Emerging Evaluation Issues Revisited

Edward Vine, Lawrence Berkeley National Laboratory and California Institute for Energy and Environment
Nick Hall, TecMarket Works
Kenneth M. Keating, Independent Consultant
Martin Kushler, American Council for an Energy-Efficient Economy
Ralph Prahl, Prahl & Associates

ABSTRACT

In this paper, we focus on a select group of technical and policy issues, which are currently important and/or are expected to become more critical in the coming years. The first set of technical issues deals with the evaluation of (1) persistence, (2) behavior and behavior change, and (3) rebound. We provide an overview of the importance of these issues, discuss key data collection and analytical challenges involved in evaluating them, and identify some recent methodological advances that have been made in these areas. These technical issues are becoming more important as energy efficiency and demand side management are increasingly being relied upon as a means of achieving long-term energy resource and environmental objectives. The second set of policy issues deals with (1) the evaluation of energy efficiency at the "policy" rather than the "program" level; (2) the use of "top-down" rather than "bottom-up" evaluation of energy efficiency programs and policies, and (3) closing the loop between evaluators and implementers. We provide an overview of the importance of these issues, particularly as seen by policymakers at the state, federal and international levels.

Introduction

The evaluation, measurement and verification (EM&V) of energy efficiency (EE) programs has a rich and extensive history in the United States, dating back to the late 1970s. In a previous paper for this conference (Vine et al. 2010), we reviewed a number of key EM&V issues facing the energy evaluation field (net energy savings calculation, market transformation evaluation, carbon emissions calculation, evaluation metrics, evaluation practice, national EM&V protocols, developing a professional evaluation community and workforce, and training the next generation of evaluators), but, due to lack of space, we were unable to cover several additional key issues. In this paper, we highlight additional issues, which are currently important and/or are expected to become more critical in the coming years.

The first set of technical issues deals with the evaluation of (1) persistence, (2) behavior and behavior change, and (3) rebound. We provide an overview of the importance of these issues, discuss key data collection and analytical challenges involved in evaluating them, and identify some recent methodological advances that have been made in these areas. The second set of policy issues deals with (1) the evaluation of EE at the "policy" rather than the "program" level; (2) the use of "top-down" rather than "bottom-up" evaluation of EE programs and policies, and (3) closing the loop between evaluators and implementers. We provide an overview of the importance of these issues, particularly as seen by policymakers at the state, federal and international levels.

Technical Issues

The first set of technical issues deals with the evaluation of (1) persistence, (2) behavior and behavior change, and (3) rebound.

Persistence

Persistence refers to the reliability of a measure over time: in particular, how long can one rely on the savings for a particular measure? The period of savings persistence is often represented by the expected life of a measure and is a critical component that is used in cost-effectiveness tests, technical reference manuals, and procurement decisions. Measure life is defined as the median effective useful life (EUL) of the measure – i.e., the median number of years that a measure is likely to remain in-place and operable. 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.

More than 100 persistence (also known as retention or measure life) studies have been conducted, examining *in-situ* median lifetimes for residential and non-residential measures (one of the earlier papers was by Keating (1991); see Skumatz et al. 2009 for a more recent and comprehensive listing). A review of results from measure-based EUL studies around North America showed that measure lifetimes for the type of measures normally offered in prescriptive programs and installed in typical end-uses are fairly consistent for many measure-based programs in commercial, residential, and industrial sectors (*ibid*).[1]

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, given typical technical and resource limitations. The lack of activity may also reflect the fact that whether or not something is cost-effective does not change significantly with small changes in measure-based EULs (because the economic value of additional savings in the latter years of a measure lifetime is typically significantly discounted). Finally, the most pragmatic reason may be that there are significant challenges to conducting measure life studies outside of the laboratory (Skumatz et al. 2009), such as the following: data analysis limitations (insufficient sample or insufficient failures to allow reasonable estimation, or weak model selection), long lifetimes of measures (making it impractical to wait for determining median survival time), incomplete data sets, high cost of data collection, and need for trained staff.

A key factor affecting how much savings is being delivered from program-related installations of EE equipment is whether the measures perform at the new efficiencies consistently over time, or whether their efficiency performance degrades over time. Decays in net technical performance could be an important issue, particularly for measures for which savings are assumed to accrue over long periods of time. Unfortunately, in a review of more than 100 EUL and degradation studies, there have been very few degradation papers within the last decade that have been based on primary data. And even these note the problem of finding enough replacements to set reliable EUL estimates (Skumatz et al. 2009). Studies conducted in the mid-1990s in California on different types of equipment (Proctor Engineering Group 1999) concluded that degradation – above the degradation pattern that would be realized in standard efficiency equipment – was very unlikely for the majority of equipment types (e.g., residential air conditioners and refrigerators), but that technical degradation was suggested for some equipment (e.g., commercial package air conditioners, oversized evaporative cooled condensers, and high intensity discharge lighting).

[1] They also noted a lack of in-depth studies in process equipment, some shell measures, and specific end-uses like cooking, refrigeration, and air compressors.

One additional complicating factor is the issue of the persistence of the associated behavior (i.e., buying the energy-efficient line of equipment) rather than just the effective useful life of a technology. Savings can live as long as the behavior is achieved, well past the life of the first purchase of that type of technology. Few studies have focused on assessing if the program-induced behavior is transformative for the participant. That is, 20 years later, are they still making the energy-efficient equipment choice extending savings beyond the EUL of the program-incented measure?

Given the lack of recent studies, primary research is needed for collecting data on both technical and behavioral degradation, particularly for measures with technical or engineering changes that may affect the net degradation rates of specific equipment types relative to the degradation that would be expected with older technology – or measures accounting for large shares of portfolio savings. Furthermore, these studies need to account for equipment with significant changes in behavioral (operational or upkeep) elements. Priority-setting for new research on this topic should take both factors into account (mechanical and behavioral), and resulting figures should be verified periodically.

Early replacement programs present their own challenges in determining the appropriate calculation of measure life (Skumatz et al. 2009), because the issue of likely remaining life of the replaced measure becomes critically important. In contrast to "naturally occurring" market transactions (e.g., new equipment purchases, "replace on burnout" decisions, etc.), early replacement programs are 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. In these early replacement programs, the critical question is: should the savings be calculated as the difference between energy use for the old measure replaced and the new EE measure ("Savings 1"), or should it be calculated as the difference between energy use for the new standard measures available on the market and the new EE measure ("Savings 2")? Hence, should early replacement programs be able to take credit for extra savings – "Savings 1" during early replacement period (its "remaining useful lifetime" (RUL) before its EUL), and "Savings 2" for the period after the normal measure lifetime for the old equipment? This is the approach that evaluators typically refer to as the "*stair step"* approach because of the two levels of savings estimated over the two different periods of time. 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? Research is clearly needed to identify best procedures for identifying the "hypothetical" expected removal date for early removals. There is some work starting in New York on this issue (see Ridge et al. 2011), and we expect more attention in the future if early replacement programs offer incentives.[2]

Finally, perhaps the most critical area of need in terms of persistence research is in regard to the effects of education / outreach / behavioral programs that do not incent a specific targeted behavior (such as in a typical energy efficiency rebate program that causes a purchase behavior change to occur). Quality studies examining the persistence of more general behavior-induced savings from education-oriented interventions are virtually absent. On the face of it, it appears that standard experimental design measurement methods may work well for quantifying the immediate impact of these programs, however, these studies require very large populations with very well controlled analysis methods. Moreover, there is a need for more detailed research to better understand how these programs work, and to identify more clearly what behaviors participants actually change in response to these interventions. These studies need to be especially rigorous and conducted independently by skilled evaluation professionals. However, this is an important gap in our library of persistence information, as behavioral and market-based programs have become a larger share of utility/agency portfolios, as described in the next section.

[2] In at least one study of an incentive-based early replacement program, high rates of free riders were found, so that early retirement was not really achieved (personal communication with Robert Wirtshafter, May 24, 2011).

Behavior and behavior change

Most program evaluations have focused on the evaluation of EE technologies in the residential, commercial and industrial sectors. There is now a growing interest in understanding the human dimensions of energy use and EE – both as it affects energy use in a building and as it affects investment decisions: for example, (1) funding provided by the California Public Utilities Commission (CPUC) on behavior and energy white papers,[3] (2) the CPUC's Decision (D.09-09-047) indicating their intent to "…consider expedited approval of new EM&V methodologies to verify savings driven by behavior-based efficiency programs (currently considered non-resource programs)," (3) the large attendance and presentations at the annual Behavior, Energy and Climate Change Conference[4], (4) the proliferation of vendors offering indirect feedback of utility data to customers (e.g., Duke Energy's Home Energy Comparison Report, OPOWER, Efficiency 2.0, and Google.org (see Donnelly 2010)), (5) the increasingly widespread implementation of utility programs using these vendor-based technologies, and (6) an extensive research program on evaluating energy behavioral change programs in Europe (Gynther et al. 2010).

While "human behavior" is ultimately involved in every type of energy program (e.g., the decision to buy a high-efficiency appliance and claim a rebate requires a change in purchasing and investment behavior), the aspect that has been experiencing a resurgence of interest in the last few years is the issue of how to influence consumers to implement EE through everyday practices and usage behaviors not involving technology replacements. It is that aspect of behavioral program evaluation which will be focused upon in this paper.

The evaluation of behavior-focused programs presents key data collection and analytical challenges. To begin, one key problem is that behavioral programs rely upon an assumption of repeated, difficult-to-observe minor behaviors by a participant, rather than single discrete actions such as the purchase of a major EE measure. More broadly, the fundamental challenge is understanding more about if, how, when and for how long behavioral programs truly impact energy consumption behavior, and separating out effects that can be attributed to that program. Some recent studies have suggested that savings from these types of programs may come from specific residential segments who typically save energy, while other participating segments may increase energy use (Integral Analytics et al. 2011). Presently, the commercialization of approaches and the reliance of administrators and states on these commercialized approaches are outrunning the ability of evaluators to demonstrate their value. Many claims are being made, but without a systematically strong scientific analysis of these claims. The use of true experimental research designs with random assignment to treatment and control is helping to overcome some of these concerns (see Vine et al. 2011), but much more evaluation work remains in order to truly understand if and how these programs affect behavior, and virtually no comprehensive studies of the nature suggested in the previous section have been done on these types of programs.

Another key issue is the reliability or persistence of the savings from these programs. Since we do not have experience in the long-term persistence of behavioral change, agencies might want to consider requiring new behavioral programs to conduct retention assessments every year or two 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 (Skumatz et al. 2009). As noted earlier, the EUL measurement approaches will need to be tested and applied to a variety of usage behaviors not involving technology replacements if EULs of these behaviors are to be calculated. Some may parallel traditional EUL estimation best practices, but the application of statistical approaches to some programs may be more challenging.

---

[3] These papers are found at: http://uc-ciee.org/index.php?option=com_content&view=article&id=18&Itemid=47)

[4] BECC: www.BECCconference.org

As noted above, quality studies examining the persistence of behavior-induced savings are virtually absent. In addition, it is important that future studies need to look at the different types of customers participating in these types of programs. For example, the EUL of behavior change savings can be substantially different for customers with different psychographic profiles – with some segments not saving energy at all, some increasing consumption, some saving a small amount, and some saving a bit more. Therefore, to understand persistence from behavior change programs, you have to also understand the psychographic profiles of the market reached by those programs and how they impact savings rates and persistence over time.

The attribution issue is critical for these types of program as more media messages on behaviors and education bleed across territories (see Vine et al. 2010). 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. Related to this issue is the issue of double counting. For example, the evaluation of a proprietary comparative information feedback program in Massachusetts found that the savings from this program appeared to be due more to the adoption of physical measures (potentially via other energy efficiency programs) than to "behavioral" changes. They plan to assess to what extent the physical measures may be occurring in conjunction with other programs. To the extent they are, and decisions to purchase were not uniquely influenced by messages and other feedback that the behavior treatment group received, there is double counting that needs to be adjusted for, although the comparative information feedback program may still be valuable as a recruitment mechanism.

Experimental research design with large populations and strong segmentation analysis will be critical to understanding the operation of these programs and their impact on consumers, as well as their inclusion as a reliable resource in procurement plans (Vine et al. 2011). The CPUC is presently conducting an evaluation of Pacific Gas & Electric Power's comparative information feedback programs using a complete experimental research design, and Duke Energy also used an experimental research design to evaluate their home energy comparison report program (Integral Analytics et al. 2011). The development of more detailed evaluation protocols for evaluating these types of programs may be needed. The CPUC is exploring the need for such a protocol in California, and the Electric Power Research Institute's (EPRI 2010) report on feedback programs provides some excellent guidance material as a starting point.

Finally, one pragmatic option that merits some consideration is to avoid the need for confronting these very difficult measurement and attribution issues by taking these behavioral initiatives out of the "energy efficiency program portfolio" and re-categorizing them as core utility "customer service" features. For example, activities such as providing periodic customer feedback, comparing consumption to neighbors, etc. could be made part of the utility billing system, paid for by regular utility funds, and any effects would simply become part of the "baseline" conditions within which the EE portfolio programs operate. While such a solution is not as exciting for evaluators, it does solve a number of practical problems, and this would avoid lengthy arguments and litigation over behavior savings estimates in regulatory forums tasked with determining the net impacts of EE portfolios and/or calculating the amount of shareholder incentives.

Rebound

For the purposes of this paper, "rebound" effects (also called takeback or substitution effects) refer to those effects that can mitigate the reductions in energy consumption associated with EE. David Owen's recent article in *The New Yorker* (2010) on EE and rebound phenomena has sparked a lively debate about the potential for improvements in EE to more than negate environmental gains. For the evaluation of EE programs, we limit our discussion to the direct and indirect rebound effects for

consumers. The direct rebound effect on the consumer side theoretically arises because an EE gain reduces the effective price of energy, potentially causing consumers to use more of it. An example is the installation of more efficient heaters or air conditioners that causes the household to heat or cool more rooms. The indirect rebound effect on the consumer side arises from consumers taking the money saved from, say, buying a more efficient refrigerator, and spending it to purchase other goods and services that require energy.

David Owen uses air conditioning as an example of indirect rebound. According to Owen, more efficient air conditioners have led to a decrease in the cost of running an air conditioner, and the decreased costs, therefore, have made air conditioners more affordable to more people. As a result, more people have bought air conditioners, leading to increased electricity usage. Owen's critics, such as Steve Nadel (2011), argue that the causes of rising use of air conditioners were due to rising household incomes and the declining price of air conditioners, not because of greater EE. Similarly, he argues that rising incomes and declining costs are driving growing saturations of microwave ovens, personal computers, and flat screen televisions, and that improved EE has contributed only marginally to the growing use of these services. He concludes that EE has helped to moderate (but not eliminate) the associated increases in energy use as these services grow.

Clearly, it is theoretically possible for some consumers in some situations to act in accord with Owen's theory, but even that is correlation, and causation is still questionable. Unfortunately, the EE evaluation industry is not very well positioned to respond to these arguments, because we have not made any significant effort to study the issue of rebound in the last 18 years. Nadel (1993) serves as the last best review of rebound studies in EE programs: from his review of 42 studies, he concluded that rebound could occur but that it was not a widespread phenomenon. Instead, he noted that rebound was more likely a localized phenomenon, largely limited to specific end uses (e.g., residential lighting (10% increase in operating hours due to the installation of CFLs), and industrial plant production (2% increase due to the installation of EE process measures)). For other end uses, he found no data or inconclusive data supporting the rebound effect.

We do not wish to revisit the methods used to derive the above estimates, or the particulars in the arguments for supporting or criticizing the rebound effect. But we do want to alert evaluators that they should be aware of: (1) these rebound studies and the implications for their work – particularly for those working on potential studies and carbon emission reduction plans and policies; and (2) the methodological issues associated with these studies – in particular, the reliance on a few questions in self-reported surveys and small samples of households or buildings for the micro effects analysis. More research is clearly needed, so that advocates and opponents of the rebound issue can have a firm basis to support their positions!

Policy Issues

The second set of policy issues deals with (1) the evaluation of EE policies, (2) top-down evaluations of EE programs and policies, and (3) closing the loop between evaluators and implementers.

Evaluation of energy efficiency policies

Experience in policy instruments for EE in the building sector goes back to the 1970s, and today more than 30 different types of policy instruments are in use around the world to overcome and eliminate barriers to EE (Kiss et al. 2010). Policy evaluation research is often concerned with the bottom-up simulation and modeling of several EE policy instruments (e.g., taxes, tax credits, building codes, and communication campaigns) (Mundaca and Neij 2010). These models (simulation, optimization, accounting, and hybrid models) often address different objectives. For example, the

simulation models try to replicate end-user behavior for technology choice considering different drivers (e.g., income, energy security, energy prices, etc.), while optimization models attempt to find least-cost solutions for technology choices for energy systems based on various policy and market constraints. Accounting models describe the physical flows of energy, and hybrid models merge different methodological components from the above models. All of the models are driven by economic and engineering principles, and, in terms of technology choice, most models use techno-economic data as the main criteria. The main function of bottom-up models is to describe and allow the examination of the current and future competition of technologies in detail, for example, by showing different technology prospects and resulting economic and environmental impacts. These bottom-up models have primarily been used to evaluate *ex-ante* the performance of EE policy instruments, and they often contribute to scenario studies on whether and how a portfolio of policies and programs can achieve energy saving or greenhouse gas (GHG) emissions targets. For example, Lin et al. (2010), using the LEAP model, estimated the reduction potential for energy consumption and GHG emissions in Xiamen for a time frame of 2007-2020: in their scenario with energy policy measures, they found that energy consumption and GHG emissions would increase (9% and 7%, respectively), but less than the business-as-usual (BAU) scenario where consumption and emissions would increase 11% and 10%, respectively.

In contrast, particularly in the U.S., *ex-post* evaluations of policy instruments have also been bottom up but primarily focused on collecting and analyzing empirical data on utility programs (rebate, audits, demonstrations, educational centers and training program - see CALMAC and CEE databases of evaluation studies[5]). We are aware of several empirical bottom-up evaluations of policies: e.g., building codes, appliance standards, 10 policies in Denmark (Togeby et al. 2010), and 20 policies across Europe (Khan et al. 2007). Many past evaluations of appliance-labeling programs have focused on consumer awareness of the label but have not explicitly linked the label to actual behavior (i.e., to the efficiency of the appliances purchased and to the most likely purchase if there had been no label) (Vine et al. 2001). However, some evaluations of appliance-labeling programs do include data on actual sales and behavior. Finally, a retrospective examination of EE policies covering appliance standards, financial incentives, information and voluntary programs, and government energy use (building and professional codes were not included) was conducted at the beginning of the last decade (Gillingham et al. 2004).

The analysis of policy is confronted with several challenges, similar to those challenges affecting the evaluation of EE programs: data problems, limits to information, and deep-seated methodological challenges and debates about how to properly measure and predict the costs, benefits, and effectiveness of past and prospective policy (Gillingham et al. 2004). The key methodological issues dealing with free riders, spillover, rebound, discount rates, non-energy benefits and costs, and attribution have either been discussed in this paper or in our previous paper (Vine et al. 2010). All of these potential sources of error suggest that considerable care must be taken in interpreting existing estimates of the costs, energy savings, and cost-effectiveness, from EE programs and other policies. And while these sources of error are shared by program and policy evaluations, the evaluation of policy instruments may have one unique analytical challenge: the inability to clearly identify and delimit the actual "policy" that is being measured. The policy logic model is very complex at many levels and, therefore, data may need to be as widely collected as the parties and channels involved are widely scattered. Thus, in contrast to single program evaluations, policy evaluation may need to (1) attribute savings to each one of the different policies and programs targeting one area (such as campaigns, audits, energy performance certificates, financial incentives, and standards on building renovation) while excluding double-counting, or (2) avoid attribution at all (if the question is what are the total savings from the portfolio), and exclude double-counting for evaluating the effect of the whole package/portfolio (Thomas et al. 2010).

The scarcity of independent academic studies that take a detailed look at the effectiveness and

[5] CALMAC: www.calmac.org; CEE: www.cee1.org.

the costs of specific programs and policies after they have been implemented needs to be rectified (Gillingham et al. 2004). More importantly, as noted by Kiss et al. (2010) in their review of policies in the Nordic countries, many policy instruments have too little focus on systematically demonstrating their effects in terms of actual energy savings. It is vital that the evidence of concrete energy savings and other desirable impacts becomes an integrated part of policy instruments. Evaluation should not be viewed as an add-on or at worst a "distraction." Evaluations should be integrated into policy instruments to provide continuous feedback. Modeling and scenario methods should be complemented with other types of methods to validate results and recommendations. A combination of methods is important.

Process evaluations of EE policies are critical: it is vitally important to examine how a policy is implemented, including how it is communicated to both administrators and market actors, and how it is (or is not) enforced in the field. Just as with "program" evaluation, it is important to understand how the policy design is actually being implemented, and how the design and/or implementation might be improved. In that same vein, most policy evaluations have had little focus on how to improve learning (Khan et al. 2007); instead, what we see are sporadic or "ad hoc" evaluations. For policy instruments to be designed and implemented successfully – resulting in the desired impacts – a long-term strategy is required that provides clear signals to actors in the building sector. Strategic policy evaluations are a vital part of efforts on EE and buildings.

The vast majority of evaluations of policy instruments in the U.S. has occurred at the state level (i.e., utility programs) rather than at the federal level. In order to conduct robust policy analysis at the national level, a political environment that welcomes objective analysis is needed. In the last few years, that political environment has changed in ways that are both more supportive of, and more challenging to, rigorous evaluation. Most recently, in President Obama's FY2012 Budget (Section 8, Program Evaluation) proposal, he noted that, historically, evaluations had been an afterthought when programs were designed, and once a program had been in place for a while, building a constituency for rigorous evaluation was hard. In his budget proposal (continuing previous guidance from his Administration), he noted that he wanted to build an evaluation infrastructure for evaluation in federal agencies so that new or expanded programs have evaluation "based into their DNA." On the other hand, we recognize the significant political challenge inherent in having policymakers agree to an independent, unbiased evaluation of a policy that they have either publicly promoted or opposed. And today's 'super-charged' partisan environment may make consensus on objective evaluation more difficult.

Finally, one particular policy-related issue that we would like to mention specifically is the strong interest in "jobs" that has emerged in the past few years, as a result of the economic recession. While we don't have the space to explore that issue in this paper, we want to acknowledge its major and growing importance to policymakers, and note that evaluators can expect to be asked what they can contribute to quantifying the economic and jobs impacts of EE programs and policies through *ex-post* evaluation. We believe this will be an important emerging area of focus for evaluators.

Top-down evaluations of energy efficiency programs and policies

In Europe, government policymakers are encouraging the use of top-down methodologies for evaluating the impacts of EE programs and policies. These top-down evaluations look at national or regional energy usage, trends, and changes, as well as the causes of these changes, some of which may be program interventions. The most extensive review of top-down methods has been conducted by the European project EMEEES (Evaluation and Monitoring for the EU Directive on Energy End-Use Efficiency and Energy Services)[6] (Lapillonne et al. 2009). While promising, the EMEEES project found that the top-down evaluation method was limited due to data problems, such as the following: (1) data

[6] The EMEEES project: http://www.evaluate-energy-savings.eu/emeees/en/the_project/project_description.php

were not available for some countries to calculate the required indicator, (2) only aggregated data were available in some countries (when disaggregated data would enable researchers to remove the effect of hidden structural factors and enable a better assessment of energy savings), and (3) for some end uses or sub-sectors, only an aggregated indicator could be used to calculate savings, and the indicator was increasing when energy savings should make it decrease (and vice versa) (*ibid*). There were also other issues and problems related to the calculation of policy-induced (additional) energy savings, such as: (1) what value should be used for price elasticity (the value obtained for most countries was not statistically significant), (2) how to precisely derive autonomous technological trends (reflecting technological progress without policies and programs), and (3) the calculation of additional energy savings from increases in market prices.

Aside from a few researchers (e.g., Horowitz 2007 and 2010), top-down evaluations have not been the focus of activity in the United States. The CPUC has recently shown interest in this approach and has approved the funding of a white paper and possibly a pilot study in looking at alternatives to estimating energy savings measure by measure, as is the current practice. In the CPUC's EM&V decision (Decision 10-10-033, October 28, 2010), the CPUC directed the Energy Division to assess, explore, and test the viability of measuring the reduction in energy consumption due to the various EE programs and efforts in California from aggregate consumption data. This interest is driven by the CPUC's need to develop robust methods to assess progress towards achieving the carbon emission reductions resulting from EE required by the California's Global Warming Solutions Act (Assembly Bill 32) and the CPUC's adoption of the California Energy Efficiency Strategic Plan, which is intended to set the utility programs on a course towards market transformation-oriented goals.

If California and other states (or the federal government) decide to pursue a top-down evaluation methodology, they will most likely need to "harmonize" the top-down and bottom-up methods – for example, this is required by the Energy End-use Efficiency and Energy Services Directive (ESD) of the European Union (Thomas et al. 2010). EMEEES has developed an integrated system of bottom-up and top-down methods for the measurement of energy savings (*ibid*). While the EMEEES project provided recommendations for harmonization, debates continue on the issues of simplicity vs. precision and which factors to include or not, the effort needed, and in which areas to use bottom-up vs. top-down calculations. This debate also appears to be based on traditions that some Member States have in using methods, or their level of policy ambition. A similar debate will undoubtedly occur in California and nationally.

Closing the loop between evaluators and implementers

"Closing the loop" refers to ensuring, at a minimum, that the results of program evaluation are provided to program planners and managers for consideration. There is also an expectation that some or all of the evaluation results are actually used by implementers to improve the design or performance of their programs. Encouraging the use of evaluation results by program managers and other stakeholders has become a more prominent and publicly visible topic as the budgets for EE program implementation and evaluation have increased. As described in Vine (2008), the type of interest in and use of evaluation varies by functional role (evaluator versus implementer), maturity of the EE market, institutional context (e.g., evaluation and implementation conducted inside the same organization, or evaluation and implementation conducted by separate entities), and by regulatory demands and interests (e.g., evaluation as an accounting audit (report card) or evaluation as an information source for helping to transform markets by providing information on lessons learned or best programs).

After reviewing the communication of evaluation results, the use of program evaluation results, the needs of program implementers, and the role of utility regulators affecting the use of evaluation results, Vine (2008) identified the following mechanisms that need to be used to facilitate the use of

evaluation:

- Implementers, evaluators and regulators must work together as a team.
- Implementers must conduct evaluability assessments.
- Implementers and regulators must track how evaluation recommendations are used by implementers.
- Evaluators must provide evaluation findings that are readable, fair, accurate, and actionable – making evaluation results more specific and directly relevant to the needs of the program.
- Evaluators and regulators must provide more real-time feedback to implementers – for example, by establishing forums for sharing evaluation information.
- Regulators must:
    - Require that implementers have reviewed, commented and indicated that they are (or not) responding to the evaluation recommendations.
    - Provide sufficient resources for evaluators to address the needs of implementers.
    - Require program administrators to support their energy-saving calculations based on findings from evaluation studies.
    - Create performance metrics that go beyond direct energy savings and include factors such as non-energy impacts, customer satisfaction, and market effects.

For these mechanisms to work, evaluators, implementers and regulators must explicitly endorse learning from evaluation by presenting evaluation as a win-win collaboration, rather than a win-lose proposition. Regulators and implementers also need to make an explicit commitment to make evaluation results a part of their decision process regarding the planning, design and operation of programs. And sufficient resources must be dedicated to evaluation, so that people understand that the state (or sponsoring organization) has identified evaluation as a critical activity. This will help institutionalize the value and use of evaluation – as discussed above, this is what the Obama Administration is attempting to do at the federal level.

Conclusions

The evaluation of EE programs is at a critical stage. The last few years has seen a resurgence in the interest and attention paid to the evaluation of EE programs at both the state and federal level. At the same time, the recent economic meltdown has made policymakers and other stakeholders wary of spending too much money on evaluation (even in California!). The value of evaluation is being questioned by non-evaluators. And while the evaluation challenges described in this paper and our previous paper are meant to be used as a foundation for improving the practice of evaluation, the "critics" of evaluation see these challenges reflecting weaknesses in the conduct of evaluation, making them more wary of supporting such an enterprise. We hope that these critics understand that similar challenges are inherent in most other professions – in fact, some argue that the "evaluation community" is more self-critical than any other field. We hope that the use of evaluation will continue to be supported in these tough economic times, for the benefits of conducting evaluation are immense.

References

Donnelly, K. 2010. "The Technological and Human Dimensions of Residential Feedback: An Introduction to the Broad Range of Today's Feedback," in K. Ehrhardt-Martinez and J. Skip Laitner (Eds), *People-Centered Initiatives for Increasing Energy Savings*, Washington, D.C.: American Council for an Energy-Efficient Economy.

Electric Power Research Institute (EPRI). 2010. *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols.* Report 1020855. Palo Alto, CA: EPRI.

Gillingham, K., R. Newell and K. Plamer. 2004. *Retrospective Examination of Demand-Side Energy Efficiency Policies*. Report RFF DP 04-19 REV, Washington, DC: Resources for the Future.

Gynther, L., I. Mikkonen, and A. Smits. 2010. Evaluation of European Energy Behavioural Change Programmers," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm.

Horowitz, M. 2007. "Changes in Electricity Demand in the U.S. from the 1970s to 2003," *The Energy Journal* 28(3): 93-119.

Horowitz, M. 2010. "Measuring the Savings from Energy Efficiency Policies: A Step Beyond Program Evaluation," *Energy Efficiency* 4(1): 43-56.

Integral Analytics, SageView Associates, and TecMarket Works. 2011. *Evaluation of the Home Energy Comparison Report*, Oregon, WI: TecMarket Works.

Keating, K. 1991. "Persistence of Energy Savings," in E. Hirst and J. Reed (Eds.), *Evaluation Handbook.* Report ORNL/CON-33, Chapter 6, pp. 89-99. Oak Ridge, TN: Oak Ridge National Laboratory.

Khan, J., M. Harmelink, R. Harmsen, W. Irrek and N. Labanca. 2007. *From Theory Based Policy Evaluation to SMART Policy Design: Summary Report of the AID-EE Project*. Utrecht, the Netherlands: ECOFYS Netherlands bv. Available at: http://www.aid-ee.org/documents.htm.

Kiss, B., K. McCormick, L. Neij, and L. Mundaca. 2010. "Policy Instruments for Energy Efficiency in Buildings: Experiences and Lessons from the Nordic Countries," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm.

Lapillonne, B., D. Bosseboeuf, and S. Thomas. 2009. *Top-down Evaluation Methods of Energy Savings: A Summary Report*, Grenoble, Switzerland: Enerdata; Paris, France: ADEME; and Wuppertal, Germany: Wuppertal Institute.

Lin, J., B. Cao, S. Cui, W. Wang, and X. Bai. 2010. "Evaluating the Effectiveness of Urban Energy Conservation and GHG Mitigation Measures: The Case of Xiamen City, China," *Energy Policy* 38(9): 5123-5132.

Mundaca, L. and L. Neij. 2010. "A Meta-Analysis of Bottom-Up Ex-Ante Energy Efficiency Policy Evaluation Studies," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm.

Nadel, S. 1993. *The Takeback Effect: Fact or Fiction?* Washington, DC. American Council for an Energy-Efficient Economy.

Nadel, S. 2011. "Our Perspective on the 'Rebound Effect' – Is It True That the More Efficient a Product Becomes, the More Its Owner Will Use It?" January 12. Available at: http://www.aceee.org/blog/2011/01/our-perspective-rebound-effect-it-true-more-efficient-pro.

Owen, D. 2010. "The Efficiency Dilemma," *The New Yorker*, Dec. 20.

Proctor Engineering Group. 1999. *Summary Report of Persistence Studies: Assessment of Technical Degradation Factors, Final Report*. San Rafael, CA: Proctor Engineering Group.

Ridge, R., P. Jacobs, H. Tress, N. Hall and B. Evans. 2011. "One Solution to Capturing the Benefits of Early Replacement: Approximately Correct Is Good Enough," *Proceedings of the 2011 International Energy Program Evaluation Conference*, Boston, MA: IEPEC.

Skumatz, L., M. Khawaya, and J. Colby. 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*, available at: http://uc-ciee.org/energyeff/energyeff.html.

Thomas, S., P. Boonekamp, H. Vreuls, J. Broc, D. Bosseboeuf, B. Lapillonne, and N. Labanca. 2010. "How to Measure the Overall Energy Savings Linked to Policies and Energy Services at the National Level," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm

Togeby, M., K Dyhr-Mikkelsen, A. Larsen, and P. Bach. 1010. "Portfolio Evaluation and Its Impact on Energy Efficiency Policy," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm

Vine, E. 2008. "Strategies and Policies for Improving Energy Efficiency Programs: Closing the Loop Between Evaluation and Implementation," *Energy Policy* 36(10): 3872-3881.

Vine, E., P. du Pont, and P. Waide. 2001. "Evaluating the Impact of Appliance Efficiency Labeling Programs and Standards: Process, Impact, and Market Transformation Evaluations," *Energy - The International Journal*, 26 (11): 1041-1059.

Vine, E., N. Hall, K. Keating. M. Kushler and R. Prahl. 2010. "Emerging Issues in the Evaluation of Energy Efficiency Programs," *Proceedings of the 2010 International Energy Program Evaluation Conference*, Paris, France: IEPEC. Available at: http://www.iepec.org/2010PapersTOC/2010TOC.htm.

Vine, E., M. Sullivan, L. Lutzenhiser, C. Blumstein, and B. Miller. 2011. "Experimentation and the Evaluation of Energy Efficiency Programs: Will the Twain Meet?" *Proceedings of the 2011 International Energy Program Evaluation Conference*. Boston, MA: IEPEC.