of energy efficiency equipment is whether the measures perform at the new efficiencies consistently over time, or whether their efficiency performance degrades over time (or potentially in given installations). Unexpected decay in performance could be an important issue, particularly for measures for which savings are assumed to accrue for upwards of 15 years. Unfortunately, in reviewing more than 100 EUL and TDF studies, we found very few TDF papers within the last decade that had been based on primary data. A paper by Jump et. al. (2008) applied 1998 laboratory measurements of lamp median life (from Rennselaer Polytechnic Institute Lighting Lab) to residential logger data collected by KEMA in

2003-2004, and derived an average CFL normalized lamp life. Using the rated life from the lamp packaging allows computation of an observed life.[140] An engineering study (Proctor 1997) provided valuable information suggesting that only a few measures may most likely have been affected in a positive or negative way relative to the decay in performance of standard measures.[141] Primary research could well be justified, particularly for measures with technical or engineering changes that may affect the degradation of specific equipment types relative to the degradation that would be expected with older technology – or measures accounting for large shares of portfolio savings.

Of course, it is worth noting that the performance degradation is the combination of two effects – technical degradation, as well as behavioral / operational component including the quality of use and quality of upkeep of the equipment [see California EM&V Protocols on EUL]. Studies that look at degradation *in situ* need to account for the influence of both these factors. Engineering studies that examine potential technical / mechanical reasons for differences in the relative performance decay over time may miss changes in the effects on the behavioral components. Therefore, setting priorities for future TDF studies will need to examine both technical and behavioral elements.

We do not separately address TDF related to behavioral programs, as we consider the concept in tandem with EUL. The behavior (as a measure or a performance) decays and ceases. Studies of both elements are needed, but the topic was addressed under EUL.[142]

5.2 Overall Findings and Patterns

Retention Results for Measure-Based Programs

The authors conducted a review of half a dozen recent studies of EUL summaries, as well as examining more than 100 EUL studies conducted on a host of programs in California, and we examined measure lifetime assignments for hundreds of measures. We found that deemed measure lifetimes or EUL values used by many different areas of the country seemed to have similar values, as illustrated in Table 5.2.

---

[140] This is a result distinct from an EUL because EULs include early burnout or removal, which is not captured by this method.

[141] The few studies identified included work by Jump et al. (2008), and research in the 1990s by Heschong-Mahone, SBW, and others. We also pursued leads for work by Ecotope, Stellar Processes, and others. In addiiton, according to one interviewee, the CPUC has approved a relatively large study for a team to begin testing some lamps that are currently on the market. Testing of this nature hasn't been conducted since the 1990s, and studies of CFLs at that time indicated that the results fell short of rated lifetimes for some lamps. The manufacturer's rated life is based on a schedule of three hours on, 20 minutes off (which does not parallel common usage patterns). (Personal communication with Corina Jump, 2009). A series of studies by Proctor Engineering (1997) on different equipment types was also identified. The studies generally concluded that degradation – above the degradation pattern that would be realized in standard efficiency equipment – was very unlikely for the majority of equipment types (examples included residential air conditioners and refrigerators). In some cases, the analysis suggested that the degradation associated with new efficient equipment might be less than standard / traditional equipment, leading to "negative" degradation or higher / increasing savings relative to traditional equipment. In the case of high intensity discharge (HID) lighting, small quantifiable technical degradation was suggested. The study noted that for several measures (commercial package air conditioners and oversized evaporative cooled condensers), the engineering analysis suggested that potentially significant relative technical degradation could occur – and primary research may be needed to further explore the issue. Many measures were likely to experience absolute technical degradation, but that it leads to stable or increasing savings over time compared to the parallel standard measure.

[142] Technically, each behavior for each person educated by the program has a presence, in place and operating, parallel to an EUL. There is also a TDF associated with the behavior – for example, when that person does the behavior only a share of the time or begins to forget the learned behavior. However, for ease, we treat it all under EUL. Given partial adoption, both issues will need to be considered as part of any credible EUL or TDF study.

Table5.2: Range of EUL Values Used in the US

Residential Measures	Commercial Measures
• Lighting – CFL Bulbs: 6-8 years, with some recent work starting to incorporate variations based on assumptions about hours per day that the bulb operates • Hardwired fixtures – 15-20 years for interior and exterior fixtures • Lamps (table or touchier) – 5-10 years for most studies[143], depending on type • Occupancy sensors – 10-15 years • HVAC replacement – 15-25 years • HVAC and water heating in Energy Star – 15-25 years • Room A/C – 11-15 years • Programmable thermostat – 10-12 years • Whole house fans – 25 years • Attic ventilation fans with thermostat controls – 19 years • Duct sealing and air sealing – each 15-20 years • Insulation – 20-25 years • Duct insulation – 20 years • Windows – 20-35 years • Pipe wrap – 10-20 years • Tank temperature turn down – 4-7 years[144] • Weatherization (combination measures) – 20-25 years[145]	• Lighting – CFL Bulbs – 3.4-6 years, with some recent work starting to incorporate variations based on assumption on hours per day bulb that operates in business locations • Fluorescent fixture – 11-16 years • Hardwired CFL – 10-15 years • HID (interior and exterior) 13-15 years • Occupancy sensors – 8-15 years • Daylighting dimming – 9-10 years[146] • Packaged AC/Heat Pump – 12-15 years • Chillers 19-23 years • Economizers – 7-15 years • Programmable thermostat – 5-10 years • Energy Management Systems (EMS) – 10-15 years • Motors – 13-20 years.

Our review of EULs identified several issues:[147]

- Process equipment lacks EUL studies in many cases, largely because each specific measure has a small sample size. Some utilities or agencies "assign" a 10 year lifetime, assuming that progress in the industry leads to reconfiguring of equipment on that kind of schedule. This issue may bear additional research, especially since lifetimes are likely dependent on the pace of innovations in the particular industry.[148]
- Some equipment may require evaluations of operating assumptions: for example, CFLs and other lighting equipment in commercial establishments, variable speed drives (VSD)s when applied to agricultural milking that endure harsh outside conditions, etc.[149] Lighting logger studies are particularly important given that huge shares of utility programs and savings are based on lighting measures.
- Reliable EUL estimates are missing in many key end uses: e.g., cooking, air compressor equipment, chillers, adjustable speed drives (ASDs)/VSDs, refrigeration equipment and freezers in some sectors. In addition, there is only limited information available on the increasingly important – and targeted – plug loads sector (e.g., copiers and office equipment) and unless very short lifetimes are assigned, these measures may need to have EUL studies conducted to provide justifiable savings estimates.

[143] But longer for California (9-16 years). All California numbers from the Database on Energy Efficiency Resources (DEER).
[144] One study (GDS 2007) suggested this measure needed additional study.
[145] California: 13 years
[146] California: 16 years.
[147] Based on our review of national and California EUL studies (Skumatz 2008).
[148] Think of the difference between high-tech computer chip manufacture vs. traditional steel or paper manufacture, as hypothetical extremes.
[149] In addition, some lifetimes may specifically need to be adjusted based on the influence of behavioral programs. For instance, if a program suggests relying on daylighting and leaving lights off until really needed, the operating hours for CFLs may need to be adjusted in accordance with the success of such a hypothetical program.

- There are few retention studies on building shell measures. Building shell measures are not generally assumed to be subject to widespread failure / removal, but this assumption should be verified, potentially in different parts of the country.[150]
- There has been a trend in the field to move toward simplified EUL tables, but this is a problem. Even some of the earliest research (Skumatz and Hickman 1991) found significant variations in business turnover by business type, and this turnover has a direct effect on retention of measures (particularly lighting). These variations are important factors in program savings computations and program design.[151]

The urgency of the need for additional EUL research in specific measures should be weighted by the expected future savings to be derived from the measures. For those that are rare and low savings, the priority is low. Similarly, for measures unaffected by operating hours and climate (e.g., exit signs), priority for investment of additional research budget should probably also be low. Measures subject to climate and operating hours assumptions may be higher priority (e.g., HVAC).[152] And again, waiting for failures hurts the timeliness of EUL studies, making the results less applicable to current and next generation measures that are being installed.

Retention for Non-Widget-Based Programs - Education / Training / Behavioral

Probably the single biggest gap in lifetime studies is the virtual non-existence of studies examining the retention of education, training, and behavior-focused "measures". On the behavioral side, programs tend not to get energy savings credit, so EULs / retention / persistence has not been much studied, even though the programs and their outcomes presumably do have lifetimes. Reviewing more than 100 studies in education / training (Skumatz and Green 2000; Freeman and Skumatz 2009), we found only a couple that even mentioned the topic of the retention of savings. Almost all studies examined savings for the first year of the program, which makes it hard for potentially important and dynamic education programs to receive high benefit/cost ratios, reducing likelihood of funding.

There are two studies available that addressed retention of educational messages and installation of low-cost energy-efficiency measures delivered through energy education programs. The Energy Smart Program conducted in Oregon with low-income households found strong to mild retention (about 40% after 3 years) of behavioral changes. Especially successful have been those energy education efforts that provide quality education over a longer period of time. Three energy education programs delivered in schools: the Kentucky NEED Program, the Iowa LivingWise Program and the Washington Energy Education in Schools Program show the importance of quality education and reinforcement of behavioral change messages over time. Of these three programs, the highest institution of behavioral changes are found from the Washington program where teachers conduct at least three different classroom sessions and

[150] Peach (2009) also expresses concern that within the TRC environment, new construction, design and shell measures are generally assigned lifetimes no longer than about 20 years, even though many of these measures last perhaps 100 years or longer. He feels this 20-year horizon is a problematic artifact and that the future is too discounted to reflect the actual climate imperative.

[151] Without consideration of variations by business type, programs could keep replacing measures continually in the same business types, and fixed or deemed EULs that don't vary by business type would miss this effect and keep counting streams of savings that never materialize over time.

[152] The question arises whether lifetimes for HVAC equipment should be similar between two very different areas of the country; say, the Northwest vs. Florida. Behavioral considerations should be expected to matter.

one assembly with kids over the course of an entire school year. These efforts, along with an early study by Harrigan and Gregory (1994), which found 85%-90% of the savings from the education portion of a weatherization program was retained after three years, few studies have conducted primary data analysis of the topic. Even well-funded multi-year statewide outreach programs have not examined the persistence of behavior change.

This oversight, along with the omission of cost information from most social marketing and outreach / behavior studies (Freeman and Skumatz 2009), represent significant issues associated with the evaluation of behavioral programs.[153] Unfortunately, even if first year annual savings estimates are available, it is not possible to develop reliable estimates of the benefit-cost ratio, nor is it possible to rely on long-term savings from programs that are not continually refreshed. For this reason, many utilities assign retention values no higher than three years in most cases.[154] Also of concern is that the savings and potentially the persistence may be highly variable depending on the specific program, specific media, quality of the campaign, and many other factors. It may be that every program will require its own persistence study for at least a while, until there is time to develop reliable best practices and "template" programs. The behavioral persistence topic is gaining interest,[155] and should be among the highest priorities for new research.[156]

Attributing behavioral changes or energy savings effects to particular campaigns or programs is becoming more complex as more agencies work toward similar energy efficiency behavior changes. Generally, this factor has minimal effect on the measurement of energy savings lifetimes; however, it does tend to affect in a significant way the estimate of (the share of) behavior-induced energy savings that can be clearly attributed to a specific program or intervention.[157] Research is exploring several options for behavioral programs, as noted below.

One avenue is identifying cases for which it is suitable to apply the measure-based "Best Practices" methods to the development of measure life estimates for behavioral programs. This may work, in general, with revisions to questions to ask about the persistence / presence of behaviors.[158] There are nuances related to "partial retention" (e.g., some household members), but, conceptually, this approach can apply to some programs.[159] However, there may be problems in using this approach. For example, there may be more problems with bias[160] since behaviors may not be as observable as measures. Also, the costs for conducting this type of research may be even higher than traditional EUL work, because behavioral programs may not be easily associated with specific businesses or homes. Large-scale survey approaches may be one of the few data collection options available, and these are costly.[161] Finally, the traditional EUL approach is most suited to longer-lived measures (assuming behavioral measures are less

[153] There is not a great deal of information on this point in other fields beyond energy efficiency either (Freeman and Skumatz 2009).
[154] Although it is unclear if a median EUL of 3 years can be justified given that there is minimal research for this estimate.
[155] It is gaining mention in more and more policy documents, and, for example, the Canadian Association of Evaluators has established a working committee on the topic.
[156] The associated issue of technical degradation (TDF) is probably best represented in behavioral programs as "retention", and there are certainly no extant studies of this topic separately.
[157] This topic is the subject of Skumatz (2009). [Note: this paper is not in References]
[158] For behavioral or market-based outreach / education programs that influence a home or business to purchase a measure (e.g., Energy Star programs advocating purchase of CFLs or Energy Star refrigerators, etc.), the traditional approach is appropriate.
[159] However, from a data analysis point of view, it may provide more failure data, which can assist model-fitting!
[160] Bias may depend on who in the house / business is being interviewed, and how the information is obtained (e.g., surveys where the respondent may be trying to please the interviewer).
[161] Residential appliance saturation surveys could be expanded to include behaviors, or large-scale surveys incorporating interviews on behaviors from several programs could be conducted. Surveys of rolling segments of the population may also be appropriate.

long-lived than measure-based programs).[162] Given that the lifetimes may be short, data collection might also have to be more frequent. In conclusion, simplified approaches – perhaps as straightforward as the kind of retention study conducted by Harrigan (1994) – may be more appropriate for lower-budget programs. And random assignment, follow-up of test and control groups, and similar methods to estimate retained shares of savings and behaviors are the least that is needed. Large scale surveys of households or business populations may be needed until reasonable estimates can be derived and some kind of convergence in results by type of program emerges.

Upstream

This previous discussion largely considered "direct" behavioral / educational programs – those related to occupants of the home or business. However, there is the issue of retention of "upstream" behavioral / educational programs. Technical degradation of upstream training programs offered to agents that do not actually operate the measures (e.g., equipment vendors, manufacturers, commissioning agents, builders, and architects and engineers) is another matter. To the extent that these programs work to influence second and third round and future savings, then EUL and TDF is an important consideration and very hard to measure (VHTM).[163] The TDF may decay, but presumably, it may increase as the builder / agent is inspired to take on more and more (self-) education and measures as a result. This delves into the realm of spillover, but it can also be viewed under the subject of EUL / TDF. This is a topic that has not been studied and represents a particular challenge for developing credible methods.

Summary

Table 5.3 summarizes key patterns in EUL results.

Table 5.3: Variations in EULs by Program Type and Region

	EULs
General results	After early work in the Northwest, results broadly have gravitated toward values fairly similar to those in California's protocols, with some variations elsewhere. The State of California required *ex post* statistical verification, leading to minor refinements. There are a number of measures for which there are missing or inadequate data; the most glaring example is the nearly complete omission of retention information or estimates for behavioral programs.
Variations by Program type	Almost all EUL results are by measure, not by program design or incentive provided. Therefore, although measures have EULs, there are no variations for measures installed from programs designed as rebate vs. codes / standards, etc. Any program delivering a measure receives basically the same retention value for that measure.
Variations for behavioral vs. measure-based programs	There is almost no information for retention of behavioral programs including education / training, commissioning training, and similar programs. Widget-based programs have fairly thorough EUL information, with omissions for some measures (cooking, some shell, and others listed in the tables).

[162] This seems sensible because people move from the home (and potentially the service territory) about every 5 years and take their program-influenced behaviors with them. In contrast, most measure-based programs are permanent to the home (except refrigerators and CFLs) and remain after the occupant moves.

[163] A step beyond the hard to measure (HTM) effects.

5.3 Issues / Problems Identified

This chapter reviewed the literature and status of work on measure lifetimes and provided information on a number of key topics in persistence. The research found:

- **Problems and best practice suggestions for EUL studies:** The study addressed some of the key issues that have hampered EUL studies in the past. Of particular note are assuring that implementation databases are better structured to support evaluation research; using appropriate sampling approaches when bundled programs are implemented; using phone data collection only when measures are unique or memorable; using panel surveys if possible; testing multiple model specifications; and, most importantly, benchmarking the results against the findings for earlier years of the program and for similar programs around the nation.

- **Results and Gaps in EULs**: A review of results from measure-based EUL studies around North America showed that measure lifetimes exist and are fairly consistent for many measure-based programs in residential and non-residential sectors. Relatively similar EUL values are being assigned by utilities across the country – perhaps with not enough recognition of the operational hours that vary by climate zone. The review also shows a lack of depth in studies in process equipment, some shell measures, and specific end-uses like cooking, refrigeration, and air compressors. There was some concern expressed about the trend toward aggregating lifetimes to broader categories. Measure-specific information is important, as the lifetime can be application- and usage- specific (e.g. agricultural pumps, lighting, etc.). Suggestions were made that EUL results would benefit from more frequent post installation surveys or measurement for certain measures (Ogle 2009, Blasnik 2009, Mengelberg 2009).

- **Technical Degradation**: The issue of technical degradation was discussed, and there is a shortage of primary research on this topic. Certainly, engineering-type studies can help to identify research priorities to some extent, noting which technologies have had engineering, mechanical, or process change that will more likely significantly change their performance relative to standard equipment. However, equipment with significant changes in behavioral (operational or upkeep) elements may also see changes in performance. Priority-setting for new research on this topic should take both factors into account (mechanical and behavioral), and resulting figures should be verified periodically.

- **RUL Issues**: Regarding the topic of Remaining Useful Lifetimes (RULs), some utilities argue RULs are critical to certain programs, while others don't feel the estimation complexity is a worthwhile expenditure. The jury is still out on the policies to be applied broadly, but if a program is designed as early replacement, a credible case could be made that its savings pattern is significantly altered from end-of-lifetime programs. Perhaps in the short run, presenting benefit cost figures including and excluding the enhanced savings could be presented to identify whether the programs are moving decisions forward enough to make a difference. There are potentially cases in which this analysis would also be applied to behavioral programs.

- **Retention of Behavioral Changes Results and Needs**: Of particular note is the virtual absence of studies addressing retention or persistence of education / outreach / behavioral programs. This is an important gap as behavioral and market-based

programs become a larger and larger share of utility / agency portfolios. Further research in best practices for the array of behavioral programs or "types" would be a useful addition to the literature, and agencies should consider requiring new behavioral programs to conduct retention assessments every year or two for a period reaching on the order of three or more years out. This may be the only way to gain enough information to develop credible estimates of the persistence of savings from behavioral programs, and allow more serious consideration of them as reliable resource substitutes. The issue of retention of behaviors and savings for "upstream" education and training programs is particularly troublesome and, to the degree that these programs are part of portfolios, retention work is needed where there currently is none. Finally, EUL measurement approaches will need to be tested and applied to a variety of behavioral programs. Some may parallel traditional EUL estimation best practices, but the application of statistical approaches to some programs may be challenging. This research should be a priority for the near term.

Measure lifetimes are a key element in the computation of program savings. It is important to assure that new programs are developed – including creative programs and programs that encourage new measures and behaviors and are not the "same old same old". However, if measure lifetimes, TDF, and other factors are known for some programs and unknown up front for others, there will be a bias away from developing new (more uncertain) programs. Risk is an issue affecting investment and development.

Risk needs to be considered from two perspectives - providing up-front information on computational elements encourages program development. "True-up" is needed for credibility and reliability of savings estimates for EE relative to generation capacity. One suggestion may be that new programs are assigned a deemed lifetime by general "type" up front, and then after 1-2 years, a true-up is prepared that does not readjust program incentives retroactively, but does refine the estimate of future savings from a resource perspective.

Identifying the lifetimes or EULs of behavioral or information programs is complicated as more media on behavioral and education bleeds across territories. This affects retention of the messages and behaviors because behaviors originally attributable to the program may be "refreshed" from other sources. It may not be possible to separate these out cleanly; how large a practical problem this is requires research. The priority depends on the ranking of estimated savings and costs from these programs

5.4 What Has Been Learned: Emerging Approaches and Experience

The literature on approaches to measure lifetimes has been fairly dormant over the past few years, reflecting the fact that there seems to be fairly consistent agreement that the estimation approaches are defensible and appropriate. Some issues were raised in the best practices discussion (test other distributions, incorporate explanatory variables, etc.). However, the biggest challenge has yet to be addressed: whether current measurement methods are well-suited to behavioral programs, particularly when behavioral programs can be implemented part-time by participants. On the face of it, it appears that standard measurement methods may work; however, whether appropriate statistical properties are retained and exactly how some of the imprecision of partial participation is incorporated into the modeling needs further exploration.

Most importantly, any retention work – even fairly simple work – is a very high priority for behavioral and related programs.

5.5 Conclusions and Additional Research Needed

Conclusions, recommendations, needed research, and other key issues uncovered in the analysis are detailed below.

5.5.1 Conclusions

- EUL research methods are fairly established, defensible, tested, and consistent. California's practice of checking *ex post* results against *ex ante* figures provides some confidence in deemed lifetime assumptions for key measures (those top four representing 50% or more of the savings in various programs).
- Some variations in EULs from climate are appropriate (space conditioning), and a few are represented in differences in lifetimes assumed in the northwest vs. New England.
- Rather than moving toward simplified EULs (e.g. assigning the same EUL for all non-residential applications), numbers tailored to some degree are necessary to limit "churning." For example, research consistently shows greater turnover in some business sectors which can affect measure presence and usage, and thus, attributable savings.
- There is considerable difficulty identifying retention of behavior and education associated with upstream actors. It seems unfair to assume zero retention, or these programs potentially may not receive appropriate funding (high or low). However, evidence is needed to identify other values, particularly as there is no reason to believe the retention patterns will follow those of the direct education to savings program participants.
- Technical degradation in a behavioral or educational program may definitionally be indistinguishable from the EUL. In fact, an EUL may better be described as TDF in behavioral and educational circumstances. Each behavior for each person educated by the program has a presence, in place and operating, parallel to an EUL. There is also a TDF associated – for example, when that person undertakes the behavior only a share of the time or begins to forget the learned behavior.
- Remaining useful lifetimes (RUL) has been the focus of some discussion, but generally has not elevated to a priority by industry.
- Remaining useful lifetimes (RUL) are a valid consideration. Logic dictates that programs that remove equipment early should be credited with higher savings during the period between the date of removal and the date equipment otherwise would have been removed, falling back to the difference between new savings and new "standard" equipment for the remaining EUL period.
- In the short run, planners might be asked to submit benefit cost figures including and excluding the enhanced savings to identify whether the programs are moving decisions forward enough to make a difference. If the program is on the brink, negotiations may be used to identity appropriate assumptions for the specific program.

- Early adoption of new behavior is an appropriate concept. Current behaviors are not going to stay the baseline forever, especially with mass media attention on green behaviors and potential for mandates, etc. Although theoretically, this will tend to decrease savings associated with programs, it will depend on whether behavioral persistence (EUL) is longer than the baseline new behavior adoption.
- Near term policy tradeoffs probably dictate assuming programs that encourage adoption of new behaviors get full credit compared to today's behavior for a "safe" period – perhaps on the order of up to 3 years. This is practical, but gives credit to these programs. As an alternative, perhaps in the short run, presenting benefit cost figures including and excluding the enhanced savings could be presented to identify whether the programs are moving decisions forward enough to make a difference.

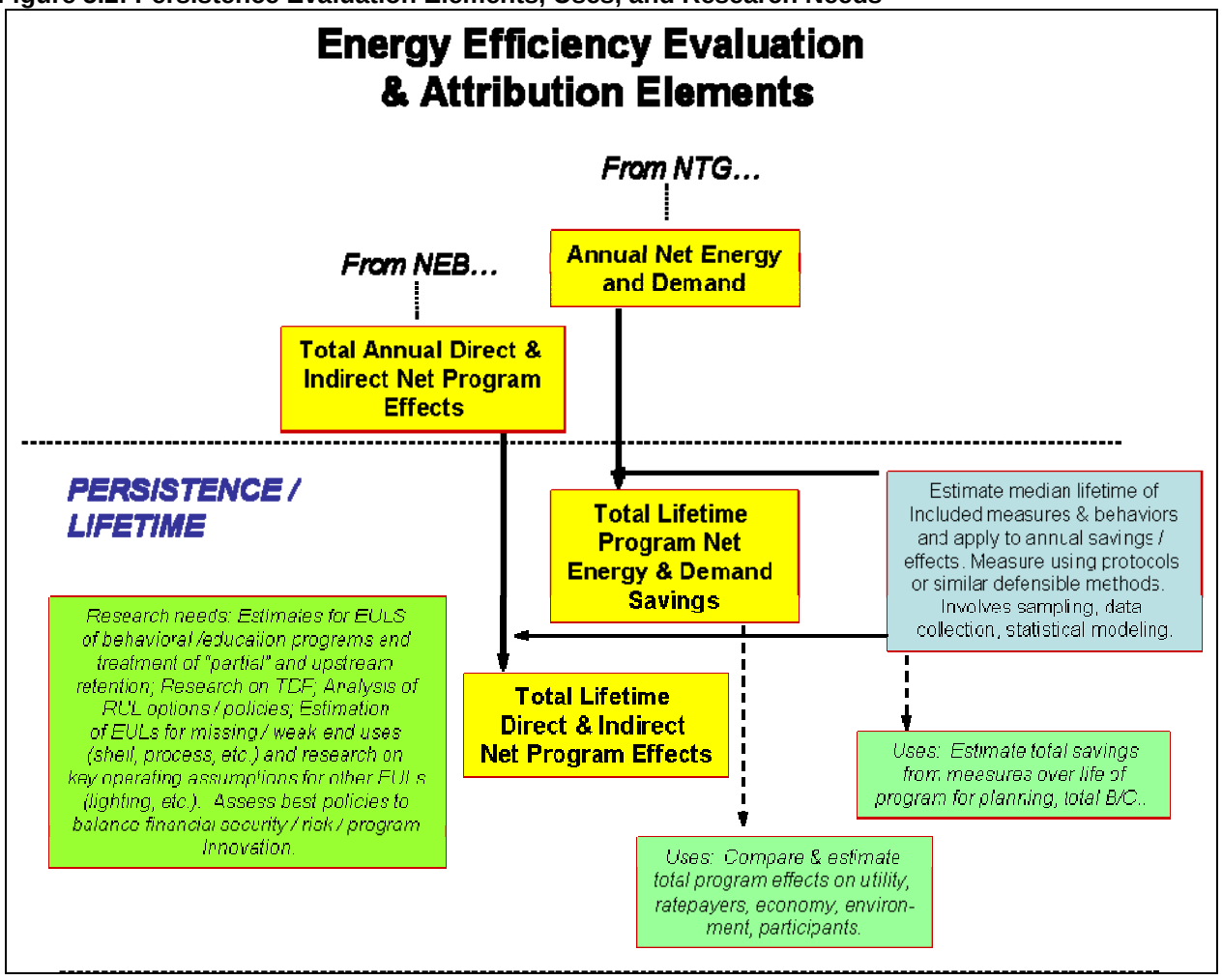
5.5.2 Additional Research Needed

- Improvements in standard practice should center around better design of upfront data bases for evaluation, improved sampling, and especially, benchmarking against previous research on the program, and on similar programs elsewhere to explain similarities/differences.
- EUL estimates are missing or not strong in several areas and need study, potentially at the expense of yet more EUL research on measures with already strong results. Research is needed on: cooking, air compressor equipment, chillers, adjustable speed drives/variable speed drives (ASDs/VSDs), and refrigeration equipment and freezers in some sectors.
- There is only limited information available on the increasingly important – and targeted – plug loads sector (including copiers and office equipment) and unless very short lifetimes are assigned, EUL studies are needed to provide justifiable savings estimates. Many of these measures are also critically linked to behavioral issues and need refined research methods incorporating factors addressing these sources of differences in performance.
- There are few retention studies on building shell measures. They are not generally assumed to be subject to widespread failure/removal, but this assumption should be verified by studies, potentially in different parts of the country.
- More study of usage (behavioral) patterns needs to be incorporated for some measures (e.g., lighting logger studies, space conditioning variations, and VSDs in varying operating conditions).
- More use of explanatory variables and consideration of different distributions in EUL modeling is needed and important. This can incorporate behavioral and climate patterns that are linked closely with lifetimes for many measures and allow better benchmarking and comparison across regions – and allow understanding and rationalization of appropriate differences.
- Periodic checking of measure lifetimes into the future is essential to provide confidence that measures are being retained and remain operating. It is necessary to allow programs to spot problems like early removals or failures of new technologies to allow correction and to appropriately account for these issues in savings computations (resource needs implications). At least one of the retention checks should be fairly early in the lifetime, and of course, some will be needed in outlying years to confirm model selection and decay pattern.

- There is virtually no work on retention of education and behavioral measures with the exception of the lifetimes of the equipment purchased as a result of marketing – e.g., refrigerators. This is a very large gap in the literature and should be a focus of future work.
- Research is essential on retention of messages and behavior and associated energy savings. As these sorts of EE programs become increasingly important in the portfolio mix, and as EE works to continue as a substitute to generation, then confidence in the performance and lifetimes of proposed initiatives is needed.
- Research is needed on retention of commissioning and other training programs. This is needed to identify the retention of the training message (for the multiple levels of agents trained), and for the ultimate on-site operation of the buildings once built (for the custodial and operational staff).
- Measure lifetime methodologies parallel to those used for widget-based programs may generally work but will be complicated by the treatment of partial behavior retention (half of the household members all of the time, etc.). Recommended methods should be developed and tested / explored, and the size in variation of results based on variation in models should be clearly delineated to identify whether simple approaches are only theoretically flawed vs. flawed in ways affecting meaningful application of the results to evaluation.
- Simple studies of retention should be conducted as soon as possible for potentially all major types of education and behavioral programs that have been offered for the last several years for which past participants can be located (e.g., weatherization/education programs, commissioning, etc.). The targeted programs definitely should include those being counted on for large savings. Basic survey methods or interviews may provide some indicative and useful information related to retention of messages and behaviors.
- It is expected that "lifetimes" may ultimately be established for various program "types," and maybe even included in databases like DEER. However, there will always be good and poor quality educational efforts, and this will require the use of ranges of months or years for savings at the very least, with some protocol for correcting within (or outside) the range based on actual performance for the specific program or in the shorter term, based on early indicators of other types.
- This is a topic in which further research is needed – both engineering and primary performance data. If funds are short, it can particularly be targeted to several key cases: (1) when there are suspicions that equipment performance may change because of substantial engineering, design, mechanical, material, manufacture, or lifetime changes; (2) when performance of the standard equipment changes; and (3) when a measure represents a significant share of the energy savings for programs, portfolios, or beyond. Even an updated engineering analysis of key residential and commercial measures would be helpful.
- TDF studies should be conducted if new technologies are incorporated (like LEDs).
- TDF studies need to incorporate specific analyses of the role and influence of behavioral characteristics in the performance of measures – and the influence that behavioral factors have on the performance.
- TDF factors should be considered in computations for those (significant) measures that decay differently than older technologies. This is a key part of providing a reliable and well-counted generation alternative.

- Research on the years before a new behavior would otherwise have been adopted is complicated and completely program and message dependent. It is unlikely any standard or deemed numbers could be generated, but it is hard to imagine how measured numbers for each practice could be devised. Adoption curve research may be an appropriate method (needs research), but the scope is difficult to pin down and potentially vast.

Figure 5.2: Persistence Evaluation Elements, Uses, and Research Needs



6. REFERENCES

6.1 Impact Evaluation

Agnew, Burke & Ham-Su,2007. "Participation of Demand-Response Resources in ISO New England's Ancillary Service Markets", 2007 Energy Program Evaluation Conference, Chicago.

Barata, S, 2006. "The New Hampshire Electric Utilities' Low Income Retrofit Program - Impact Evaluation". Final Report by Opinion Dynamics, January 2006.

Barbeiri & Swan, 2007. "Myth Busting Savings Calculations", 2007 Energy Program Evaluation Conference, Chicago.

Bernier, Clark & Metoyer, Jarred, 2007. "Cracking the Code for Residential New Construction: Using End-Use Metered data to Revise Energy Estimates of Compliance Models". 2007 Energy Program Evaluation Conference, Chicago.

Bruchs, Doug, Michelle Levy, and Swarupa Ganguli. 2006. "Redefining Homework: Are the Green Schools Program Students Taking Energy Efficiency Home with Them?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Burnham, Kenneth and David Anderson. 2002. "Model selection and multimodal inference". New York: Springer-Verlag.

Cooney, Kevin. 7/31/08. Summit Blue, Boulder, CO, Personal Communication with the Author.

Degens, Phil. 4/21/09. Oregon Trust, Portland, OR, Personal Communication with the Author.

Dimetrosky, Scott, Bicknell, Charlie, and Titus, Elizabeth. 2007. "Filling Gaps In The Story Of Energy-Efficiency Program Success: Evaluating the Availability of Market Penetration Tracking Data for the Residential Sector". AESP White Paper: March 2007

Dimetrosky, Scott. 4/20/09. Cadmus Group, Boulder, Co, Personal Communication with the Author.

Dohrmann, Don, John Peterson, John Reed, Shahanna Samiullah, and Steve Westberg. 2007. "Net Savings Estimation in Appliance Recycling Programs: A Review and Empirical Analysis with Recent CA Data", Proceedings of the IEPEC Conference.

Drakos, Khawaja, & West, 2007. "Flipping the Switch: Evaluating the Effectiveness of Low-Income Residential Programs". 2007 Energy Program Evaluation Conference, Chicago.

Efficiency Valuation Organization (EVO). 2002. "International Performance Measurement & Verification Protocol, Volume I: Concepts and Options for Determining Savings", www.evo-world.org/ipmvp.php

Hall, Nick, Pete Jacobs, and Steve Kromer. 2006. "Improving the Reliability of Energy Efficiency Portfolio and Program Evaluation Via a Risk Assessment Approach". Prepared for the Energy Division of the California Public Utilities Commission, 2006.

Heschong Mahone Group, 2006. "Sidelighting Photocontrols Study", Prepared for Northwest Energy Efficiency Alliance, March 22, 2006.

Khawaja, M. Sami, 10/3/08, 11/23/09. Cadmus Group, Portland, OR Personal Communication with the Author.

Kmenta, Jan, and James B. Ramsey. 1980. Evaluation of econometric models. New York: Academic Press.

Lee, A.D., D. Kavanaugh, M.K. Gobris, S. Boughen, and J. Staples. 2002. "Searching for Impacts: Two Innovative Approaches to Measure the Effects of a Residential Energy-Efficiency Program", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

McQuarrie, Allan D. R., Tsai, C.L. 1998. Regression and time series model selection. Singapore: World Scientific.
Messenger, Michael. 7/29/08 and 4/3/09. Itron, Sacramento, CA, Personal Communication with the Author.
Michaud, Norman, Lori Megdal, Pierre Baillargeon, and Carl Acocella. 2009. "Billing Analysis & Environment that "Re-Sets" Savings for Programmable Thermostats in New Homes", Proceedings of the IEPEC Conference.
Mulholland, Carol. 4/21/09. Cadmus Group, Wa DC, Personal Communication with the Author.
Nadel, Steven, Anna Monis Shipley, and R. Neal Elliott. 2004. "The Technical, Economic, and Achievable Potential for Energy Efficiency in the United States: A Meta-Analysis of Recent Studies", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Nexus Market Research, Inc. et al. 2005. "Market Progress and Evaluation Report (MPER) for the 2004 Massachusetts ENERGY STAR® Appliances Program", May 23. Prepared for Cape Light Compact, Massachusetts Electric Company, Nantucket Electric Company, NSTAR Electric, Western Massachusetts Electric Company, Fitchburg Gas and Electric Light Company.
Parlin, Kathryn and Larry Haugh, 2007. "Eliminating the Guesswork: The Information-Theoretic Approach to Model Selection", 2007 International Energy Performance and Evaluation Conference.
Patil,Yogesh, Dan Barbeiri, and Gail Azulay. 2009. "Taking Engineering Savings to the Next Level", Proceedings of the IEPEC Conference.
Peach, Hugh "Gil". 4/21/09. H Gil Peach & Associates, Personal Communication with the Author.
Peters, J., McRae, M., Morander, L. & D. O'Brien. 2000. "Detecting Behavioral Change from a Visit to a Children's Museum Energy Conservation Exhibit," Proceedings of the 2000 ACEEE Summer Study, pp. 8.281-292, Asilomar, CA.
Pratt & Miller,1998. "Estimated Refrigerator Loads in Public Housing". Prepared for U.S. DOE, August 1998.
Sabo, Carol, Andrews, Lee & Bakalars, 2007. "Benchmarking and Best Practices in Power Management of Computers and other Plug Loads on Campus", 2007 Energy Program Evaluation Conference, Chicago.
Sabo, Carol. 11/21/09. PA Government Services, VI, Personal Communication with the Author.
Schonder, Hughes, Sweitzer & Schmoyer, 2007."Methodology for the Evaluation of an Energy Savings Performance Contracting Program for the Federal Government", 2007 Energy Program Evaluation Conference, Chicago.
Scott, Steven & Stout, Jennifer, 2005. "Impact Evaluation of Oregon Industrial Transition Projects". Final report submitted to Energy Trust of Oregon, January, 2005.
Select Energy Services, Inc., 2004. "Analysis of Cooler Control Energy Conservation Measures". Final report Submitted to NSTAR Electric, March 3, 2004.
Skumatz, Lisa. 2006. "Tracking Market Progress: Addressing Traditional Measurement Flaws with an Innovative Proxy -the "NEEPP" Tracking Approach", AESP White Paper: January 2007.
Skumatz, Lisa and John Green. 2000. "Evaluating the Impacts of Education/Outreach Programs - Lessons on Impacts, Methods, and Optimal Education". Proceedings from ACEEE Conference, Asilomar, CA.
Skumatz, Lisa A., and John Gardner, 2005. "Decomposing price differentials due to ENERGY STAR® lavels and energy efficiency features in appliances: Proxy for market share tracking?", ECEEE 2005 Summer Study proceedings, Cote d'Azur, France.

Skumatz, Lisa A.,et.al. 2006. "Incremental / Hedonic Price Analysis: Cost-Effective Methods of Tracking Program Impacts Over Time", Proceedings for the ACEEE Summer Study on Buildings, Asilomar, CA, August 2006.

6.2 Net-To-Gross / Attribution

Albert, Scott, 4/9/09. GDS Associates, Marietta, GA, Personal Communication with the Author.

Albert, Scott, Lori Megdal, and Victoria Engel. 2006. "Using Residential Sector-Level Logic Models to Improve the Design, Implementation, and Evaluation of EE Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Anderson, Marge. 2004. "Education by Design: Creating Lasting Market Behavior Change through Education & Training", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Austin, Cynthia, Catherine Chappell, Bill Knox, and Marshall Hunt. 2005. "A Regional Approach to Energy Efficiency", Proceedings of the IEPEC Conference.

Baker, David S., Illinois Department of Commerce and Economic Opportunity, Energy Division. Interview with the author, November 20, 2008.

Bender, Sylvia L., Mithra Moezzi, Marcia Hill Gossard and Loren Lutzenhiser. (no date). "Using Mass Media to Influence Energy Consumption Behavior: California's 2001 Flex Your Power Campaign as a Case Study", Internet search.

Bender, Sylvia, Adrienne Kandel, and Sy Goldstone, 2004. "Behavioral Economics: The Link Between Human Dimensions and Market Transformation", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Bensch, Ingo, 3/31/09. Energy Center of Wisconsin, WI, Personal Communication with the Author.

Bensch, Ingo, Scott Pigg, and Marge Anderson, 2006. "How Much Is That Training Program Worth? Quantifying the Value of Training and Other 'Fuzzy' Education Events", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Bensch, Ingo. 2008. "Is Climate Change a Good Thing? Opportunities and Barriers to Using Climate Change to Motivate Efficiency", Proceedings of the ACEEE Summer Study on Energy Efficiency in Buildings.

Bicknell, Charles, Scott Dimetrosky, and Jim Thayer. 2008. "National Efficiency Benchmarking Study for Residential Central Air-Conditioning", Proceedings of the AESP Conference.

Blonz, Josh, Danny Morris, and Andy Stevenson. 2008. "Energy Efficiency and Behavioral Economics; Common Tragedies", From Web Page printout, April.

Bordner, Robert D., Robert M. Wirtshafter, and Mary Wold. 2004. "Multifamily Markets: Hard-to-Reach and Hard-to-Serve", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Brateng, Eric, 4/3/09. Puget Sound Energy, Bellevue, WA, Personal Communication with the Author.

Brown, Marilyn A., Linda G. Berry, Richard A. Balzer, and Ellen Faby, 1993, National Impacts of the Weatherization Assistance Program in Single-Family and Small Multifamily Dwellings, Oak Ridge National Laboratory, Oak Ridge, TN, ORNL/CON-379, May 1993.

Chappell, Catherine, Doug Mahone, Marian Brown, Kenneth Keating, and Lori Megdal. 2005. "Net Savings in Non-Residential New Construction: Is a Market Based Approach the Answer?", Proceedings of the IEPEC Conference.

Coito, Fred, 7/25/08. KEMA, Oakland, CA, Personal Communication with the Author.

Cook, Gay, 4/5/09. Summit Blue Canada, Vancouver, WA, Personal Communication with the Author.

Cook, Gay. 2008. "Attribution Methodology Wars: Self-Reported Methods versus Statistical Number Crunching- Which Should Win?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Degens, Phil. 4/21/09. Oregon Trust, Portland, OR, Personal Communication with the Author.
Dimetrosky, Scott, Cadmus Group, Boulder, CO, Personal Communication with the Author, 4/20/09.
Dohrmann, Don, John Peterson, John Reed, Shahanna Samiullah, and Steve Westberg. 2007. "Net Savings Estimation in Appliance Recycling Programs: A Review and Empirical Analysis with Recent CA Data", Proceedings of the IEPEC Conference.
Dougherty, Anne, Katherine Van Dusen Randazzo, and Pamela Wellner. 2009. "Using Structural Equation Modeling (SEM) to Identify, Tease Out, and Quantify a Marketing Program's Influence on Energy Efficiency Intentions and Behaviors", Proceedings of the IEPEC Conference.
Drakos, Jamie, 3/31/09. Cadmus Group, Portland, OR, Personal Communication with the Author.
Dyson, Christopher, Shahana Samiullah, Tami Rasmussen, and John Cavalli. 2005. "Can Programmable Thermostats Be Part of a Cost-Effective Residential Program Portfolio?", Proceedings of the IEPEC Conference.
Ehrhardt-Martinez, Karen. 2008. "Dollars or Sense: Economic versus Social Rationality in Residential Energy Consumption", From Energy Collaborative Analysis Initiative, Web Forum, August.
Erickson, Jeff, Mary Klos, and ValyGoepfrich. 2009. "Free Ridership: Arbitrary Algorithms vs. Consistant Calculations", Proceedings of the IEPEC Conference.
Erickson, Jeff. 2008. "Preaching to the Choir: Are Repeat Participants Free Riders?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Erickson, Jeff. 2008. "Preaching to the Choir: Are Repeat Participants Free Riders?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Fagan, Jennifer, Itron, Madison, WI, Personal Communication with the Author, 11/17/08 and 12/4/08.
Fagan, Jennifer, Mike Messenger, Mike Rufo, and Peter Lai. 2009. "A Meta-Analysis of Net to Gross Estimates in California", Proceedings of the AESP Conference.
Friedmann, Rafael, and Chris Ann Dickerson. 2009. "Options for Improving Energy Efficiency Evaluation in California: "Houston, We Have a Problem", Proceedings of the IEPEC Conference.
Friedmann, Rafael. 2007. "Maximizing Societal Uptake of EE in the New Millennium: Time for NTG to Get Out of the Way?". Proceedings of the IEPEC Conference.
Friedmann, Rafael. 2008. "Energy Efficiency Best Practices: What's New?", From The National Energy Efficiency Best Practices Study, July.
Gardner, John, and Lisa Skumatz. 2006. "Actual Versus Perceived Energy Savings: Results from a Low-Income Weatherization Program", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Gordon, Fred, 7/25/08. Oregon Trust, Portland, OR, Personal Communication with the Author.
Gordon, Susie and Lisa A. Skumatz, Ph.D., 2007. "Integrated, Real Time (IRT), On-Going Data Collection For Evaluation – Benefits And Comparative Results", Proceedings for the European Council for an Energy Efficient Economy (ECEEE), June 2007, France.
Hedman, Brian. 3/31/09. Cadmus Group Portland, OR, Personal Communication with the Author.
Hoefgen, Lynn, Susan Oman, Angela Li, Gail Azulay, and Ralph Prahl. 2008. "Market Effects: Claim Them Now or Forever Hold Your Peace", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Jackson, Carla, Jane Peters, Mersiha Spahic, and Susan Lutzenhiser. 2009. "Trends in ENERGY STAR Awareness: Results from Four National Surveys, 2002-2008", Proceedings of the IEPEC Conference.
Jelsma, Jaap. 2004. "The Engineering Approach and Social Aspects of Energy Use: Mind the Gap, but Can It Be Closed?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Johnson, Alissa, and Kathleen Gaffney. 2009. "Residential Lighting Metering Study: Detailed Methods and Preliminary Lighting Inventory Results", Proceedings of the IEPEC Conference.
Kandel, Adrienne V. 2002. "Theory-Based Estimation of Energy Savings from DSM, Spillover, and Market Transformation Programs Using Survey and Billing Data", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Keating, Kenneth M. 2009. "Freeridership Borscht: Don't Salt the Soup", Proceedings of the IEPEC Conference.
Khawaja, M. Sami, 10/3/08, 11/23/09. Cadmus Group, Portland, OR Personal Communication with the Author.
Kim, Helen, and Richard Ridge. 2005. "Benefit Cost Analysis of a Portfolio of Energy Efficiency Programs B/C Ratios Calculated at the Program, Sector, and Portfolio Levels", Proceedings of the IEPEC Conference.
LeBlanc, Bill, Heather Ramsey, Ray Kolynhuk, and Professor Loren Lutzenhiser. 2007. "Program and Literature Summaries in Support of Social Marketing: Market Review", Prepared for Ontario Power Authority by Summit Blue Consulting, LLC, January.
Ledyard, Thomas, Dimple Gandhi, and Ralph Prahl. 2009. "In It for the Long Haul: The Challenges of a Seven- Year Effort to Assess the Market Effects of a Non- Residential New Construction Program", Proceedings of the IEPEC Conference.
Lee, Lark, Lynn Westerlind, and Laura Schauer. 2009. "Stay Ahead of the Curve! Responding to Shifting Baselines", Proceedings of the IEPEC Conference.
Lutzenhiser, Susan, Jane Peters, Mithra Moezzi, and James Woods. 2009 "Beyond the Price Effect in Time-of-Use Programs: Results from a Municipal Utility Pilot, 2007-2008", Proceedings of the IEPEC Conference.
Mahone, Douglas. 2008. Email from Douglas Mahone to Robert Kasman (PG&E), June 12, 2008, provided to author.
Martinez, Mark S., and Craig Williamson. 2005. "California Information Display Pilot (Energy Orb) Evaluation. What Effect Does Enhanced Information Have on Critical Peak Price Response?", Proceedings of the IEPEC Conference.
McRae, Majorie, Jane S.Peters, Mary Sutter, Richard Ridge, and Ben Bronfman. 2005. "Efficient Building Equipment in Oregon: What They Got & How They Got It", Proceedings of the IEPEC Conference.
Megdal, Lori, Yogesh Patil, Cherie Gregoire, Jennifer Meissner, and Kathryn Parlin. 2009. "Feasting at the Ultimate Enhanced Free-Ridership Salad Bar", Proceedings of the IEPEC Conference.
Meissner, Jennifer, Cherie Gregoire, Steven Meyers, Lori Megdal, and Kathryn Parlin. 2008. "Allocating Impact Evaluation Resources: Using Risk Analysis to get the Biggest Bang for your Buck", Proceedings of the AESP Conference.
Mengelberg, Ulrike, 3/31/09. Cadmus Group, Portland, Or, Personal Communication with the Author.
Messenger, Michael. 7/29/08 and 4/3/09. Itron, Sacramento, CA, Personal Communication with the Author.
Moran, Dulane, Jane Peters, Shahana Samiullah, Corina Jump, and James Hirsch. 2008. "CFL Program Strategy Review: No Programmatic 'Silver Bullet", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Mulholland, Carol. 4/21/09 and 3/31/09. Cadmus Group, Wa DC, Personal Communication with the Author.
Myers, Jody, and Lisa Skumatz. 2006. "Evaluation Attribution, Causality, NEBs, and Cost Effectiveness in Multifamily Programs:Enhanced Techniques", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Ogle, Rick, 3/31/09. Cadmus Group, Portland, Or, Personal Communication with the Author.
Peters, Jane and Majorie McRae. 2008. "Free-Ridership Measurement Is Out of Sync with Program Logic…or, We've got the structure built ,but what's its foundation?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Peters, Jane. Research into Action, Portland, OR, Personal Communication with the Author, 10/3/08
Quantec, Scott Dimetrosky. 2008. "Assessment of Energy and Capacity Savings Potential in Iowa", Prepared for the Iowa Utility Association, February 15, 2008.
Rasmussen, Tami, KEMA, Oakland, CA, Personal Communication with the Author, 4/20/09.
Ridge, Richard, Steve Kromer, Steve Meyers, et. al. 2007. "EE Portfolio Risk Mgmt: A Systematic Data-Driven Approach for Timely Interventions to Maximize Results", Proceedings of the IEPEC Conference.
Ridge, Richard, Phillipus Willems, and Jennifer Fagan. 2009. "Self-Report Methods for Estimating Net-to-Gross Ratios in California: Honest!", Proceedings of the AESP Conference.
Ridge, Richard, Phillipus Willems, Jennifer Fagan, and Katherine Randazzo. 2009. "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ration", Proceedings of the IEPEC Conference.
Riggert, Jeff, Nick Hall, John Reed, and Andrew Oh. 2000. "Non-Energy Benefits of Weatherization and Low-Income Residential -Programs:The 1999 Mega-Meta-Study", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Rosenberg, Mitchell, Ivin Rhyne, and Sandra Fromm. 2009. "What's the NPV of R&D? Benefit-Cost Assessment of a Comprehensive Energy Research and Development Program", Proceedings of the IEPEC Conference.
Ross, Lynn, National Grid. 2008. "As the World of Commercial HVAC Turns…", Proceedings of the AESP Conference.
Rufo, Mike, 2009. "Evaluation and Performance Incentives: Seeking Paths to (Relatively) Peaceful Coexistence." From the proceedings of the IEPEC Conference, August 2009, Portland, OR.
Rufo, Michael, Mary 0 'Drain, Allen Lee, John Cavalli, and Julia Larkin. 2000. "Market Assessment and Evaluation of California's 1999 Small and Medium Nonresidential Energy Efficiency Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Sabo, Carol. 11/21/09. PA Government Services, VI, Personal Communication with the Author.
Saxonis, William. 2007. "Free Ridership and Spillover: A Regulatory Delimma", Proceedings of the IEPEC Conference.
Schare, Stuart and Jennifer Ellefsen. 2007. "Advancing the 'Science' of Free Ridership Estimation: An Evolution of the Self-Report Method for New York Energy $mart Programs", Proceedings of the AESP Conference.
Sebold, Fredrick D., Alan Fields, Lisa Skumatz, Shel Feldman, Miriam Goldberg, Ken Keating, and Jane Peters. 2001. "A Framework for Planning and Assessing Publicly Funded Energy Efficiency", From PG&E Study ID PG&E-SW040, March.
Sipe, Brien, and Sarah Castor. 2009. "The Net Impact of Home Energy Feedback Devices", Proceedings of the IEPEC Conference.

Skumatz, Lisa A., Ph.D., 2005. "Techniques for Getting the Most from and Evaluation: Review of Methods and Results for Attributing Progress, Non-Energy Benefits, Net to Gross, and Cost-Benefit", proceedings of the ECEEE Conference, Cote d'Azur, France, May 2005.
Skumatz, Lisa A., Ph.D., 2005. Comparing Award Mechanisms - What Works?, Proceedings of the 2005 International Energy Program Evaluation Conference, Brooklyn, NY, August 2005
Skumatz, Lisa. 2007. "Attributable Effects from Information and Outreach Programs: Net to Gross, NEBs and Beyond", Proceedings of the ECEEE Conference.
Skumatz, Lisa A., and Owen Howlett, 2006. "Findings and Gaps in CFL Evaluation Research", Proceedings of the 2006 EEDAL Conference, London, England, June 2006.
Skumatz, Lisa, Dan Violette, and Rose A. Woods, 2004. "Successful Techniques for Identifying, Measuring, and Attributing Casualty in Efficiency and Transformation Programs." From proceedings of ACEEE Conference, Asilomar, CA.
Stern, Paul C. 2006. "Why Social and Behavioral Science Research is Critical to Meeting California's Climate Challenges", From Talk to the California Energy Commission, December.
Sulyma, Iris M., Power Smart, BC Hydro, Vancouver, Canada, Personal Communication with the Author, 4/28/09.
Tiedemann, Ken, Maliki Nanduri, Jean-Francois Bilodeau, and Jack Habart. 2005. "Home Energy Audits, EE and Carbon Dioxide Emissions", Proceedings of the IEPEC Conference.
Tiedemann, Ken, Iris Sulyma, and Mark Rebman. 2009. "Measuring the Impact of Time of Use Rates on Peak and Off-peak Energy Consumption: Some Results from a Randomized Controlled Experiment", Proceedings of the IEPEC Conference.
Titus Elizabeth and Julie Michals. 2008. "Debating Net versus Gross Impacts in the Northeast: Policy and Programs Perspectives", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Torok, and Bradley. 2009. "Nonresidential Audit Impacts: Digging Deeper Toward Causality", Proceedings of the AESP Conference.
Train, Kenneth, UC Berkeley and N/E/R/A, CA, Personal Communication with the Authors. 10/6/09.
Urge-Vorsatz, Diana, Kristina Sroukanska, and Szilard Asztalos. 2002. "Standing By in Central Europe: A Survey of Hungarian, Romanian, and Bulgarian Residences", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Wahlstrand, Garrick, Jennifer Mitchell-Jackson, Megan Campbell, and Pamela Wellner. 2009. "Measuring the Impact of Mass Media Campagins: What Do You Get for Your Research Dollars? ', Proceedings of the IEPEC Conference.
Weitzel, David, and Lisa A. Skumatz, 2004. "Efficient Techniques for Estimating Baseline and Market Shares Projections from Market Transformation Interventions", Proceedings of the 2004 ACEEE Summer Study, Asilomar, CA, August 2004.
Wong, Crispin, Hossein Haeri, Kerstin Rock, Steven Chamberlin, Ben Bronfman, and Edward Lovelace. 2009. "Using Experimental Design to Assess the Impacts of Education and Rate Design: The PEAK Plus Pilot Project", Proceedings of the IEPEC Conference.

6.3 Non-Energy Benefits

Arrow, Kenneth et al. 1993. Report of the NOAA Panel on Contingent Valuation. January. http://www.darrp.noaa.gov/library.
Barkett, Brent, 7/25/08. Summit Blue Consulting, Boulder, CO, personal conversation with the author.

BC Hydro, 2008. "BC Hydro Discussion Paper on Counting Participant Non-Energy Benefits in the Total Resource Cost Test", BC Hydro, Vancouver BC, April 15.
Becker, Gary. 1962. "Irrational Behavior and Economic Theory", Journal of Political Economy, 70:1-13.
Becker, Gary. 1976. "Economic Approach to Human Behavior", University of Chicago Press, Chicago, IL.
Bensch, Ingo, 3/31/09. Energy Center of Wisconsin, WI, Personal Communication with the Author.
Blasnik, Michael, 4/3/09. M Blasnik & Associates, Boston, MA, Personal Communication with the Authors.
Caplan, Arthur J., Therese C. Grijalva, and Paul M. Jakus, 2002. "Waste not or want not? A contingent ranking analyiss of curbside waste disposal options", Ecological Economics. 43(2-3) December.
Coito, Fred, 7/25/08. KEMA, Oakland, CA, Personal Communication with the Author.
Collins, Stephanie, 9/2009, Cadmus Group, Personal Communication with the Author.
Cooper, Joseph C., William Michael Hanemann, and Giovanni Signorello, 2002. One-and-One-Half Bound Dichotomous Choice Contingent Valuation. CUDARE Working Paper series. 921, University of California at Berkeley, Department of Agricultural and Resource Economics and Policy.
Degens, Phil. 4/21/09. Oregon Trust, Portland, OR, Personal Communication with the Author.
Dickerson, Chris Ann and Mike McCormick, 2005. How Will Energy Efficiency Evaluation Protocols Measure Up? International Energy Program Evaluation Conference, Brooklyn, NY, August 2005,
Dunn, Gordon. 11/17/08. Iowa Utilities board, Des Moines, IA, Personal Communication with the Author,
Fagan, Jennifer, 11/17/08 and 12/4/08. Itron, Madison, WI, Personal Communication with the Author.
Fisk, William J. 2000. "Health and Productivity Gains from Better Indoor Environments and Their Relationship with Building Energy Efficiency", Annu. Rev. Energy Environ. 2000, 25:537-566.
Fuchs, Leah, and Lisa Skumatz. 2004. "Non-Energy Benefits (NEBs) from Energy Star: Comprehensive Analysis of Appliance, Outreach and Homes Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Gandhi, Nikhil, Floyd Keneipp, Dulane Moran, Jane Peters, Shahanna Samiullah, and Anne West. 2007. "Product Selection- A Forgotten Vital Component of Program Design", Proceedings of the IEPEC Conference.
Geller, Howard, Stephen Bernow, and William Dougherty. 2000. "Meeting America's Kyoto Protocol Target: Policies and Impacts", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Gordon, Fred, 7/25/08. Oregon Trust, Portland, OR, Personal Communication with the Author.
Graves, Phillip, 2003. "The Simple Analytics of the WTA-WTP Disparity for Public Goods", Center for Environmental and Resource Economics Working Paper. Www2.ncsu.edu/unity/lockers/user/v/vksmith/opportunities/Graves_paper.pdf
Gunn, Randy, 7/25/08. Summit Blue Consulting, Chicago, IL, Personal Communication with the Author.
Hall, Nick, and Carmen Best. 2006. "Framework for NEBs in the Next generation of Evaluation and Program Design", Proceedings of the AESP Conference.
Harris, Jeffrey. 1996. (Northwest Power Planning Council), Personal communications with author, Lisa Skumatz, 1996.

Heschong, Lisa, Dr. Roger Wright, and Stacia Okura. 2000. "Daylighting and Productivity: Elementary School Studies", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Hill, David G., Tom Buckley, Mark Eldridge, Debra Sachs, and Abby Young. 2000. "Implementing and Monitoring Community Based Climate Action Plans", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Hill, David G., John Plunkett, Lawrence J. Pakenas, R. Neal Elliot, Christine Donovan, Phil Mosenthal, and Chris Neme. 2004. "Cost Effective Contributions to New York's Greenhouse Gas Reduction Targets from Energy Efficiency and Renewable Energy Resources", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Imbierowicz, Karen, and Lisa A. Skumatz. 2004. "The Most Volatile Non-Energy Benefits (NEBs): New Research Results "Homing In" on Environmental And Economic Impacts", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Imbierowicz, Karen, Lisa A. Skumatz, and John Gardner. 2006. "Net NEB Multipliers for Economic Impacts: Detailed Analysis of Differences by Program Type and State", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Jennings, John, and Lisa A. Skumatz. 2006. "Non-Energy Benefits (NEBs) from Commissions in Schools, Prisons, and Other Public Buildings", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Josephson, Alec, Stephen Grover, Ben Bronfman, and Fred Gordon. 2004. "Reaping the Wind: The Economic Impacts of Spending on Renewable Energy and Energy Efficiency Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Kempton, Prof. Willett. 2007. "Conservation and Renewable Energy Policy", June.
Khawaja, M. Sami, 10/3/08, 11/23/09. Cadmus Group, Portland, OR Personal Communication with the Author.
Knight, Robert. 2006. "Home Performance Retrofit Contracting and Non-Energy Benefits", For the California Building Performance Contractors Association (CBPCA).
Knight, Robert, Fran Curl, Subid Wagley, and Ganesh Venkat. 2008. "Home Performance with Energy Star in CA: Moving into the Spotlight", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Knight, Robert, Loren Lutzenhiser, and Susan Lutzenhiser. 2006. "Why Comprehensive Residential Energy Efficiency Retrofits are Undervalued", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Li, Feldman, Lynn Hoefgen, and Thomas Ledyard. 2004. "Beyond Clean: Customer Views of NEBs of Clothes Washers.", Proceedings of the AESP Conference.
Low, Jon, Blaine Collison, and Don Anderson. 2004. "Intangibles and Corporate Value: How Can Energy Efficiency Differentiate Corporate Performance?", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Lutzenheiser, Loren. 10/3/08. Portland State University, Portland, OR, Personal Communication with the Author.
Mallory, Jillian, 10/31/08, 11/6/08. BC Hydro, Vancouver BC, Canada, Personal Communication with the Author.
Markowitz, Ezra M and Bob Doppelt. 2009. "Reducing Greenhouse Gas Emissions Through Behavioral Change: An Assessment of Past Research On Energy Use, Transportation and Water Consumption", January.
McHugh,Jonathon, Lisa Heschong, Nehemiah Stone, Abby Vogen, Daryl Mills, and Cosimina Panetti. 2002. "Non-Energy Benefits As a Market Transformation Driver", Building Codes Assistance Project. Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Megdal, Lori, 11/6/08. Megdal & Associates, Acton, MA, Personal Communication with the Author.
Messenger, Michael. 7/29/08 and 4/3/09. Itron, Sacramento, CA, Personal Communication with the Author.
Mills, E., and A. Rosenfeld. 1994. Consumer Non-Energy Benefits as a Motivation for Making Energy-Efficiency Improvements, LBL Report 35405, Lawrence Berkeley Laboratory, Berkeley, CA, 1994.
Mulholland, Denise, John A. "Skip" Laitner, and Nikolaas Dietsch. 2004. "Exploring the Economic Development Implications of Capacity Building within State and Local Energy Efficiency Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Murtishaw, Scott, Lee Schipper, and Fridtjof Unander. 2000. "The "Mine/Yours" Method of International Comparisons of Carbon Emissions", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Nemtzow, David and Omar Siddiqui. 2008. "Giving Credit Where Credit is Due: EE in CO2 Emission Trading", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Nevius, Monica, Maureen McNamara, Jocelyn Spielman, and Ryan Barry. 2009. "Progress Towards Loyalty: Trends in ENERGY STAR Awareness and Brand Equity Among U.S. Households, 2000-2008", Proceedings of the IEPEC Conference.
Newberger, Jeremy, Nick Hall, Johna Roth, Paul Horowitz, David Weber. 2007. "Custom NEBs: Are They Worth It? Experiences, Challenges, and Directions in MA", Proceedings of the IEPEC Conference.
No Author Cited. (no date). "Non-Energy Benefits of Energy Saving and Energy Efficient Renovation of Houses: An introduction to the concept of Non-Energy Benefits (NEB) to the Danish debate, concerning obstacles to energy savings and possible ways to overcome them", From Internet search.
NYSERDA. 2005. "New York Energy $mart(sm) Program Evaluation and Status Report, Final Report", Albany, NY, May 2005.
O'Drain, Mary, Nick Hall, and Lisa Skumatz. 2001. "Valuing Hardship: Developing a New Cost Effectiveness Test for Low Income Energy Efficient Programs", Proceedings of the IEPEC Conference.
Ottinger, et.al., 1990, Environmental Costs of Electricity, PACE University Center for Environmental Legal Studies, for New York State Environmental Research And Development Authority and US Department of Energy, Oceana Publications, Inc., 1990.
Pearson, Dennis and Lisa Skumatz. 2002. "Non-Energy Benefits including Productivity, Liability, Tenant Satisfaction, and Others: What Participant Surveys Tell Us About Designing and Marketing Commercial Programs", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Pigg, Scott, et.al., 1994, An Evaluation of Iowa's Low Income Weatherization Program SLICE, WECC, MidIowa Community Action League, August 9, 1994.
Pomerantz, Melvin, Hashem Akbari, and John T. Harvey. 2000. "Cooler Reflective Pavements Give Benefits Beyond Energy Savings: Durability and Illumination", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Price, Lynn, Chris Marnay, Jayant Sathaye, Scott Murtishaw, Diane Fisher, Amol Phadke, and Guido Franco. 2002. "The California Climate Action Registry: Development of Methodologies for Calculating Greenhouse Gas Emissions from Electricity Generation", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Raynolds, Ned. 2004. "Out of the Closet: Climate Change as a Driver for Energy Efficiency", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Rogers, Edmunds, and Knight. 2006. "Home Performance with Energy Star Delivering Savings with a Whole House Approach", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Sabo, Carol. 11/21/09. PA Government Services, VI, Personal Communication with the Author.
Sanstad, Alan H., and John A. "Skip" Laitner. 2004. "A Multi-Agent, Multi-Attribute Policy Model for Analyzing the Adoption of Energy Efficiency Technologies", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Schauer, Laura, and Lark Lee. 2005. "Evaluating Wisconsin's Low Income Programs-final Results of the Longitudinal Study and Resulting Changes", Proceedings of the IEPEC Conference.
Schiller, Steven R, Edward Vine, and William Prindle. 2005. "Evaluating the Emission reductions for EE and Renewable Energy Projects and Programs", Proceedings of the IEPEC Conference.
Schweitzer, Martin, and Bruce Tonn. 2002. "Non-Energy Benefits from the Weatherization Assistance Program: A Summary of Findings From the Recent Literature", Prepared for U. S. Department of Energy Office of Building Technology Assistance, April.
Shepler, Nicole, 2001. "Developing a hedonic regression model", http://www.bls.gov/cpi/cipcamco.htm, accessed July 2006.
Skumatz, Lisa A., Ph.D., 1997, "Recognizing All Program Benefits: Estimating the Non-Energy Benefits of PG&E's Venture Partner's Pilot Program (VPP)", Proceedings of the 1997 Energy Evaluation Conference, Chicago, Illinois, 1997.
Skumatz, Lisa A., Ph.D., 1998, "Non-Energy Benefits (NEBs) Swamp Load Impacts - Results for Multiple Residential Programs", Skumatz Economic Research Associates, Inc. (SERA) Research Report NEB9802, April 1998.
Skumatz, Lisa. 2001. "The New "Standard" in Comprehensive Estimation and Modeling of NEBs for Commercial & Residential Programs", Proceedings of the IEPEC Conference Salt Lake City, Utah.
Skumatz, Lisa A., Ph.D., 2001, Non-Energy Benefits for Northeast Utilities, Draft Report, prepared for Northeast Utilities, 2001.
Skumatz, Lisa. 2002. "Comparing Participant Valuation Results using Three Advanced Survey Measurement Techniques: New Non-Energy Benefits Computations of Participant Value", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Skumatz, Lisa A. 2003. "The 'Mother' of Non-Energy Benefits (NEBs) Studies -Comprehensive Analysis and Modeling of NEBs for Resource Acquisition and Market Transformation Programs", Proceedings of the EEDAL conference.
Skumatz, Lisa. 2007. "Attributable Effects from Information and Outreach Programs: Net to Gross, NEBs and Beyond", Proceedings of the ECEEE Conference.
Skumatz, Lisa. 2007. "Commissioning in Public Sector Building- NEBs, Not Savings, are the Selling Point", Proceedings of the ECEEE Conference.
Skumatz, Lisa. 2007. "Economic Impacts from Energy Efficiency Programs- Variations in Multiplier Effect by Program Type and Region", Proceedings of the ECEEE Conference.
Skumatz, Lisa. 2007. "Measuring NEBs; Valuation Approaches for Participant NEBs", Proceedings of the ECEEE Conference.
Skumatz, Lisa. 2007. "Zero and Low Energy Homes in New Zealand: The Value of NEBs and their use in Attracting Homeowners", Proceedings of the ECEEE Conference.
Skumatz, Lisa A., Ph.D. and Chris Ann Dickerson, 1998, "Extra! Extra! Non-Energy Benefits of Residential Programs Swamp Load Impacts!", Proceedings of the 1998 ACEEE Conference, Asilomar, California, August 1998.
Skumatz, Lisa A., Chris Ann Dickerson, and Brian Coates, 2000, "Non-Energy Benefits In The Residential And Non-Residential Sectors - Innovative Measurements And Results For

Participant Benefits", Proceedings of the 2000 ACEEE Summer Study, Asilomar, CA, 2000.
Skumatz, Lisa A., Ph.D. and John Gardner, 2006. "Non-Energy Benefits Valuation Mechanisms: Survey and Results", Presented at Western Economics Association International, San Diego, CA.
Skumatz, Lisa, and John Gardner. 2006. "Differences in the Valuation of NEBs According to Measurement Methodology: Causes and Consequences", Proceedings of the AESP Conference. Clearwater Beach, FL.
Skumatz, Lisa A., Ph.D., and M. Sami Khawaja, Ph.D. 2009. "Non-Energy Benefits: Status, Findings, Next Steps, and Implications for Low Income Program Analyses in California", Prepared for Sempra Utilities, Draft, San Diego, CA, November 17, 2009.
Smith-McClain, Lisa, Lisa Skumatz, and John Gardner. 2006. "Attributing NEB Values to Specific Measures: Decomposition Results from Programs with Multiple Measures", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Stoecklein, Albrecht, and Lisa A. Skumatz. 2004. "Using Non-Energy Benefits (NEBs) to Market Zero and Low Energy Homes in New Zealand", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Stolarski, Richard, Smith, Kyle McDonald, and Contreras. 2008. "Total Energy and Emissions Perspectives for Utility EE Initiatives", Proceedings of the AESP Conference. Dallas, TX.
Sumi, David, 7/25/08. PA Consulting, Madison, WI Personal Communication with the Author.
Sumi, David, 2009. Quantifying and Valuing Displaced Power Plant Emissions as a Greenhouse Gas Mitigation Option, From Conference Proceedings IEPCE, Portland, OR.
Sumi, David, Oscar Bloch, and Jeff Erickson. 2005. "How to Balance Green House Gas Mitigation Strategies Across Programs with Near-term and Long-term Impacts for Public Benefits Programs", Proceedings of the IEPEC Conference.
Sumi, David, Jeff Erickson, and Jim Mapp. 2002. "Wisconsin's Public Benefits Approach to Quantifying Environmental Benefits: Creating Different Emissions Factors for Peak/Off-Peak Energy Savings", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Sumi, David, Meyers, Marnay, Fisher, and Jeff Erickson. 2001. "Quantification of Environmental Benefits for WI's Focus on E Pilot Programs", Proceedings of the IEPEC Conference.
Sumi, David, Bryan Ward, and Nick Hall. 2007. "Building Bridges Between EE Program Evaluation and GHG Mitigation Quantification Protocols", Proceedings of the IEPEC Conference.
Sumi, David and Bryan Ward. 2008. "Selecting an Appropriate Approach for Calculating Displaced Emissions for Different EE Projects and Program Types", Proceedings of the AESP Conference.
Sumi, David, Glen Weisbrod, Bryan Ward, and Miriam L. Goldberg. 2003. "An Approach to Quantifying Economic and Environmental Benefits for Wisconsin's Focus on Energy ", Presented at The International Energy Program Evaluation Conference, Seattle, WA, August.
TecMarket Works, Skumatz Economic Research Associates, and Megdal and Associates. 2001. "Low Income Public Purpose Test (LIPPT) Report", Prepared for RRM Working Group Cost Effectiveness Committee, San Francisco, CA.
Tolkin, Betty, William Blake, Elizabeth Titus, Ralph Prahl, Dorothy Conant, and Lynn Hoefgen. 2009. "What Else Does an ENERGY STAR Home Provide? Quantifying Non-Energy Impacts in Residential New Construction", Proceedings of the IEPEC Conference.
Vine, E., and J. Sathaye. 1997. "The Monitoring, Evaluation, Reporting, and Verification of Climate Change Mitigation Projects", Discussion of Issues and Methodologies and

Review of Existing Protocols and Guidelines, LBNL-40316, Lawrence Berkeley National Laboratory, Berkeley, CA.
Vine, Edward, Gregory Kats, Jayant Sathaye, and Hemant Joshi. 2003. "International Greenhouse Gas Trading Programs: A Discussion of Measurement and Accounting Issues", Energy Policy 31 (2003) 211-224.
Vine, E., and J. Sathaye. 1999. "Guidelines for the Monitoring, Evaluation, Reporting, Verification, and Certification of Energy-Efficiency Projects for Climate Change Mitigation", LBNL-41543, Lawrence Berkeley National Laboratory, Berkeley, CA.
Vine, E., and J. Sathaye. 2000. "The Monitoring, Evaluation, Reporting, Verification, and Certification of Energy-Efficiency Projects", Mitigation and Adaptation Strategies for Global Change 5, 189-216.
Wietzel, David and Lisa Skumatz. 2001. Measure Retention Study: Revised Lifetimes for a Residential Weatherization Program. Proceeding from the ACEEE Conference, Asilomar, CA.
Wobus, Nicole, Jennifer Meissner, Barkett, Waldman, Train, Thacher, Daniel Violette. 2007. "Exploring the Application of Conjoint Analysis for Estimating the Value of Non-Energy Impacts", Proceedings of the IEPEC Conference. Chicago, IL.
Woods, Rose A., and Lisa A. Skumatz. 2004. "Self-Efficacy in Conservation: Relationships between Conservation Behavior and Beliefs in the Ability to Make a Difference", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Woolf, Tim, 1999. "Environmental Benefits of Efficiency Programs", memorandum to DTE-100 Cost-effectiveness working group on behalf of Cape Light Compact, Cambridge, Massachusetts, March 31.

6.4 Persistence / Lifetimes / EULs

Albert, Scott, 4/9/09. GDS Associates, Marietta, GA, Personal Communication with the Author.
Barkett, Brent, 7/25/08. Summit Blue Consulting, Boulder, CO, Personal Communication with the Author.
Bobker, Michael. 2005. "Existing Building Commissioning: Market Transformation for Persistence of Savings; Recognizing and Formalizing the Role of Operator Training", Proceedings of the IEPEC Conference.
Bond, Murray. 5/5/08. BC Hydro, Vancouver, BC. Personal Interview.
Bourassa, Norman J., Mary Ann Piette, and Naoya Motegi. 2004. "An Evaluation of Savings and Measure Persistence from Retrocommissioning Of Large Commercial Buildings", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.
Campoy, Leonel. 5/2/08. Southern California Edison, Rosemead, CA. Personal Interview.
Canseco, Jennifer, Tami Rasmussen, and Anu Teja. 2009. "A Market Transformed: But Will the Impacts Be Sustained?", Proceedings of the IEPEC Conference.
Coito, Fred, 7/25/08. KEMA, Oakland, CA, Personal Communication with the Author.
Eijadi, David A. 2005. "Performance Persistence What happens to predicted energy savings from Design Assistance Programs after several years of building operation?", Proceedings of the IEPEC Conference.
GDS Associates, 2007. "Measure Life Report - Residential and Commercial / Industrial Lighting and HVAC Measures", Prepared fro the New England State Program Working Group (SPWG), Manchester NH.
Haasl, Tudi, Hannah Friedman, and Amanda Potter. 2004. "Strategies for Improving Persistence of Commissioning Benefits: Making Lasting Improvements in Building Operations", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Harrigan, Merrilee, and Judith M. Gregory, 1994, "Documenting Energy Savings Enhancements from Energy Education Components of a Low Income Weatherization Program", Proceedings of the 1994 ACEEE Summer Study, Asilomar, CA, 1994.

Jump, Corina, James Hirsch, Jane Peters, and Dulane Moran. 2008. "Welcome to the Dark Side: The Effect of Switching on CFL Measure Life", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Keating, Ken, 5/2/08. Bonneville Power Administration, Portland, OR. Personal communication with the Author.

Messenger, Michael. 7/29/08, 4/3/09. Itron, Sacramento, CA. Personal Communication with the Author.

Richardson, Valerie, and Lisa Skumatz. 2000. "Measure Retention in Residential New Construction", Proceedings of the ACEEE Summer Study on Building Conference, Asilomar, CA.

Skumatz, Lisa A. Ph.D;, and John Gardner, 2005. "Revised / Updated EULs Based on Retention and Persistence Studies Results", Prepared for Southern California Edison.

Skumatz, Ph.D., Lisa, and John Gardner. 2005. "Best Practices in Measure Retention and Lifetime Studies: Standards for Reliable Measure Retention Methodology Derived from Extensive Review", Proceedings of the IEPEC Conference.

Skumatz, Lisa A., Ph.D. and Allen Lee, 2004. "Attachment G - Assessment of Technical Degradation Factor (TDF) Study", prepared as Attachment to "Review of Retention and Persistence Studies for California Public Utilities Commission (CPUC)", San Francisco, CA.

Skumatz, Lisa A. Ph.D., Rose Woods, and Scott Dimetrosky, 2004. "Review of Retention and Persistence Studies for the California Public Utilities Commission (CPUC), San Francisco, CA.

Tellerico, Tom. 5/2/08. Glacier Consulting Madison, WI. Personal Communication with the Author.

APPENDIX A: SUMMARY OF KEY ELEMENTS OF CALIFORNIA PROTOCOLS

1. California Protocols – Key Notes, Volume II (Research Methodologies)

- The overall Impact Evaluation Protocol contains one subset of 3 Protocols for estimating direct energy and demand impacts and one for estimating indirect impacts.

    - Direct Impact Evaluation Protocols:

        a. The Gross Energy Impact Protocol has two levels of rigor (Basic and Enhanced) for developing gross energy estimates;

        b. The Gross Demand Impact Protocol has two levels of rigor (Basic and Enhanced) for developing gross demand estimates; and

        c. The Participant Net Impact Protocol has three levels of rigor for developing net impact estimates (Basic, Standard and Enhanced).

    - The Indirect Impact Evaluation Protocol has three levels of rigor (Basic, Standard and Enhanced). The Basic Rigor level is reserved for those programs or program components that cannot be linked to energy savings but where net behavior changes need to be estimated to measure program impacts.
    - Other Protocols:
        a. The Measurement and Verification (M&V) Protocol
        b. The Emerging Technology Protocol
        c. The Codes and Standards Protocol
        d. The Effective Useful Life Protocol
        e. The Process Evaluation Protocol
        f. The Market Effects Protocol
        g. The Sampling and Uncertainty Protocol

2. Minimum Allowable Methods for Gross Energy Evaluation

Basic Rigor

- The primary difference between the Basic and Enhanced rigor levels is that the minimum allowable methods in the Enhanced rigor level directly address or control for the more likely sources of potential bias in that class of methods (e.g., regression-based versus engineering-based).
- Simple Engineering Model (SEM) with M&V equal to IPMVP Option A and meeting all requirements in the M&V Protocol for this method. Sampling according to the Sampling and Uncertainty Protocol.
- Normalized Annual Consumption (NAC) using pre- and post-program participation consumption from utility bills from the appropriate meters related to the measures undertaken, normalized for weather, using identified weather data to normalize for

heating and/or cooling as is appropriate to measures included. Twelve (12) months pre-retrofit and twelve (12) months post-retrofit consumption data is required. Sampling must be according to the Sampling and Uncertainty Protocol.

Enhanced Rigor

- A fully specified regression analysis of consumption information from utility bills with inclusion/adjustment for changes and background variables over the time period of analysis that could potentially be correlated with the gross energy savings being measured. Twelve (12) months post-retrofit consumption data are required. Twelve (12) months pre-retrofit consumption data are required, unless program design does not allow pre-retrofit billing data, such as in new construction. In these cases, well-matched control groups and post-retrofit consumption analysis is allowable.[164] Sampling must be according to the Sampling and Uncertainty Protocol utilizing power analysis as an input to determining required sample size(s).
- Building energy simulation models that are calibrated as described in IPMVP Option D requirements in the M&V Protocols. If appropriate, may alternatively use a process-engineering model (e.g., AirMaster+) with calibration as described in the M&V Protocols. Sampling according to the Sampling and Uncertainty Protocol.
- Retrofit Isolation engineering models as described in IPMVP Option B requirements in the M&V Protocols. Sampling according to the Sampling and Uncertainty Protocol.
- Experimental design established within the program implementation process, designed to obtain reliable net energy savings based upon differences between energy consumption between treatment and non-treatment groups from consumption data.[165] Sampling must be according to the Sampling and Uncertainty Protocol.
- All impact evaluations should employ a research design that has properly identified participants made available from the program database(s). The regression methods of pre- and post-consumption and the calibrated engineering model equivalent to Option D could yield results not restricted to the program being evaluated if participation in multiple programs occurs around the same time period or overlaps in influence. This could contribute to double counting at the portfolio level. Evaluators are required to ensure that their methodologies and analysis account for any overlap in program participation and measures that could potentially bias the program evaluation results.

[164] Post-retrofit only billing collapses the analysis from cross-sectional time-series to cross-sectional. Given this, even more care and examination is expected with regard to controlling for cross-sectional issues that could potentially bias the savings estimate.
[165] The overall goal of the Direct Impact Protocols is to obtain reliable net energy and demand savings estimates. If the methodology directly estimates net savings at the same or better rigor than the required level of rigor, then a gross savings and participant net impact analysis is not required to be shown separately.

- All impact evaluations must meet the requirements of the Sampling and Uncertainty Protocol. Regression analysis of consumption data requires addressing outliers, missing data, weather adjustment, selection bias, background variables, data screens, heterogeneity of customers, autocorrelation, truncation, error in measuring variables, model specification and omitted variable error, heteroscedasticity, collinearity and influential data points.
- Engineering analysis and M&V-based methods are required to address sources of uncertainty in parameters, construction of baseline, guarding against measurement error, site selection and non-response bias, engineering model bias, modeler bias, deemed parameter bias, meter bias, sensor placement bias and non-random selection of equipment or circuits to monitor.
- Experience in energy efficiency program evaluation has shown that there are cases where some methods are more likely to yield defensible results than others for certain sectors or program designs. Experience to date in energy efficiency impact program evaluation has generally shown the following:

    a. NAC methods are most applicable to residential and small commercial efforts where the expected energy savings are at least 10 percent of pre-installation usage;

    b. NAC methods are not well suited to handle significant issues with heteroscedasticity, truncation, self-selection or changes in background issues (e.g., significant change in economic conditions-large recession, recovery or economic growth);

    c. SEM methods are not well suited for whole building measures with interactive effects or commissioning/retro-commissioning efforts;

    d. The heterogeneity and multitude of background variable issues for industrial customers and unique commercial (e.g., ski resorts and amusement parks/facilities) or institutional (e.g., water/wastewater and prisons) customers make the use of any regression-based consumption analysis difficult and potentially less reliable than engineering-based methods;

    e. Regression-based consumption analyses are less likely to be able to obtain definitive energy savings estimates where the expected energy savings are not at least 10 percent of pre-installation usage; and

    f. Regression-based consumption analysis is quite difficult for new construction programs due to the lack of pre-retrofit consumption data and the consequential greater burden for controlling for cross-sectional issues for comparing participants and non-participants (and self-selection bias, particularly if the non-participants are any form of rejecters of program participation). New construction program impact evaluations are generally conducted using engineering models (such as those described in IPMVP Option D).

3. Minimum Allowable Methods for Gross Demand Evaluation

Basic Rigor

- Reliance upon secondary data for estimating demand impacts as a function of energy savings. End-use savings load shapes or end-use load shapes from one of the following will be used to estimate demand impacts:

    a. End-use savings load shapes, end-use load shapes or allocation factors from simulations conducted for DEER, or
    b. Allocation factors from CEC forecasting models or utility forecasting models with approval through the evaluation plan review process, or
    c. Allocation based on end-use savings load shapes or end-use load shapes from other studies for related programs/similar markets with approval through the evaluation plan review process

Enhanced Rigor

- Primary demand impact data must be collected during the peak hour during the peak month for each utility system peak. Estimation of demand impact estimates based on these data is required. If the methodology and data used can readily provide 8,760-hour output, these should also be provided.[166] Sampling requirements can be met at the program level but reporting must be by climate zone (according to CEC's climate zone classification).

    a. If interval or time-of-use consumption data are available for participants through utility bills, these data can be used for regression analysis, accounting for weather, day type and other pertinent change variables, to determine demand impact estimates. Pre- and post-retrofit billing periods must contain peak periods. Requires using power analysis, evaluations of similar programs, and professional judgment to determine sample size requirements for planning the evaluation. Needs to meet the requirements of the Sampling and Uncertainty Protocol.

[166] This includes the use of 15-minute interval data or Building Energy Simulation models whose output is 8,760 hourly data.

- Spot or continuous metering/measurement of peak pre and post-retrofit during the peak hour of the peak month for the utility system peak to be used with full measurement Option B or calibrated engineering model Option D meeting all requirements as provided in the M&V Protocol. Pre-retrofit data must be adjusted for weather and other pertinent change variables. Must meet the Sampling and Uncertainty Protocol with a program target of 10% precision at a 90% confidence level. Experimental design established within the program implementation process, designed to obtain reliable net demand savings based upon differences between energy consumption during peak demand periods between treatment and non-treatment groups from consumption data or spot or continuous metering.[167] Sampling must be according to the Sampling and Uncertainty Protocol.
- For the purposes of the Gross Demand Impact Protocol, demand impacts must be reported as energy savings estimates for six time periods for each of four months as follows: noon-1 p.m., 1-2 p.m., 2-3 p.m., 3-4 p.m., 4-5 p.m. and 5-6 p.m. for June, July, August and September for each climate zone in which there are program participants. These demand savings are to be estimated using the Typical Meteorological Year from the National Oceanic & Atmospheric Administration (NOAA), the CEC CTZ long-term average weather data, the Administrator's long-term average weather year or the CEC's rolling average weather year.
- A regression model specified to measure program impacts for peak time periods (via analysis of interval data) or TOU/demand[168] consumption metering can be used to estimate program gross demand. This regression analysis must properly account for weather influences that are specific to the demand estimation and other pertinent change variables (e.g., day-type and hours of occupancy).
- Regression analysis with interval data should focus on obtaining direct demand impacts. If demand consumption data are used, a methodology to estimate demand savings based upon the demand regression analysis must be detailed in the evaluation plan and approved through the evaluation planning review process. Pre- and post-retrofit billing periods must contain peak periods within this analysis. A power analysis in combination with evaluations of similar program and professional judgment must be used to select and justify the proposed sample sizes.[169]
- The second class of primary data collection for the Enhanced Gross Demand Impact Protocol is to conduct field measurement of peak impacts within the evaluation effort. Spot or continuous metering/measurement at peak pre- and post-retrofit will be conducted during the peak hour in the peak month for the utility system peak. These data will be used with one of two engineering modeling approaches: (1) full measurement Option B or (2) calibrated engineering model Option D, where the modeling approach must meet all requirements as provided in the M&V Protocol.
- Both of these engineering methods need to be designed to a program target of 10 percent precision at a 90 percent confidence level and must meet the requirements of the Sampling and Uncertainty Protocol.

---

[167] The overall goal of the Impact Protocols is to obtain reliable net energy and demand savings estimates. If the methodology directly estimates net savings at the same or better rigor than the required level of rigor, then a gross savings and participant net impact analysis is not required to be shown separately.

[168] If demand billing is used, the research design must address the issues of building demand versus time period for peak and issues with demand ratchets and how the evaluation can reliably provide demand savings estimates.

[169] Power analysis is a statistical technique that can be used (among other things) to determine sample size requirements to ensure statistical significance can be found. There are several software packages and calculation Web sites that conduct the power analysis calculation. Power analysis is only being required in the Protocol for determining required sample sizes. .

- The third class of allowable methods is experimental design with primary data collection. Experimental design with demand measurement comparisons between customers randomly assigned to the treatment and non-treatment groups meets the Enhanced Gross Demand Protocol rigor level. Experimental design will need to measure energy savings during peak periods either through interval data or spot or continuous monitoring of comparison treatment and non-treatment groups to calculate demand savings estimates. Currently, experimental design has not been widely used within efficiency evaluation.

4. Participant Net Impact Protocol

- The intent is to provide reliable estimates of program level net energy and demand impacts when combined with the results from work complying with the Gross Energy Impact Protocol and the Gross Demand Impact Protocol.

Basic Rigor

- Participant self-report.

Standard Rigor

- Participant and non-participant analysis of utility consumption data that addresses the issue of self-selection.
- Enhanced self-report method using other data sources relevant to the decision to install/adopt. These could include, for example, record/business policy and paper review, examination of other similar decisions, interviews with multiple actors at end-user, interviews with mid-stream and upstream market actors, Title 24 review of typically built buildings by builders and/or stocking practices.
- Econometric or discrete choice[170] with participant and non-participant comparison addressing the issue of self-selection.

Enhanced Rigor

- "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 estimates obtained.

- Participant net impact analysis must address the following issues:

    a. Probability that the participant would have adopted the technology or behavior in the absence of the program (participant free-ridership);
    b. If adopted in the absence of the program, the probability or proportion (partial free-ridership) of expected savings induced by the program given its ability to:
        i. Increase the efficiency of what would have been adopted;
        ii. Make the adoption occur earlier than when it would have occurred; and
        iii. Increase the quantity of efficient equipment that would have been adopted.
    c. The estimation of participant net is consistent with decision-making behavior;

[170] The instrumental-decomposition (ID) method described and referenced in the *Evaluation Framework* (page 145) is an allowable method that falls into this category. A propensity score methodology is also an allowable method in this category as described in: Itzhak Yanovitzky, Elaine Zanutto and Robert Hornik,"Estimating causal effects of public health education campaigns using propensity score methodology." *Evaluation and Program Planning* 28 (2005): 209–220.

d. Consistency is assessed to ensure that other forms of bias, such as, centrality bias, are not introduced;
e. If survey methods are used, ensuring that survey questions (instrumentation) and techniques are employed to minimize social desirability bias;
f. Results that include only free-ridership adjustment are clearly labeled as such;
g. Report participant free-ridership and participant spillover separately where the methodologies selected allow this to be done;
h. If at least some portion of participant spillover may be embedded within the gross savings estimates cannot be separated out using the estimation method chosen (e.g., a regression approach is used and the spillover behavior is simultaneous with program participation), clearly present why participant spillover may be present within these estimates and a qualitative assessment of whether these might be expected to be significant or not compared to the program savings estimate;
i. And if only participant free-ridership is presented in the report without a reporting of participant spillover savings, clearly discuss that this presents a downwardly biased presentation of overall true net savings.

- The research design, selected method, survey instrument design or modeling specification(s) must also address participant self-selection bias(es). Overall sample size targets can be by program. However, all survey or interview inquiries concerning participant net (free-ridership and spillover, and application to gross impacts to obtain net savings) need to be conducted and measured by measure or end-use. Considerations of uncertainty should guide the sample stratification plan.
- Like the other approaches to estimating the NTGR, there is no precision target when using the self-report method. However, unlike the estimation of the required sample sizes when using the regression and discrete choice approaches, the self-report approach poses a unique set of challenges to estimating required sample sizes. These challenges stem from the fact that the self-report methods for estimating free-ridership involve greater issues with construct validity and often include a variety of layered measurements involving the collection of both qualitative and quantitative data from various actors involved in the decision to install the efficient equipment. Such a situation makes it difficult to arrive at a prior estimate of the expected variance needed to estimate the sample size.
- This Protocol, instead, establishes a minimum sample size for end-use participants: a sample of 300 participant decision-makers for at least 300 participant sites (where decision-makers may cover more than one site) or a census attempt, whichever is smaller. Sample sizes of other actors, engineering work or record review need to be described in the evaluation plan and approved through the evaluation planning review process.

5. Minimum Allowable Methods for Indirect Impact Evaluation

- The primary uncertainty within the logic chain of obtaining energy and demand savings from these types of programs is the estimation of the program-induced impact on the behavior of participants. Therefore, the primary focus of the Indirect Impact Evaluation is in evaluating and estimating the program's net impact on behavioral change.

Basic Rigor

- An evaluation to estimate the program's net changes on the behavior of the participants is required; the impact of the program on participant behavior.

Standard Rigor

- A two-stage analysis is required that will produce energy and demand savings. The first stage is to conduct an evaluation to estimate the program's net changes on the behavior of the participants/targeted-customers. The second is to link the behaviors identified to estimates of energy and demand savings based upon prior studies (as approved through the evaluation planning or evaluation review process).

Enhanced Rigor

- A three-stage analysis is required that will produce energy and demand savings. The first stage is to conduct an evaluation to estimate the program's net impact on the behavior changes of the participants. The second stage is to link the behavioral changes to estimates of energy and demand savings based upon prior studies (as approved through the evaluation planning or evaluation review process). The third stage is to conduct field observation/testing to *verify* that the occurrence of the level of net behavioral changes.

- Indirect impact evaluation design, analysis and reporting must address the following issues:

  a. Expected impacts and the target audience for these impacts;

  b. How the expected impacts will be measured;

  c. Identification and measurement of baseline (and where baseline would have been in the absence of the program, i.e., forecasted, dynamic baseline or estimated counter-factual from research design) or identification and measurement of well-matched non-treatment comparison group over time;

  d. Extent of exposure/treatment and how this is being measured in the evaluation; and

  e. Self-selection bias and how this is being controlled for to obtain an unbiased estimate of the program-induced impact.

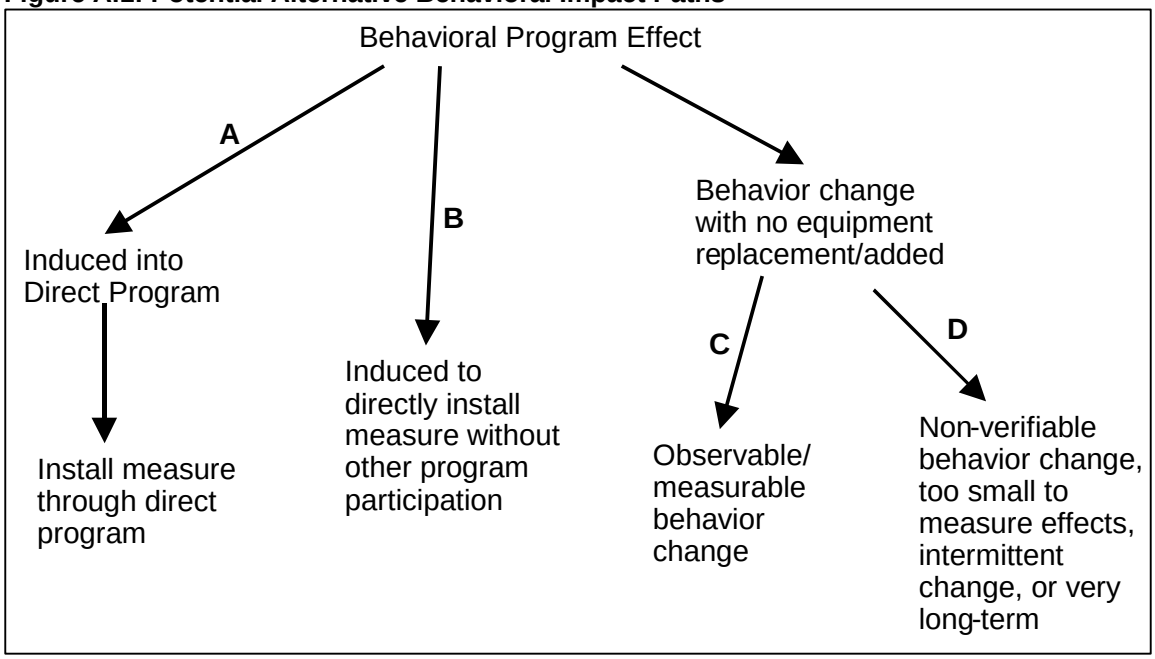
- The assessment or development of a program theory and logic model (PT/LM) is recommended. The PT/LM could be particularly useful if expanded to include the expected interactions with the market or the use of behavioral change models. These can be valuable as a foundation for the evaluation research design, researchable questions and basis for developing survey/interview questions.
- In the Standard and Enhanced rigor levels, evaluation studies are conducted to link net behavioral impacts to energy and demand saving impacts based upon prior studies. These prior studies do not need to be previously completed evaluations (however this is preferred if they are available). For example, linking net behavior change savings estimates using DEER will meet the Indirect Impact Evaluation Protocol. Linking savings estimates to past evaluations of similar programs, new engineering models for savings estimates or other studies must be approved by the Joint Staff through the evaluation review process.

- A behavioral impact program (through information, education, training, advertising or other non-monetary incentive efforts) may be part of a portfolio to lead customer/market actors into other programs. This program/program component could be assigned an Indirect Impact Evaluation to determine the impact the program(s) is having on the portfolio and to provide input for the process evaluation

of the program. An assignment of the Standard rigor level requires that an impact evaluation be conducted and linked to energy and demand savings estimates. (The energy and demand savings, however, would not, in this case, be added to the portfolio level savings unless a method is used and approved by the Joint Staff to ensure that these savings are not double counted with those attributed to other programs).

- Four types of impacts from a behavioral change program are shown in Figure A.1.
- Inducing customers into other programs is shown as Path A. Savings from this path are not direct savings due to the information, education, training or advertising program under study. The savings are those obtained through the direct program. However, documenting the impacts of this effort is important to estimate the various components that contribute to generating a portfolio's savings and to aid in making investment decisions. An example might be customers who participate and obtain high-efficiency room air conditioners through a rebate program due to behavioral impacts from the program being evaluated.

Figure A.1: Potential Alternative Behavioral Impact Paths



- Programs or program components that directly influence customer behavior to purchase high efficiency replacement equipment or add equipment that can save energy (e.g., timers) are shown as Path B. If assigned an Indirect Impact Evaluation with a Standard or Enhanced rigor level, these programs would be expected to undertake similar evaluation designs to those in Path A. The energy and demand savings for these, however, are *directly* attributable to the program effort being evaluated. The research design may need to estimate and find the proportion of customers that take these actions outside of other programs. An example might be customers who purchase high efficiency room air-conditioning due only to this program and who did not receive any financial incentives from other portfolio efforts to do so.

- Path C refers to those program-induced behavioral changes that can be observed or measured but are not tied to equipment replacement or the addition of equipment. This could include such changes as those to business policies regarding energy efficiency, architects' decisions on when to test daylighting alternatives, and/or plant managers' operating and maintenance schedules.
- Path D represents behavioral changes that are too small, long-term or intermittent to be cost-efficiently verified through observation, field-testing or surveying with enough reliability to measure any energy and demand impacts. Depending on the level of investment and the advances made in the evaluation of behavioral change, the programs or program components that fall into this category could vary over time. Path D examples include residential behavior of turning off lights, educating children through school programs to changing their energy-use behavior when they are adults, and changes in residential thermostat set points. The Joint Staff will only assign a Basic rigor level for this category if meeting a higher rigor level would not be possible. This could occur because a specific estimate of the degree of the impact cannot be obtained cost-effectively or the link and translation to energy and demand savings is not available or cost-effective to develop.
- Every program evaluation is required to demonstrate that the program is accomplishing its primary goals of affecting behavioral change, as stated in its PT/LM.
- It is expected that the Indirect Impact Evaluation for paths A, B and C will be assigned either a Standard or Enhanced rigor level depending upon the size of resources being invested and the importance of the anticipated outcomes to the overall success of the portfolio. The indirect impact evaluation for an Enhanced rigor level is distinguished from a Standard rigor level by the requirement to conduct field observation/testing to verify net changes in behavior. For Path D it is expected that only a Basic rigor level will *usually* be assigned. The evaluation design for each path is briefly described below:

    a. Path A: The evaluation design to verify these actions is most straightforward for Path A. Verification through program participation is sufficient given these programs are conducting their own verification and impact evaluation.

    b. Path B: The evaluation design for Path B requires the additional step of finding effected customers. This step would have to be part of the evaluation design when estimating the proportion affected in the impact evaluation.

    c. Path C: The evaluation research design needed to accomplish an Enhanced rigor indirect impact evaluation following Path C is more challenging. Examples of Path C activities include review of pre- and post-program architectural plans, review of government policy, planning and hearing documents and their dates of adoption along with interview support, examination of business policy manuals, and review of business programs created due to education efforts and testing subsequent employee knowledge and reported actions.

    d. Path D: For path D, the Basic level rigor indirect impact evaluation must be used to demonstrate that the program has carried out specific activities that are designed to produce behavioral change.

6. Measurement and Verification (M&V) Protocol

- M&V will typically be used to support impact studies by providing measured quantitative data from the field. One of the primary uses is to reduce uncertainty in baselines, engineering calculations, equipment performance and operational parameters. However, M&V can be used in process and market effects evaluations as well, when such data are useful for understanding issues such as measure quality and suitability for particular applications, installation practices and quality, baseline equipment efficiency and operation practices, and other issues identified by the process and/or market effects evaluation plan.
- How M&V differs from impact evaluation: M&V refers to data collection, monitoring and analysis activities associated with the calculation of gross energy and peak demand savings from individual customer sites or projects. Gross and net impacts at the program level will be guided by the Impact Evaluation Protocol, where results from M&V studies conducted on a sample of sites will be combined with other information to develop an overall estimate of savings by program or program component.
- Sources of uncertainty in engineering estimates: Engineering estimates are based on the application of the basic laws of physics to the calculation of energy consumption and energy savings resulting from the implementation of energy-efficient equipment and systems. Engineering models range from simple one-line algorithms to systems of complex engineering equations contained within a building energy simulation program such as DOE-2. Uncertainty in engineering estimates stems from uncertainty in the inputs to an engineering model and the uncertainty in the ability of the algorithms to predict savings.
- Uncertainty analysis and M&V planning: Energy efficiency programs utilize a wide range of technical and behavioral tools and concepts as “measures.” The likelihood of success of the measure depends on a large number assumptions, many of which can be verified through measurement. Measured data from field studies are used to quantify and reduce the uncertainty in energy and peak demand impact calculations.
- Uncertainty analysis conducted during the planning phase shall be used to identify the assumptions that have the greatest contribution to the overall savings uncertainty and allocate resources in an appropriate manner to address these uncertainties.
- The objectives of measure installation verification are to confirm that the measures were actually installed, the installation meets reasonable quality standards, and the measures are operating correctly and have the potential to generate the predicted savings. Installation verification shall be conducted at all sites claiming energy or peak demand impacts where M&V is conducted.
- Measure existence shall be verified through on-site inspections of facilities. Measure, make and model number data shall be collected and compared to participant program records as applicable. Sampling may be employed at large facilities with numerous measures installed. As-built construction documents may be used to verify measures such as wall insulation where access is difficult or impossible. Spot measurements may be used to supplement visual inspections, such as solar transmission measurements and low-e coating detection instruments to verify the optical properties of windows and glazing systems.

- Measure installation inspections shall note the quality of measure installation, including the level of workmanship employed by installing contractor toward the measure installation and repairs to existing infrastructure affected by measure installation, and physical appearance and attractiveness of the measure in its installed condition. Installation quality guidelines developed by program implementer shall be used to assess installation quality. (If such guidelines are not available, they shall be developed by the M&V contractor and approved by the Joint Staff prior to conducting any verification activities. Installation quality shall be determined from the perspective of the customer).
- Correct measure application and measure operation shall be observed and compared to project design intent. For example, CFL applications in seldom used areas or occupancy sensors in spaces with frequent occupancy shall be noted during measure verification activities. At enhanced rigor sites, commissioning reports (as applicable) shall be obtained and reviewed to verify proper operation of installed systems. If measures have not been commissioned, measure design intent shall be established from program records and/or construction documents; and functional performance testing shall be conducted to verify operation of systems in accordance with design intent.

Table A.1: Summary of M&V Protocol for Enhanced Level of Rigor

Provision	Requirement
Verification	Physical inspection of installation to verify correct measure installation and installation quality. Review of commissioning reports or functional performance testing to verify correct operation
IPMVP Option	Option B or Option D
Source of Stipulated Data	DEER assumptions, program work papers, engineering references, manufacturers catalog data, on-site survey data
Baseline Definition	Consistent with program baseline definition. May include federal or Title 20 appliance standards effective at date of equipment manufacture, Title 24 building standards in effect at time of building permit; existing equipment conditions or common replacement or design practices as defined by the program
Monitoring Duration	Sufficient to capture all operational modes and seasons
Weather Adjustments	Weather dependent measures: normalize to long-term average weather data as directed by the Impact Evaluation Protocol
Calibration Criteria	Option D building energy simulation models calibrated to monthly billing or interval demand data. Optional calibration to end-use metered data
Additional Provisions	Hourly building energy simulation program compliant with ASHRAE Standard 140-2001

IPMVP Option

The Enhanced rigor M&V Protocol shall conform to IPMVP Option B - Retrofit Isolation or IPMVP Option D - Calibrated Simulation. Under Option B, savings are determined by field measurement of the energy use of the systems to which the ECM was applied separate from the energy use of the rest of the facility. Savings are estimated directly from measurements. Stipulated values are not allowed. Under Option D, savings are determined through simulation of the energy use of components or the whole facility. Simulation routines should be demonstrated to adequately model actual energy performance measured in the facility. Savings are estimated from energy use simulation, calibrated with hourly or monthly utility billing data, and/or end-use metering.

7. Emerging Technologies Protocol

- The Statewide Emerging Technologies Program (ETP) is an information-only program that seeks to accelerate the introduction of innovative energy efficient technologies, applications and analytical tools that are not widely adopted in California. The overall objective of the ET Program is to verify the performance of new energy efficiency innovations which can be transferred directly into the marketplace and/or integrated into utility portfolios in support of resource acquisition goals for energy efficiency. Emerging technologies may include hardware, software, design tools, strategies and services.
- Finally, it is recognized that such programs are expected to have a number of failures[171] (technologies that do not perform as expected) given the inherent risks[172] associated with the technologies selected for investigation.
- Because of the absence of energy and demand goals and the longer lead time required to introduce new technologies directly into the market and/or into utility energy efficiency programs, a separate Protocol has been prepared to guide the ETP evaluation. The evaluation approach in this Protocol is theory-driven and is based on monitoring the full range of activities, outputs, and immediate, intermediate and long-range outcomes. This approach explicitly recognizes that while many, if not all, of these outputs and outcomes are difficult, if not impossible, to monetize, they can be documented and monitored over time to assess whether the program is on track to achieve the ultimate impacts[173].

Table A.2: Sample of Available ETP Evaluation Methods

Method	Brief Description	Example of Use
Analytical/conceptual modeling of underlying theory	Investigating underlying concepts and developing models to advance understanding of some aspect of a program, project, or phenomenon.	To describe conceptually the paths through which projects evolve or through which spillover effects may occur and validate the underlying theory.
Survey	Asking multiple parties a uniform set of questions about activities, plans, relationships, accomplishments, value, or other topics, which can be statistically analyzed.	To find out how many members of a given target audience have been informed about a given technology through the dissemination efforts of the ETP.
Case study - descriptive	Using single-case or multiple-case designs with single or multiple units of analysis for investigating in-depth a program or project, a technology, or a facility, describing and explaining how and why developments of	To recount how a particular joint venture (e.g., between the ETP and a customer who hosts a technology demonstration; between the ETP and a manufacturer) was formed, how parties shared research tasks, and why the collaboration was

[171] There are two types of failure: 1) failure of the technology to perform as expected (note: such failures can provide valuable information to members of the various target audiences), and 2) the failure of the utility to select promising technologies such that a reasonable number of new technologies are not being funneled into utility energy efficiency programs. This Protocol will address both types of failure.

[172] Risk involves the exposure to a chance of injury or loss. Hardware, software, design tools, strategies and services (products) have varying levels of uncertainty as to whether they will perform as expected. Thus, investing in these products assumes varying levels of risk that the return on these investments might not be fully realized (i.e., there will be a loss).

[173] Unlike the methods identified in the Impact Protocol, the methods for evaluating the benefits of public investment in RD&D and related emerging technology programs are not nearly as advanced. However, it has been recognized by many that stakeholders should not have to wait three to five to ten years before discovering whether projects with relatively long times are successful. There is agreement among many researchers that one should be able to identify immediate and intermediate indicators that can reassure stakeholders that the efforts are on track to achieve such objectives as successful deployment of new technologies into utility energy efficiency programs and the bridging of the "chasm", leading eventually to significant energy and demand impacts.

Method	Brief Description	Example of Use
	interest have occurred.	successful or unsuccessful.
Sociometric and social network analysis	Identifying and studying the structure of relationships by direct observation, survey, and statistical analysis of secondary databases to increase understanding of social/organizational behavior and related economic outcomes.	To learn how projects can be structured to increase the diffusion of resulting knowledge.
Bibliometrics - counts	Tracking the quantity of research outputs.	To find how many publications per applied research dollar a technology assessment generated.
Bibliometrics - citations	Assessing the frequency with which others cite publications or patents and noting who is doing the citing.	To learn the extent and pattern of dissemination of a technology assessment's publications and patents.
Bibliometrics - content analysis	Extracting content information from text using techniques such as co-word analysis, database tomography, and textual data mining, supplemented by visualization techniques.	To identify a project's contribution, and the timing of that contribution, to the evolution of a technology.
Historical tracing	Tracing forward from research to a future outcome or backward from an outcome to precursor contributing developments.	To identify apparent linkages between a ratepayer-funded applied research project and something of significance that happens later or has already occurred.
Expert judgment/Peer Review	Using informed judgments to make assessments.	Experts can be called upon to give their opinions about the technical quality and effectiveness of a technology assessment. The experts generally render their verdict after reviewing written or orally presented evidence.

Source: Adapted from Ruegg and Feller (2003)

- The aggregate analysis involves the analysis of a variety of data collected for *all* of the projects in each utility's ETP portfolio. Such a level of analysis provides a statistical overview of the ETP portfolio (e.g., frequencies, cross tabulations, means etc.) across multiple projects and participants. The analysis of these aggregate data will allow one to address a number of contextual, program and policy questions, such as:

1. What are the various sources of funding, (PGC, academic institutions, manufacturers, government agencies, etc.), by type of technology assessment?
2. How many full-time equivalent ETP employees are involved by type of technology assessment?
3. How does PGC funding and co-funding vary by type of technology assessment by sector over time?
4. How does PGC funding and co-funding vary by end use and/or by sector over time?
5. What is the frequency of the various types of technology assessments, by end use, over time?
6. How is risk being balanced (e.g., measures that do not perform as expected versus those that do)?
7. What is the average duration of a technology assessment?
8. Are the technology assessments proportionately focused on sectors and end-uses in which there are the greatest expected potential energy and demand benefits?

9. How many technology assessments are launched annually?
10. How many technology assessments are currently active?
11. What percent of the technologies sponsored by the ETP have been deployed into utility energy efficiency program and/or directly into the marketplace?
12. Are there imbalances in the types of projects funded?
13. Are the needs of all the sectors being adequately addressed?

- Those technologies that have been deployed to utility energy efficiency programs must be tracked over time to determine their adoption rates[174] and resulting energy and demand impacts. Adoption rates and energy and demand impacts are useful indicators of how well the ETP screened promising technologies and developed strategies, in close collaboration with the utility-sponsored energy efficiency programs, to cross the "chasm". The goal of this component of the Protocol is ***not*** to attribute these savings directly to ETP as a resource, but to show a clear trail of which ETP technologies are being accelerated into utility energy efficiency programs.

8. Codes and Standards and Compliance Enhancement Evaluation Protocol

- This Protocol covers approaches for evaluating codes and standards programs, and for evaluating code compliance enhancement programs. The primary focus of this Protocol is to present the approach for documenting savings from the California Codes and Standards Program and the evaluation of Code Compliance Programs yet to be developed and implemented.
- The Code Compliance Enhancement Protocol is being added at this time because the IOUs are considering the addition of compliance enhancement programs into their energy efficiency program portfolio. The Compliance Enhancement Program Evaluation Protocol is new and has never before been applied within the evaluation community. As a result it is designed to be flexible, allowing a wide range of approaches to be conducted once they are approved by the Joint Staff.
- This Protocol describes how gross and net energy savings will be estimated for programs that change or contribute to a change in building codes or appliance standards that are expected to result in energy savings and programs that are implemented to increase the level of compliance with code requirements.
- We note early in the Protocols that codes and standards evaluations that follow this Protocol are best contracted prior to and launched at the same time that the CEC is assessing which technologies should be considered for the next round of codes or standards changes. This effort is launched approximately three years before a change begins producing energy savings.
- The evaluation contractor selected to conduct the evaluation of the Codes and Standards Programs will need to realize that the change theories and logic models developed by the program will be adjusted and expanded or contracted from time to time as new change-related causal relationships are identified and as program activities are modified to meet the program's objectives.
- These conditions will require a multi-year evaluation effort that is timed to the code program's change process rather than the program implementation cycles, so that

[174] Adoption rates (e.g. the number of measures adopted on an annual basis) for various measures installed through utility resource acquisition programs and associated energy and demand impacts will be obtained from utility program tracking databases. This is generally considered as distinct from a market penetration rate or a saturation rate.

the evaluation contractor can be charged with the responsibility to evaluate a specific set of assigned code or standard changes.

- The evaluation activities conducted under this Codes and Standards Protocol are established to be both prospective and retrospective. They are designed to assess events and conditions that occur in the future, such as the projected energy savings to be achieved. However, they are also designed to be retrospective, with true-up efforts that look back over time and adjust evaluation findings to reflect actual market conditions. As such the evaluations may be contracted in two phases, with the first phase being the assessment and projection of current and future savings, followed by true-up studies that look back and adjust the projected findings and energy savings to reflect actual construction, retrofit, and purchase patterns.
- The first adjustment to the gross energy savings estimate is an adjustment to account for the naturally occurring market adoption rates. New energy efficient products are likely to penetrate and be adopted by at least a portion of the market even without the Codes and Standards Program. As a result, the projected naturally occurring adoption and penetration, which would occur without the program, needs to be subtracted from the program's gross energy impacts.
- Naturally occurring adoption rates for premium energy efficient products typically occur in an "*S"* shape pattern that never reaches 100 percent penetration as long as there are alternative technologies in the market. This is especially true when the alternatives are lower cost technologies. Some energy efficient technologies may never capture a majority of the market share without a mandatory code or standard. Others may move to capture the majority of the market without a code or standard. However, there is likely to always be some level of increased penetration of a superior product that delivers benefits to a user, up to a point of product demand saturation, based on the characteristics of the product and the alternative choices in the market. Similarly, some customers never adopt a new product regardless of the benefits of the product. These customers are typically labeled as "laggards" within the technology adoption literature.
- This step requires the evaluation contractor to establish expected adoption curves for each technology included in the impact assessment. The evaluation contractor will use a range of approaches to establish the estimated penetration curves, including conducting literature searches on the penetration rates of similar technologies with similar product characteristics, the use of expert opinions on the expected penetration rates in the absence of a requirement to use the technology, relevant market data and other approaches as deemed appropriate in the evaluation planning effort.
- The second adjustment to gross savings is an adjustment for non-compliance. Since not all buildings or appliance decision makers will fully comply with the newly adopted codes or standards, these lost savings must be subtracted from the gross estimate.
- In the real world, there is often a range of appliances or measures present in the market, some falling below the standard and some above the standard in their energy efficiency levels. Similarly, technologies that do not comply with the new code or standard are often stocked and sold in the market regardless of the requirements adopted. For example, while programmable thermostats are now required in California for most space heating and cooling applications, it is easy to acquire and install non-compliant thermostats because of the stocking and sales patterns of a wide variety of wholesale and retail outlets, including internet sales.

9. Effective Useful Life Evaluation Protocol (Retention and Degradation)

- One of the most important evaluation issues is how long energy savings are expected to last (persist) once an energy efficiency measure has been installed. The Effective Useful Life (EUL) Evaluation Protocol was developed to address this issue and should be used to establish the period of time over which energy savings will be counted or credited for all measures that have claimed savings. This Protocol contains requirements for the allowable methods for three types of evaluation studies: retention, degradation, and EUL analysis studies.
- A persistence study measures changes in the net impacts that are achieved through installation/adoption of program-covered measures over time. These changes include retention and performance degradation. The definition of retention as used in this Protocol is the proportion of measures retained in place and that are operable. Effective useful life (EUL) is the estimate of the median number of years that the measures installed under the program are still in place and operable (retained).
- The primary purpose of this Protocol is to provide ex-post estimates of effective useful life and performance degradation for those measures whose estimates are either highly uncertain and/or have not been covered in studies over the past 5 years. These results will be used to make prospective adjustments to the measure level EUL estimates and performance degradation estimates for Program Years 2009 and beyond, but will not be used for retroactive adjustments of the performance of the 2006-2008 portfolios.
- Many past persistence studies were unable to provide results that were significantly different (statistically) from the ex-ante results, so that most of the current ex-post EULs are the same as the ex-ante estimates. Besides finding relatively high retention rates in most cases, a consistent and important finding in these studies is that a longer period of time is needed for conducting these studies, so that larger samples of failures are available, and so that technology failure and removal rates can be better documented and used to make more accurate assessments of failure rate functions. The selection of what to measure, when the measurements should be launched, and how often they should be conducted are critical study planning considerations that Joint Staff will direct to ensure reliable results are achieved.
- Performance degradation includes both (1) technical operational characteristics of the measures, including operating conditions and product design, and (2) human interaction components and behavioral measures. This Protocol refers to these two different components of performance degradation as technical degradation and behavioral degradation, respectively. (Performance degradation studies are also referred to in this Protocol more simply as degradation studies.)
- Performance degradation accounts for both time-related and use-related change in the energy savings from an energy efficient measure or practice relative to a standard efficiency measure or practice. It is important to note that the energy savings over time is a difference rather than a straight measurement of the program equipment/behavior. It is the difference over time, between the energy usage of the efficient equipment/behavior and the standard equipment/behavior it replaced that is the focus of the measurement.
- Energy efficiency in both standard and high efficiency equipment often decreases over time. The energy savings over time is the difference between these two curves. The technical degradation factor is a set of ratios for each year after installation/adoption as the proportion of savings obtained in that year compared to the first-year savings estimate, regardless of the retention estimate or EUL (which is

applied separately to obtain overall savings persisted). The technical (or behavioral) degradation factor could be 1.0 for each year in the forecast (often 20-year technical degradation factors are estimated) if the energy efficiency decreases (energy usage increases) by the same percentage each year as the standard equipment. This is the case where technical degradation rates are the same for both types of equipment. The technical (or behavioral) degradation factor would be higher if the efficient equipment holds its level of efficiency longer/better than the standard equipment[175] and lower if there is more relative degradation.

Table A.2: Required Protocols for Measure Retention Study

Rigor Level	Retention Evaluation Allowable Methods
Basic	1. In-place and operable status assessment based upon on-site inspections. Sampling must meet the Basic Rigor Level requirements discussed in this Protocol and must meet the requirements of the Sampling and Uncertainty Protocol. (The sampling requirements of this Protocol may need to meet the sampling requirements for the subsequent EUL study. See below specification.) 2. Non-site methods (such as telephone surveys/interviews, analysis of consumption data, or use of other data, e.g. from EMS systems) may be proposed but must be explicitly approved by Joint Staff through the evaluation planning process. Sampling must meet the Basic Rigor Level requirements discussed in this Protocol and must meet the requirements of the Sampling and Uncertainty Protocol. (The sampling requirements of this Protocol may need to meet the sampling requirements for the subsequent EUL study. See below specification.)
Enhanced	1. In-place and operable status assessment based upon on-site inspections. Sampling must meet the Enhanced Rigor Level requirements discussed in this Protocol and must meet the requirements of the Sampling and Uncertainty Protocol. (The sampling requirements of this Protocol may need to meet the sampling requirement for the subsequent EUL study. See below specification.)

Table A.3: Required Protocols for Degradation Study

Rigor Level	Allowable Methods for Degradation Studies
Basic	1. Literature review required for technical degradation studies across a range of engineering-based literature, to include but not limited to manufacturer's studies, ASHRAE studies, and laboratory studies. Review of technology assessments. Assessments using simple engineering models for technology components and which examine key input variables and uncertainty factors affecting technical degradation. 2. Telephone surveys/interviews with a research design that meets accepted social science behavioral research expectations for behavioral degradation.
Enhanced	1. For technical degradation: field measurement testing. 2. For behavioral degradation: field observations and measurement.

[175] This was found to be the case in 3 of the 25 measures studied in the five persistence studies conducted under the prior M&E Protocols: residential d/x air-conditioning, residential refrigerators, and agricultural pumps.

Table A.4. Required Protocols for EUL Analysis Studies

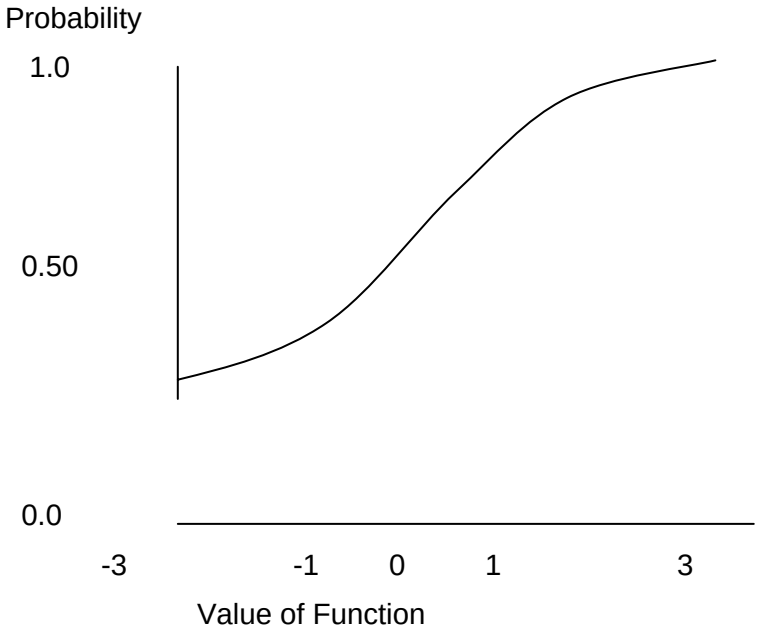
Rigor Level	Allowable Methods for EUL Analysis Studies
Basic	1. Classic survival analysis (defined below) or other analysis methods that specifically control for right-censored data (those cases of failure that might take place some time after data are collected) must be attempted. For methods not accounting for right-censored data, the functional form of the model used to estimate EUL ("model functional form") must be justified and theoretically supported. Sampling must meet the Basic Rigor Level requirements discussed in this Protocol and must meet the requirements of the Sampling and Uncertainty Protocol. Sample size requirements will be determined through the use of power analysis, results from prior studies on similar programs, and professional judgment. Power analysis used to determine the required sample size must be calculated by setting power to at least at 0.7 to determine the sample size required at a 90% confidence level (alpha set at 0.10). Where other analyses or combined functional forms are used, power analysis should be set at these parameters to determine required sample sizes for regression-based approaches and a 90% confidence level with 30% precision is to be used for non-regression components.
Enhanced	1. Classic survival analysis (defined below) or other analysis methods that specifically control for right-censored data (those cases of failure that might take place some time after data are collected) must be attempted. The functional form of the model used to estimate EUL ("model functional form") must be justified and theoretically supported. Sampling must meet the Enhanced Rigor Level requirements discussed in this Protocol and must meet the requirements of the Sampling and Uncertainty Protocol. Sample size requirements will be determined through the use of power analysis, results from prior studies on similar programs, and professional judgment. Power analysis used will set power to at least to 0.8 to determine the sample size required at a 90% confidence level (alpha set at 0.10). Where other analyses or combined functional forms are used, power analysis should be set at these parameters to determine required sample sizes for regression-based approaches and a 90% confidence level with 10% precision is to be used for non-regression components.

- Engineering analysis and M&V observations suggest that energy efficiency measures generally last a certain average length of time and then rapidly move out of use as the measures reach their end of life service. However, these approaches have generally not considered retention and behavioral degradation in establishing the EUL estimates. Similarly, a few measures may continue to last significantly beyond their expected lifetime.
- An initial approximation for most types of EUL forecasts efforts involve some form of a linear estimate, even if the estimate is not linear during the first years of use, or during the later years. This typically involves trying to fit a line to the observed data and use this to predict EUL estimates. Yet, the engineering experience for efficiency measures suggests that a linear model may not represent the survival function of many energy efficiency measures.
- Common alternative models include logistic and exponential models. A variation of the logistic function can be used to describe a pattern of little loss in the early years with increasing loss as the measure approaches its expected life, with a flattening loss occurring thereafter.
- The standard cumulative logistic probability function is:

$$Pi = F(Z_i) = F(\alpha + \beta X_i) = 1/(1 + e^{-(\alpha + \beta X_i)})$$

The logistic model is generally used to measure and predict probabilities that an event will occur. This model limits the end points to zero and one. The cumulative logistic, the logistic model, looks like the curve shown in Figure A.2.

Figure A.2: Cumulative Logistic Function



10. Process Evaluation Protocol

- The process evaluation's primary objective is to help program designers and managers structure their programs to achieve cost-effective savings while maintaining high levels of customer satisfaction. The process evaluation helps accomplish this goal by providing recommendations for changing the program's structure, management, administration, design, delivery, operations or targets.
- The Process evaluation is not a required evaluation activity in California. It is, however, often critical to the successful implementation of cost-effective and cost-efficient energy efficiency programs.
- Process evaluations identify improvements or modifications to a group of programs, individual programs or program components, that directly or indirectly acquire or help acquire, energy savings in the short-term (resource acquisition programs) or the longer-term (education, information, advertising, promotion and market effects or market transformation efforts).
- The primary purpose of the process evaluation is an in-depth investigation and assessment of one or more program-related characteristics in order to provide specific and highly detailed recommendations for program changes. Typically, recommendations are designed to affect one or more areas of the program's operational practices. Process evaluations are a significant undertaking designed to produce improved and more cost-effective programs.

Process Evaluations

- For process evaluations, the focus is on reliability at the program level, with the level of evaluation rigor varying as a function of evaluation priorities and budgets. However, because each program is somewhat unique, with respect to the data being collected and the various sources of bias, there is no specific set of required methods and level of effort for minimizing bias that can be assigned to a program that has been assigned a given level of evaluation rigor.
- *Requiring* 90/10 precision, for example, for all inquiries is very likely infeasible and not cost-efficient because budgets are limited, there is often a large set of evaluation questions to be addressed (i.e., many different questions and parameters for which some level of precision could be desired), not all of which are quantitative, and the information sought from different survey and interview groups might not be equally valuable.
- For example, one might want to field a small survey to get a sense of the motivation of a particular market actor. Again, it is important for the evaluator to have the flexibility to maximize the reliability of their findings. However, the 90/10 level of precision should be adopted as a minimum precision target for the most important data collection efforts on its most important variables. Which data collection efforts and variables are considered to be the most important for process evaluations will be determined by the independent evaluator in close collaboration with utility EM&V staff.
- There are circumstances when it might be desirable to use M&V as input to the analysis of a problem being investigated in a process evaluation. If M&V is not conducted by the Joint Staff evaluations, utility evaluation staff may chose to specify M&V activities within the process evaluation RFP.[176] If the M&V Protocol is used for purposes outside impact, indirect impact and verification analysis, a target precision should, at a minimum, be 30 percent precision at a 90 percent confidence level (or 90/30 precision).
- The evaluator must prepare a detailed plan that allocates resources in order to maximize reliability for the findings and for key parameter estimates for each program in the group. As part of this plan, the evaluator must specifically address the various sources of error that are relevant and explain how the resources allocated to each will minimize and/or mitigate the error.[177] They must also estimate the statistical precision that the planned evaluation will achieve on selected primary quantitative measurements.
- System Learning. The hallmark of any learning system is that feedback is processed and any necessary course corrections are made. Once a particular evaluation is launched, it's certainly possible that mid-course adjustments will be made to the initial plan to maximize savings reliability. For example, the coefficients of variation (CVs)[178] for certain key parameters, measures, end-uses or programs might actually be smaller than anticipated or the random and/or systematic measurement error might be worse. As data are collected and assessed, decisions can be made regarding the reallocation of resources.

[176] Coordination of M&V studies for process and impact purposes is a key issue that must be addressed by the evaluation plans for both process and impact evaluation.
[177] In the pre-1998 Protocols, there was no requirement to address these sources of error in the research plan. Evaluators only had to describe in the final report whether they had to address these various errors and, if so, what they did to mitigate their effects.
[178] The sample standard deviation divided by the sample mean. See page 320 of the *Evaluation Framework*.

- Acceptable Sampling Methods. It is rarely possible, for a variety of different reasons, to conduct a census of any population (e.g., program participants, programs non-participants or lighting vendors).[179] Especially in a state the size of California, this is due largely to the fact that many of the populations are quite large and the cost of attempting a census study would be prohibitive. Instead, random samples drawn from these populations are almost always used as a way to estimate various characteristics of these populations. The specific approaches to maximizing precision are left up to the independent evaluator. For example, one can choose from a variety of sample procedures recognized in the statistical literature, such as sequential sampling, cluster sampling, multi-stage sampling and stratified sampling with regression estimation. There are many available books on sampling techniques that can be used as reference.

11. Market Effects Evaluation Protocol

- The Market Effects Protocol is designed to measure net market effects at a market level when one or more of the Protocol-covered energy efficiency funded program efforts target a market. Net market effects are those effects that are induced by Protocol-covered energy efficiency programs and are net of market activities induced by non-energy efficiency programs including normal market changes.

- The application of the Market Effects Protocol should result in an estimate of the energy (kWh), peak (kW) or therm impacts associated with the net market effects resulting from Protocol-covered energy efficiency program interventions. These net energy market effects are referred to in A Framework for Planning and Assessing Publicly Funded Energy Efficiency (2001 Framework Study) as "ultimate market effects" or "ultimate indicators" because they are the desired indicator of whether net energy efficiency changes are occurring in the market.[180]

- The Market Effects Protocol is designed, therefore, to facilitate not just the estimate of net market effects but also, and primarily, the estimate of net energy market effects. That is, a market effects study both quantifies the changes occurring in the market caused by the energy efficiency programs and provides an estimate of the energy impacts associated with them.

- The Market Effects Protocol does not apply to the measurement of individual program-level market effects or direct program savings typically used for program-level cost-effectiveness assessments and refinement decisions. Rather the focus of the market effects evaluation is at a market level in which may different energy efficiency programs can operate. Yet, the Protocol applies to program-induced market changes that could be missed or double counted if measured program by program. As a result, the use of the Market Effects Protocol should focus on the effects of groups of programs within a market over multiple program cycles.
- Typically these efforts are designed to increase the adoption of energy efficient products, services, or practice and are causally related to market interventions. This definition,

[179] In process evaluations, a census is possible in some more limited populations such as staff and program contractors.

[180] Frederick D. Sebold et al. *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*. (Pacific Gas and Electric Company. 2001): 6-4.

however, was created within the context of guidance for conducting program evaluation of a market transformation style program. A market transformation program is one that is specifically designed and fielded for the purpose of changing the way a market operates so that energy savings are achieved at a market level.

- A more effective definition for the Market Effects Protocol for assessing the market effects from multiple programs that may or may not be designed to change market operations is that in *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs (the Scoping Study)*:

    i. "A change in the structure of a market or the behavior of participants in a market that is reflective of an increase in the adoption of energy-efficient products, services, or practices and is causally related to market intervention(s)."[181] This definition stresses the market rather than the program nature of market effects, and is the working definition for this Protocol.

- Market effects include both short-term and long-term effects. The long-term effects are the most difficult to capture at a program level because they broadly affect a market not just the specific participants in a program or in a grouping of programs. This Protocol targets those long-term effects.
- A market-level evaluation effort is recommended when there are multiple statewide or local interventions in a market such as those of California's energy efficiency programs and where other efforts are also acting to change that market. Other efforts can be associated with the normal operations of the market or when other non-California energy efficiency efforts are changing markets, such as with the national ENERGY STAR ® program, manufacturer promotions and retail sales efforts. A market level effort is also appropriate when a single large and particularly effective program is expected to have broad and long-term market effects in a single market.
- There are two types of market effects discussed in the energy efficiency industry. There are those that are occurring now as a result of how programs are changing markets. And there are those that are forecasted to occur later (after the program has been discontinued) due to the changes established or put into motion by the program. The Protocol recognizes that the methodologies to estimate each of these types of market effects can differ and that potential issues of bias that must be identified, mitigated and minimized are also different. The Market Effect Protocol is designed to measure only the current market effects and not those forecasted to occur at some future point.
- A great deal of effort has been expended over the past 10-15 years to estimate market effects, yet most of these efforts did not estimate net energy market effects, but concentrated on measurement of indicators such as awareness, sales and changes in practices by market actors. Evaluations estimating net market effects with energy estimates, the focus of this Protocol, are at an early stage of development. A variety of studies have been conducted, but only a limited number at the highest levels of rigor.
- However, this is a critically important field of research since the market effects of energy savings caused by California's energy efficiency programs are likely to be substantial once documented. Given the early stage of development of methods, it

[181] Joe Eto, Ralph Prahl, and Jeff Schlegel. *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs*. (Berkeley, CA: Ernest Orlando Lawrence Berkeley National Laboratory, 1996). LBNL-39059 UC-1322, 9.

is important that this Protocol encourage the continued advancement of the field and not prescribe or limit methodological approaches.

- Key to a successful market effects evaluation will be the initial scoping study. The scoping study will define the market to be studied, develop a market theory to test in the analysis, assess data availability for the market effects study, develop a methodology for additional data collection and recommend an analysis approach.
- For programs that are specifically designed to change the way a market operates, the program theory should also be considered in developing the initial scoping study. However, for standard programs that are not designed to change market operations, the program theory is not a significant consideration in the development of the scoping study.
- Because market effects evaluation is still evolving there are a limited, but clearly defined, set of activities that should be considered. Market effects evaluations should be developed using experimental or quasi-experimental designs whenever possible and the approach should be peer reviewed prior to implementing the study to ensure that it will provide valid and reliable results. Triangulation of data and analysis approaches is preferred when possible and teaming with industry organizations and professionals can be beneficial.
- The studies should also take into account regional differences within the market being studied and will at times need to move beyond California boundaries to the regional or national level to collect data. Finally, allocation to utility service territory will be a challenge and dependent on data availability, but should be an important consideration in the scoping study.

Table A.5: Required Protocols for Market Effects Evaluation Scoping Studies

Level of Rigor	Scoping Study Requirements
Basic	Define the market by its location, the utilities involved, the equipment, behaviors, sector and the program years of interest. Develop market theory. Identify available secondary data and potential sources for primary data. Outline data collection and analysis approaches
Enhanced	Define the market by its location, the utilities involved, the equipment, behaviors, sector and the program years of interest. Develop market theory and logic model. Detail indicators. Identify available secondary data and primary data that can be used to track changes in indicators. Outline data collection approach. Recommend hypotheses to test in the market effects study. Recommend the analysis approach most likely to be effective.

- The evaluation contractor will need to articulate a market theory in order to proceed with baseline measurement for market effects evaluation. At a minimum, this market theory shall describe how the market operates and articulate market assumptions and associated research questions. This must be done at a level of detail sufficient to develop data collection instruments for baseline measurement. If the assessment includes programs that are designed specifically to change the way a market operates the program theory should also be consistent with and embedded in the theory of how the market operates.[182]
- Market-level evaluations seek to document the changes in adoption behavior that cause changes in energy savings.[183] It is important, therefore, to clearly articulate the assumed changes in the market, so they can be measured for the market effects

[182] Nicholas P. Hall & John Reed. "Merging Program-Theory and Market-Theory in the Evaluation Planning Process." *Proceedings of the International Energy Program Evaluation Conference* (2001).
[183] Sebold et al., page 6-9, Figure 6-2.

study. If this is done properly the market effects evaluation can document changes in adoption, efficiency and provide an estimate of savings. This process also facilitates model specification.

- A higher level of rigor is achieved when the market theory can be described in a narrative and/or a graphic logic model. A narrative or graphic logic model permits a greater depth of understanding of the indicators driving anticipated market outcomes. It can also help to identify the various sources of influence on market effects outside of the program efforts. The simplest approach to a logic diagram is to view the boxes as potential measurement indicators and the arrows as a hint to questions regarding causal links, program implementation theory, where to examine underlying behavioral change assumptions, and areas for researchable questions.

12. Sampling and Uncertainty Protocol

- The pre-1998 protocols require 90/10 precision for estimates of annual energy use while the 2006 Protocols set precision targets[184] whenever possible for a variety of parameters including savings.[185] Precision targets are set rather than required since, as discussed in the *Evaluation Framework* and its cited study of this issue by Sonnenblick and Eto, bias could be much more important than precision for the reliability of the savings estimates or the cost-effectiveness calculations.
- In addition, as any evaluation study proceeds, the data collected could contain much more error than originally thought, requiring more resources to be devoted to reducing this bias and fewer resources devoted to achieving the required statistical precision. Or, the variability in the savings could be so great that it would be impossible to meet the precision requirement.
- The evaluator must have the flexibility to respond to data issues as they arise in order to maximize the reliability of the savings. Therefore, focusing on sample error, while giving relatively little attention to these other sources of error, would compromise the CPUC's objective of obtaining *reliable* estimates of kWh and kW impacts.
- Finally, the guidelines regarding sampling and uncertainty must be followed for each utility service territory. For example, precision targets, when specified for a particular level of rigor, must be set for *each* utility service territory.

[184] A precision target is a goal established at the beginning of an evaluation based in large part on initial estimates of uncertainty. If an evaluator fails to actually achieve the targeted level of precision, there will be no penalties since the assumptions underlying the sample sizes proposed in each evaluation plan will have been *clearly presented and carefully documented*. A failure to meet the precision target for a given program will only require an adjustment of the input assumptions prior to the next evaluation cycle and, if necessary, a reallocation of evaluation dollars to support increased sample sizes.

[185] The *Evaluation Framework* proposed no precision targets or requirements for savings or for any other parameters associated with such studies as process and market effects evaluations.

Table A.6: Required Protocols for Gross Impacts[186]

Rigor Level	Gross Impact Options
Basic	Simplified Engineering Models: The relative precision is 90/30[187]. The sampling unit is the premise. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
	Normalized Annual Consumption (NAC) Models: There are no targets for relative precision. This is due to the fact that NAC models are typically estimated for all participants with an adequate amount of pre- and post-billing data. Thus, there is no sampling error. However, if sampling is conducted, either a power analysis[188] or justification based upon prior evaluations of similar programs must be used to determine sample sizes. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
Enhanced	Regression: There are no relative precision targets for regression models that estimate gross energy or demand impacts. Evaluators are expected to conduct, at a minimum, a statistical power analysis as a way of initially estimating the required sample size.[189] Other information can be taken into account such as professional judgment and prior evaluations of similar programs. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
	Engineering Models: The target relative precision for gross energy and demand impacts is 90/10. The sampling unit is the premise. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.

[186] See the Impact Evaluation Protocol for a description of methods and page references in the *Evaluation Framework* for further information and examples.
[187] Also of interest, in addition to the relative precision, are the actual kWh, kW, and therm bounds of the interval.
[188] Statistical power is the probability that statistical significance will be attained, given that there really is a treatment effect. Power analysis is a statistical technique that can be used (among other things) to determine sample size requirements to ensure statistical significance can be found. Power analysis is only being required in the Protocol for determining required sample sizes. There are several software packages and calculation Web sites that conduct the power analysis calculation. One of many possible references includes: Cohen, Jacob (1989) *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Inc.
[189] Ibid.

Table A.7: Required Protocols for Gross Impacts[190]

Rigor Level	Gross Impact Options
Basic	Simplified Engineering Models: The relative precision is 90/30[191]. The sampling unit is the premise. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
	Normalized Annual Consumption (NAC) Models: There are no targets for relative precision. This is due to the fact that NAC models are typically estimated for all participants with an adequate amount of pre- and post-billing data. Thus, there is no sampling error. However, if sampling is conducted, either a power analysis[192] or justification based upon prior evaluations of similar programs must be used to determine sample sizes. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
Enhanced	Regression: There are no relative precision targets for regression models that estimate gross energy or demand impacts. Evaluators are expected to conduct, at a minimum, a statistical power analysis as a way of initially estimating the required sample size.[193] Other information can be taken into account such as professional judgment and prior evaluations of similar programs. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.
	Engineering Models: The target relative precision for gross energy and demand impacts is 90/10. The sampling unit is the premise. The sample size selected must be justified in the evaluation plan and approved as part of the evaluation planning process.

[190] See the Impact Evaluation Protocol for a description of methods and page references in the *Evaluation Framework* for further information and examples.
[191] Also of interest, in addition to the relative precision, are the actual kWh, kW, and therm bounds of the interval.
[192] Statistical power is the probability that statistical significance will be attained, given that there really is a treatment effect. Power analysis is a statistical technique that can be used (among other things) to determine sample size requirements to ensure statistical significance can be found. Power analysis is only being required in the Protocol for determining required sample sizes. There are several software packages and calculation Web sites that conduct the power analysis calculation. One of many possible references includes: Cohen, Jacob (1989) *Statistical Power Analysis for the Behavioral Sciences*, Lawrence Erlbaum Associates, Inc.
[193] Ibid.

Table A.8: Required Protocols for Net Impacts

Rigor Level	Net Impacts Options
Basic	For the self-report approach (Option Basic.1), given the greater issues with construct validity and variety of layered measurements involved in estimating participant NTGRs, no relative precision target has been established.[194] To ensure consistency and comparability a minimum sample size of 300 sites (or decision-makers in cases where decision-makers cover multiple sites) or a census[195], whichever is smaller, is required.
Standard	If the method used for estimating net energy and demand impacts is regression-based, there are no relative precision targets. If the method used for estimating NTGRs is regression-based (discrete choice), there are no relative precision targets. In either case, evaluators are expected to conduct, at a minimum, a statistical power analysis as a way of initially estimating the required sample size.[196] Other information can be taken into account such as professional judgment and prior evaluations of similar programs. For the self-report approach (Option Standard.2), there are no precision targets since the estimated NTGR will typically be estimated using information collected from multiple decision-makers involving a mix of quantitative and qualitative information around which a standard error cannot be constructed. Thus to ensure consistency and comparability, for such studies, a minimum sample size of 300 sites (or decision-makers in cases where decision-makers cover multiple sites) or a census, whichever is smaller, is required.
Enhanced	The requirements described for Enhanced apply depending on the methods chosen.

[194] This is considered the best feasible approach at the time of the creation of this Protocol. Like the other approaches to estimating the net-to-gross ratio (NTGR), there is no precision target when using the self-report method. However, unlike the estimation of the required sample sizes when using the regression and discrete choice approaches, the self-report approach poses a unique set of challenges to estimating required sample sizes. These challenges stem from the fact that the self-report methods for estimating free-ridership involve greater issues with construct validity, and often include a variety of layered measurements involving the collection of both qualitative and quantitative data from various actors involved in the decision to install the efficient equipment. Such a situation makes it difficult to arrive at a prior estimate of the expected variance needed to estimate the sample size.

Alternative proposals and the support and justifications that address all of the issues discussed here on the aggregation of variance for the proposed self-report method may be submitted to Joint Staff as an additional option (but not instead of the Protocol requirements) in impact evaluation RFPs and in Evaluation Plans. Joint Staff may elect to approve an Evaluation Plan with a well-justified alternative.

[195] A census is rarely achieved. Rather, one *attempts* to conduct a census, recognizing that there will nearly always be some sites, participants or non-participants who drop out for a variety of reasons such as refusals or insufficient data.

[196] Statistical power is the probability that statistical significance will be attained, given that there really is a treatment effect. Power analysis is a statistical technique that can be used (among other things) to determine sample size requirements to ensure statistical significance can be found. Power analysis is only being required in the Protocol for determining required sample sizes. There are several software packages and calculation Web sites that conduct the power analysis calculation.

Table A.9. Required Protocols for Measure-level Measurement and Verification

Rigor Level	M&V Options
Basic	Simplified Engineering Models: The target relative precision for gross energy and demand impacts is 90/30. The sample unit may be the individual measure, a particular circuit or point of control as designated by the M&V plan.
Enhanced	Direct Measurement and Energy Simulation Models: The target relative precision for gross energy and demand impacts is 90/10. The sample unit may be the individual measure, a particular circuit or point of control as designated by the M&V plan.

Table A.10. Required Protocols for Sampling of Measures Within a Site

The target relative precision is 90/20 for each measure selected for investigation. The sampling unit (measure, circuit, control point) shall be designated by the M&V plan. The initial assumption regarding the coefficient of variation for determining sample size is 0.5.

Table A.11. Required Protocols for Verification

Rigor Level	Verification Options
Basic	The target relative precision is 90/10. The key parameter upon which the variability for the sample size calculation is based is binary (i.e., Is it meeting the basic verification criteria specified in the M&V Protocol?).
Enhanced	The target relative precision is 90/10. The key parameter upon which the variability for the sample size calculation is based is binary (i.e., Is it meeting the enhanced verification criteria specified in the M&V Protocol?).