A National Review of Best Practices and Issues in Attribution and Net-to-Gross: Results of the SERA/CIEE White Paper

Lisa A. Skumatz, Skumatz Economic Research Associates, Inc.
Edward Vine, California Institute for Energy and Environment

ABSTRACT

Energy efficiency evaluation / attribution methods have reached a point that they must evolve in order to provide credible evaluation results for the next generation of programs. Recognizing this need, a national review was undertaken to examine the state of the art, gaps, and next steps needed to meet the evaluation needs for new programs, including behavioral and educational initiatives.

This study used interviews, a literature review, and analysis from around the United States to examine technical, research, and policy issues associated with the attribution of savings to programs – including net-to-gross (NTG) ratios and its components, free ridership, spillover, and other issues. The project reviewed results of net-to-gross (and component) estimations from around the country to identify patterns in results for "categories" of programs, and examined best practices in net savings estimation methods used to date for traditional measure-based programs.

This study found considerable variation in NTG methods, coverage, and component results. This project also examined policies used by different states related to this topic, such as whether NTG or its components are used at all, whether "deemed" levels are used, or whether the regulators endorse or include NTG estimates based on primary research. Protocols from several states were reviewed and compared, and the strengths and weaknesses of the approaches were examined.

Beyond reviewing the "state of the art" in traditional attribution work, savings and NTG issues for behavior, education, and training-based programs were also analyzed. For these programs, savings are difficult to measure, and marketplace "chatter" and overlapping programs and deliverers make measurement especially challenging. Some areas of the country are specifically addressing issues related to errors in measurement associated with NTG, and these results are highlighted. Finally, the project examined gaps in existing research, promising techniques for non-measure-based programs, and recommended next steps.

Project Introduction / Context

On behalf of the California Public Utilities Commission (CPUC), this project sought to identify current and improved techniques – and associated policy issues – related to[1]:

- **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.

[1] This paper presents the findings from one of eight white papers on behavior and energy that were funded by the CPUC and managed by the California Institute for Energy and Environment (CIEE). This work does not necessarily represent the views of the CPUC or CIEE or any of its employees. The white papers are available at: http://uc-ciee.org/energyeff/energyeff.html.

- **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.

The overall research examined four key topics in evaluation: gross savings; attribution / free ridership / net to gross (NTG); non-energy benefits; and persistence. This paper focuses on the second of these evaluation topics. The findings from these evaluation efforts play a critical role in an array of applications, from analysis to program design. Given that evaluation results are often used in making program and reward decisions that put significant investment dollars at risk, it becomes prudent to revisit methods and approaches. Further, as programs have evolved, evaluation has become more complex:

- Programs have moved away from "widget"-based programs toward behavioral, education, advertising, and upstream programs that make it harder to "count" impacts.
- 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 energy efficiency "message" are responsible for the change in energy efficiency behaviors, actions, or purchases. The increased chatter in the marketplace creates a situation in which consumers may be influenced by any number of programs by local utilities as well as influences from outside the utility (national programs, neighboring programs, movies / media, etc.).

As a result, attributing or assigning responsibility for changed behaviors and the adoption of energy efficiency measures or services is muddied and challenging.

For this project,[2] SERA[3] reviewed more than 250 conference papers and reports, and reached out to 100 professional researchers for interviews 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. The white paper addresses all four evaluation topics, but this conference paper focuses only on "net-to-gross" and its constituents, free ridership and spillover.[4]

The literature indicates that there are a number of uses to which free ridership, spillover, or NTG 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

[2] The context for this paper (California) relates to, but is not exclusive to, the situation of programs run by utilities with oversight by a public service commission and where shareholder incentives are at stake and depend on the determination of attribution. This review has relevance beyond this situation, but readers in other states may need to make a few adjustments in terminology, etc.

[3] Skumatz Economic Research Associates (SERA) was commissioned by CIEE to conduct this review. The lead author wishes to thank the following for assistance in preparing the white paper: D. Juri Freeman, Dana D'Souza, and Dawn Bement (Skumatz Economic Research Associates), Carol Mulholland, Jamie Drakos, and Natalie Auer (Cadmus Group), and Gregg Eisenberg (Iron Mountain Consulting).

[4] This paper does not discuss "takeback". An example of takeback is when a homeowner turns up the thermostat after more efficient HVAC systems are installed. This review found little recent work on this topic.

/behavioral programs,[5] 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.

Definition and Methods – Net To Gross (NTG)

Identifying the "net" effects is 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. Estimating the effects of the program above and beyond what would have happened without the program involves 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 though they do not directly receive the program's incentives for those installations and are not recorded directly in the program's "count" of installations. This is called "spillover," and there are three types of spillover:

- Inside project spillover occurs, for example, when refrigerators are rebated, and the person receives / installs that equipment, and then later installs an energy-efficient dishwasher.
- Outside project spillover occurs, for example, when a builder receives rebates on one project, but installs similar efficient measures in other homes without rebates.
- Non-participant spillover occurs, for example, when a builder hears about energy efficiency and does 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.

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.[6] The NTG ratio only equals free ridership (FR) if spillover (SO) is (or is assumed to be) zero. The NTG, or its components, have been addressed in four main ways, described below. Each approach has pros and cons. We list key strengths and weaknesses of each method based on our literature review and interviews with evaluation professionals.

[5] For some of these types of programs, spillover is actually the point of the program, and omitting it ignores important program effects. Ignoring free ridership (in favor of "deemed" NTG figures) allows the continuation of poorly-designed or implemented programs, which wastes ratepayer money.

[6] 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.

Deemed (Stipulated) NTG

A NTG ratio is assumed (1, 0.8, 0.7, etc.)[7] 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, and eliminates debate; no risk in program design or performance; inexpensive.
- Disadvantages: Does not recognize actual differences in performance from different programs, designs, or implementations.

NTG Adjusted by Models with Dynamic Baseline

A baseline of growth of adoption of efficient measures is developed, and the gross savings are adjusted by the changes in the baseline for the period.

- Advantages: Can reflect differences in performance for good or 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 versus a control group. The control group is similar to the test group but does 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 or poor designs and implementation; straightforward concept and reliable evaluation design.
- Disadvantages: Control groups can be difficult to obtain; if imperfect control groups are used, statistical corrections may be subject to protracted discussions.

Survey-Based NTG

A sophisticated battery of questions is asked about whether the participant would have purchased the measures or 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. 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.

[7] If the NTG is less than zero, then this reflects the likelihood of some free ridership.

- 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; requires well-designed survey instrument which can be long and which affects response rate.

The measurement of spillover involves different issues 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.

Our interviews and literature review suggest that a number of states consider free ridership in the calculation of NTG, but do not include spillover in their analyses of program effects, such as California. This analytic asymmetry undervalues energy efficiency by incorporating only subtractions (such as free riders) from gross savings and ignoring potential additions (such as spillover).

Issues and Controversies in NTG Determination

There is considerable – and growing - controversy regarding the use of net to gross, particularly in regulatory proceedings. As noted above, NTG ratios can be used to reduce (incorporating free ridership) or potentially expand (if spillover associated with the program exceeds free ridership) the amount of savings attributable to a program. The concern is that evaluations carefully estimate (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 methods that are less "trusted" – in other words, specifically measuring gross savings based on statistical analysis of meter readings/ billing records, compared to measuring free ridership and/or spillover based on self-report surveys of hypothetical decisions and behavior.

Another controversy relates to 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 a more comprehensive treatment of free riders, spillover, or NTG in the regulatory realm. Furthermore, 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 results in an under-investment in these efforts. Because of their spillover implications, this puts educational (and potentially behavioral) programs at a disadvantage in portfolio development, designing rewards and incentives, and in resource supply applications.

In some states (e.g., California), these measurements have huge potential financial impacts in which utilities may receive financial awards for running programs and running them well. Based on the interviews and research, the controversy seems to arise from the following main sources:

- The potential for error and uncertainty associated with these measurements, because of difficulties in (1) identifying an accurate baseline; (2) identifying and 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 the identification of a reasonable control or comparison group.
- The difficulty in separating out the effects and influences of different programs within a marketplace (own utility / agency and outside utility / agency), often called "chatter".
- Concerns that using measured NTG or free ridership ratios introduces a great deal (to some, an unacceptable level) of risk or uncertainty into the potential financial performance metrics for the program, which will lead to "same old / same old" programs and reduce innovation in program offerings. [8]

Baselines are a very important part of the problem of measuring NTG, free ridership, and spillover. The calculation of baselines is complicated by several factors, including the difference between prescribed and actual practice, and the challenge of documenting what has not happened. 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, but this is not necessarily the case. 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 sector, none of the buildings were being built to the level of baseline codes – i.e., they were underperforming, so that the actual baseline of standard practice was below the baseline of codes. In this case, NTG would be estimated as greater than "one," since the energy efficiency program improved performance over the standard practice baseline.

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.[9] More research on standard practice in the field would provide a stronger basis for baselines and provide a sounder basis for determining NTG ratios.

What Precision Is Needed?

Assuming part of the concern about NTG relates to the accuracy of its computations, two questions arise before either including or excluding NTG – and specifically free ridership - across the board. First, how accurate does the NTG need to be for different possible applications, and second, are there computation approaches that provide that – or those varying – degree(s) of accuracy?

[8] 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.

[9] For example, some upstream changes may 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, then the new product lines will become available in the potential control areas and the (important) market effect is then reduced.

The 2003 Nobel-award winning economist, W.J. Granger, noted that evaluations should be designed to the level of 'helping *avoid making wrong decisions (about programs)*'. The evaluation industry also makes a pertinent point that things that are measured tend to improve. Evaluators want to make sure that the following right decisions are made:

1) Assure public dollars are being responsibly spent;
2) Apportion dollars and efforts between alternative strategies; and
3) Help to identify the appropriate time for exit strategies (or program revisions).

This overriding principle has implications relevant to standards for evaluation in energy efficiency. It implies 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. For example, making a decision on 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. Thus, it is important to see how NTG results will be used, such as in the following activities:

- **Program planning**: Providing estimates of savings attributable to a program that can be used for program planning purposes (e.g., cost-benefit data).
- **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.
- **Integrated planning, portfolio optimization, and scenario analysis**: Providing savings and other feedback across and between programs that helps optimize program portfolios.
- **Generation alternative**: Providing an estimate of energy savings attributable to a program which may support a decision in deferring new generation.[10]
- **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.

The degree of accuracy needed in the NTG computation for these various applications are more stringent (higher) if higher dollars are involved, e.g., if shareholder incentives are involved, or if a new power supply is being sought. The accuracy needed to avoid making a wrong decision varies directly with the potential dollars associated with that wrong decision. To illustrate the point, consider the following. "One size fits all" policies are perhaps not the best approach for including or excluding spillover in NTG computations. Ignoring spillover (because we are concerned that the accuracy of the estimates is of concern) for a program for which spillover is a key goal and outcome increases the chances of making a "wrong decision" about that program investment – and eliminates the chance to improve that performance (assuming measurement breeds improvement). Estimating spillover and applying ranges or confidence intervals to the

[10] For example, if a high amount of savings or value is assigned to the program.

values in assessing the program[11] may be preferable to ignoring spillover. On the other hand, ignoring spillover for a low value program or for a program for which spillover is not an integral part may not be a significant concern.

NTG Practices, Results, and Patterns

Several states use the California Standard Practice Manual, or large portions of it, for estimating energy savings, free ridership, non-energy benefits, and benefit-cost regulatory tests, including Oregon, Washington, Idaho, Montana, Wyoming, Utah[12], 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 free ridership only in the calculation of the NTG ratio. The Nevada Power and Sierra Pacific Power collaborative examined free ridership and spillover in 23 states and/or utilities serving states. They found 15 states (69%) did not use free ridership in estimating net savings (Quantec 2008). Other states say NTG is too costly and biased. Massachusetts prefers to have utilities focus on market transformation programs and correct for factors affecting NTG savings in program design. California requires deemed free ridership values in the calculation of the NTG, but excludes spillover. Several other states say estimating NTG is not a priority - they feel free ridership is balanced by spillover and make no further efforts, argue that measurement of free ridership and spillover is unreliable, or say that when they did measure it the value was close to one.

In Illinois, NTG ratios of 0.8 are assumed for low income programs and are lower for appliance efficiency programs (Baker 2008). Washington reportedly doesn't support savings from behavioral changes or NTG allowances or disallowances (Drakos 2009).

In addition to studies reviewing state and regulatory practices or guidelines, this project also examined patterns in NTG values, results, or methods across programs and regions. The authors assembled and reviewed more than 80 evaluation studies from California, New England, and the Midwest that contained estimates of free ridership and/or other elements of NTG. The studies, which covered residential (including low income) and commercial programs, provided estimates for lighting, HVAC, new construction, appliances, motors, and other measures delivered through incentive and non-incentive programs. The studies covered programs dating from 1991 to 2008. The project examined the studies for patterns in methods between areas of the country, and in free ridership and NTG results by sector, measure, or region. Although the studies were assembled as a convenience sample, and not a statistical sample, we found the following general results, methods, and gaps presented in Table 2.

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.

[11] Or looking for that threshold value of spillover that "turns the decision" may be another way to address the accuracy issue. If the threshold is outside the estimated range for spillover or outside any credible or feasible range based on the rough estimate, the program decisionmaking is improved.

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

Table 2: NTG Results

	Net To Gross , Free Ridership, Spillover
General results	• Most utilities and regulators exclude NTG or assume values that incorporate only free riders 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 (and spillover is more commonly reported in the Northeast than in California). • Most studies rely on self-report surveys using variations in questions incorporating partial free ridership/likelihoods; only a small percent used logit/ranking/discrete choice modeling. • Some studies included both *ex ante* and *ex post* NTG figures for the same program. 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. • Gaps included: Fewer than 10% reported confidence intervals; only a small subset covered NTG for gas savings; and very few studies identified free ridership for electricity savings; most considered only kWh effects.
Variations by measure type, program type or region	• Clear patterns for free ridership, spillover, 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. • *Ex-post* free ridership clustered around 0.1-0.3 but ranged as high as 0.5 to 0.7 for some commercial HVAC / motors and refrigerator initiatives. *Ex-post* NTG clustered around 0.7-1.0, but dipped as low as 0.3 and as high as 1.3. The lowest free ridership was low income programs (as low as 0.03). • NTG for whole homes and home retrofits tended to be high (0.85 to 0.95), but ranged from 0.5 to more than 1.0. • 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 commercial HVAC rebate programs (1.07) and refrigerator rebate programs (1.15). Several programs showed net realization rates between 0.3 and 0.5 including several CFL programs, some refrigerator programs, some gas cooktop rebate programs, and some energy management system initiatives.
Variations for behavioral vs. measure-based programs	• Studies addressing NTG, 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.

Emerging Methods and Recommendations

Based on this project's analysis of the literature and interviews with evaluation professionals, the following findings and recommendations regarding NTG determination are presented:

- **Incorporate the refinements made in standard practices**. Historically, fairly simplistic measurement methods have been used to estimate free ridership. The computations have been based on self-reports. Sources of error with this method stem from faulty recall, bias toward claiming the program was not influential or influential, and from bias introduced in the form of hypothetical questions.

  The literature review noted improvements in self-report methodology including questions to distinguish "partial" free ridership. Later, studies combined partial free ridership with a review of "influencing factors" or "corroborating questions" which were used to adjust free ridership 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).

Other approaches have established multiple criteria for free ridership. In one study, free riders had to meet four criteria: 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 another example, the Energy Trust of Oregon 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 2008).

- **Recognize we may need to allow "credit-splitting or credit-sharing".** 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 and concluded 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 or mean that the program had zero effect. 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.

- **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, with random assignment of eligible participants assigned to treatment and non-treatment groups. This helps address the baseline issue in a credible way. However, to implement this option would require the regulators, utilities, or agencies to "bite the bullet" in terms of the political fallout from those that want to participate but are put into the "no treatment" bucket. Or future participants could be put "on hold" – they could be used as a control group in the short term, but can participate in the program at a later time. This approach may be especially important for outreach and behavioral programs. Train (2009) suggests pairing this with a discrete choice model to predict behavior.

- Many interviewees also agreed that well-designed randomized control and treatment groups are well-suited to impact evaluation (and attribution) for behavioral programs; however, the evaluators and regulators have not developed the kind of faith in them that they have in other programs. The use of these approaches with appropriate modeling (including mixed logit, discrete choice, etc.) shows promise (Ridge et. al. 2009, Train 2009). There is also concern that these random techniques may 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 2009). 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.

- **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 (Gordon and Skumatz 2007). 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 and behavioral programs as well as "widget" programs, but needs testing, as the approach has not been widely applied.[13]

- **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. 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 a 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.

- **Consider compromise or "hybrid" approaches for fiscal-related applications.** 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

[13] It has been suggested that the smart grid or technologies might enhance the opportunity for real time collection of some important data elements.

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, where program performance is checked every 3 years, and where programs are penalized 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. Another "tweak" to test to encourage innovation might be allowing differential rewards: upside incentives could potentially be larger than downside penalties for innovative programs. For some large, important, or innovative programs, negotiations for a priori values might be used.[14] Fiscal incentives must encourage (or at least not penalize) innovation, or only mediocre or "same old" programs will be offered – and they will be offered well past when they should be out of the market.

Reliable measurement methods are available that suit many program types, but more work remains, including research needs in the following areas: [15]

- Greater application of enhanced NTG, free ridership, and spillover methods incorporating partial (and/or deferred) free ridership and corroborating information.
- Greater use of experimental design (including random assignment for participants and non-participants) 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.
- Accumulation of results on elements of NTG in a database and continuously updated with new research and evaluations, so comparisons and tracking are facilitated.

[14] 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).

[15] And, as recognized by one of the paper's reviewers, these "methods-type recommendations" do not touch on issues such as who does the evaluation and the ability to share results for real-time program improvement.

Summary

Estimating the effects of the program above and beyond what would have happened without the program involves a relatively complicated step – identifying the share of energy-efficient measures installed / purchased that would have been installed / purchased without the program's efforts. Traditional elements include free ridership and spillover, combined into a NTG ratio. Spillover is more complicated than free ridership to measure, and as a consequence, a number of utilities that include free ridership never estimate spillover. However, given that many of the benefits from outreach and educational programs – and from a host of "non-widget-based programs – are realized from "spreading the word" (and the behaviors that follow), developing and using reliable and trusted methods that incorporate free ridership in program computations is a priority. These results are needed for applications including program design / assessment / refinement / portfolio development, program exit timing, and incentives.

Reasonable reliability is needed to provide useful information. To provide the best chance for optimal programs, several things are needed. NTG, free ridership and spillover 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, free ridership and spillover 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, free ridership and spillover 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, free ridership and spillover 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 and innovative) programs manageable and reasonably predictable.

Current incentive structures, calculating attribution among actors, and the difficulty in identifying "participants" in new programs are discouraging innovation and leading researchers to consider discarding NTG analyses as a tool in energy efficiency evaluation. This is throwing the baby out with the bathwater. Instead, more widespread application of some of the approaches summarized in this paper can preserve the positives but not be hampered by the negatives of traditional NTG assessment. These evaluations are needed to "help avoid making a wrong decision…" with the public's money. To do this effectively, we need good methods, and we need to make sure the results are fed back into programs to be used in decision-making.

References[16]

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

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

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

Drakos, Jamie, 3/31/09. Cadmus Group, Portland, OR, Personal Communication with the Author.

Gordon, Fred, 7/25/08. Oregon Trust, Portland, OR, Personal Communication with the Author.

Hedman, Brian. 3/31/09. Cadmus Group Portland, OR, Personal Communication with the Author.

Mahone, Douglas. 2008. Email from Douglas Mahone to Robert Kasman (PG&E), June 12, 2008, provided to author.

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

Quantec, Scott Dimetrosky. 2008. "**Assessment of Energy and Capacity Savings Potential in Iow**a", Prepared for the Iowa Utility Association, February 15, 2008.

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 Ratio**", *Proceedings of the IEPEC Conference*.

Saxonis, William. 2007. "**Free Ridership and Spillover: A Regulatory Dilemma**", *Proceedings of the IEPEC Conference*.

Gordon, Susie, and Lisa A. Skumatz, 2007, "**Integrated, Real Time (IRT), On-Going Data Collection For Evaluation – Benefits And Comparative Results** (Gordon & Skumatz, 4323), Proceed*ings for the European Council for an Energy Efficient Economy (ECEEE)*, France, June.

Skumatz, Lisa A., 2009. "**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 for CIEE, December. Available at: http://uc-ciee.org/energyeff/energyeff.html.

Skumatz, Lisa, Dan Violette, and Rose A. Woods, 2004. "**Successful Techniques for Identifying, Measuring, and Attributing Casualty in Efficiency and Transformation Programs**." *Proceedings of ACEEE Conference*, Asilomar, CA.

[16] The long list of references underlying this project is cited in Skumatz 2009, the full report from which this article is excerpted.

Sulyma, Iris M., Power Smart, BC Hydro, Vancouver, Canada, Personal Communication with the Author, 4/28/09.

Train, Kenneth, UC Berkeley and N/E/R/A, CA, Personal Communication with the Author. 10/6/09.