Perspectives on Program Influence and Cost Effectiveness: Moving Forward from the Recent US Debates

Michael W. Rufo, Itron Inc. Oakland, CA

ABSTRACT

Over the past decade or so, energy efficiency goals and expenditures in the United States increased dramatically from already significant activity levels. While efficiency accomplishments increased during this period of increased goals and funding, there were some significant disagreements over the magnitude of the accomplishments and the methods used to assess them, both in terms of evaluation approaches and cost effectiveness. This paper examines key issues and lessons learned from recent analyses of evaluation and cost effectiveness of efficiency programs in the US. Issues addressed include net impacts, short-term versus long-term impacts, cost effectiveness tests, incremental measure costs, and *non-energy* benefits and costs.

The paper reviews recent literature on program attribution methods and cost effectiveness tests and summarizes the range of perspectives. The paper highlights areas of general agreement and those of contention. For the key areas of disagreement and contention, the paper concludes by providing recommendations for orienting evaluation and cost effectiveness analyses to better serve multiple information, planning, and policy needs.

Introduction

Over the past decade or so, energy efficiency (EE) goals and expenditures in the United States (US) increased dramatically from already significant activity levels. These goals and expenditures were primarily associated with *voluntary* programs administered by electric and gas utilities, non-profits, and government agencies. More stringent state and federal efficiency-related codes and standards also played a significant role in energy efficiency policy during this time period. Greater energy efficiency activity led to increased involvement, attention, and expectations from regulatory and government agencies, as well as stakeholders such as consumer and industry groups, as well as private, third-party program implementers (Rufo, et al., 2008). While efficiency accomplishments certainly increased during this period of higher goals and funding, there have been some significant differences in perspective over the actual magnitude of the accomplishments and the methods and metrics used to assess them, in particular, in terms of evaluation and cost effectiveness approaches and priorities. The purpose of this paper is to summarize and offer suggestions for moving forward from recent evaluation and cost effectiveness debates in the US within the context of the European Union (EU) moving to increase its energy efficiency activities. EU reporting requirements and evaluation approaches are evolving and vary by member state as documented in several recent sources (ECEEE, 2013; Stern and Vantzis, 2014).

Current Evaluation and Cost-Effectiveness Issues and Debates in the US

In the following sections, we summarize and discuss several topics at the heart of current debates around energy efficiency program evaluation and resource valuation. Issues addressed include net impacts, treatment of spillover and market effects, short-term versus long-term impacts, incremental measure costs, non-energy benefits and costs, and cost effectiveness tests.

Attribution

Overview. Perhaps the single most contentious issue in the realm of voluntary energy efficiency programs has to do with program attribution, that is, the extent to which adoption of EE measures and associated impacts from those adoptions are attributable to a program or portfolio of programs. In the US, the attribution issue is usually discussed under the rubrics of "net-to-gross (NTG) ratios" and "free ridership." The NTG ratio refers to the fraction of program-claimed adoptions and impacts that are attributable to the program intervention. Total impacts associated with the efficiency actions of program participants are typically referred to as "gross" impacts whereas impacts estimated to be attributable to the program are referred to as net impacts. In simplest terms, gross impacts multiplied by the NTG ratio equal net impacts. "Free riders" in this context refers to program participants that would have adopted the efficiency measures anyway, that is, if there had been no program. Free ridership is often considered an element over which a program administrator has little control. However, the degree of control over free ridership likely varies across different program types, measure types, incentive strategies, and other factors (Rufo 2009).

Evaluators, planners, and regulators used the NTG and free ridership framework without much controversy from the very beginning of voluntary energy efficiency programs dating to around the early 1980s. The basic tenet was widely accepted as virtually commonsensical, namely, that the purpose of a program was to change outcomes from what they would otherwise be, and that no program could be assumed to be effective *ipso facto*, but, rather, that such effects should be based on objective measurement and analysis.

Although challenges to the free ridership and NTG concepts were relatively limited in the 1980s and early 1990s; that is not to say that there were no concerns. As with any social science program evaluation, energy efficiency evaluators were quickly faced with the difficulty of trying to determine the "counterfactual," that which would have occurred in the absence of the program. It is a daunting challenge to try to *measure* something that does not exist.[1]

Fortunately, efficiency program evaluators were not alone in this challenge and quickly borrowed and further developed methods from other disciplines to estimate net impacts. These included experimental and quasi-experimental design approaches, discrete choice techniques, and a variety of survey methods, among others. Confidence in the results from these early efficiency program evaluations may have been higher than it has been over the last decade or so. There are several possible reasons for this. First, in the first fifteen years or so of voluntary programs in the US (i.e., the late 1970s until about the mid-1990s) the efficiency market was relatively immature in almost all of its key dimensions. Efficiency-related markets during this first phase of programs were generally characterized by moderate levels of efficiency adoption, low levels of end user and supply-side[2] awareness of efficiency options, low levels of interest in efficiency as compared to other equipment- and building-related characteristics, and slow rates of change in efficiency product features and availability. Said another way, the baseline conditions during this initial period were relatively stable and low with respect to efficiency-related change. Second, efficiency programs did not grow at the same rate across the country or even within states, being clustered in regions instead, which left the resulting areas that

[1] Although the literature is full of references to "measuring" free ridership and net impacts, it is this author's opinion that "estimating" is a more appropriate term since the counterfactual cannot be directly observed. Instead, indirect measurement techniques are used to make estimates of free ridership and net impacts.

[2] "Supply side" as used here refers to the supply-side of the energy efficiency and related equipment and services markets, that is, inclusive of retailers, contractors, distributors, manufacturers and specialty providers such as energy service companies (ESCOs).

did not have programs as potential control groups in quasi-experimental research design methods. Third, program goals were relatively modest, or did not exist as aggregate targets per se, and time was not necessarily considered to be of the essence in achieving savings targets, which made for a friendlier environment for pilots and the use of experimental design. Fourth, programs had not been in existence for very much time, thus evaluations could focus on near term program impacts without much concern over whether the impacts they were observing were from the program in question or from the cumulative effect of similar programs over many years.

As a result of these early conditions, in particular, the low levels of efficiency adoption in areas without programs, and the large percentage of non-participants where programs did exist, it was relatively easy during this period to make the case that high levels of efficiency adoption that occurred along with program participation were attributable to those same programs. In contrast, for the past ten years or so, the relative opposite has been true, namely:

- a number of first generation efficiency measures have gone through a full product life cycle and reached high levels of market saturation;
- end users and supply-side actors have a much greater awareness of, and interest in, buying and selling efficient products and services;
- new products and services are developed at a much faster rate than ever before;
- programs have existed at one time or another in many, if not most, jurisdictions;
- market transformation of a number of efficiency measures (e.g., linear fluorescent and compact fluorescent lamps) has influenced virtually all US markets (regardless of local program activity levels); and
- program administrators have been under significant time pressure to ramp up programs and meet aggregate goals quickly.

In addition, there are many more programs, policies, and information campaigns (e.g., tax credits, carbon trading, corporate sustainability goals, etc.) that are outside of the typical rebate and information programs of local program administrators than ever before, making attribution to individual programs even more difficult. Finally, the existence of many programs for many years and, in some cases, decades, has made it difficult to discern near-term from long-term effects. For example, in some circumstances, there is the possibility of programs producing long-term market effects, which, if not properly accounted for, could contribute to a decrease in *observed* short-term program effects.[3] This adds to the difficulty in today's environment to provide convincing estimates of net program impacts (Friedmann, 2011).

Initially, most evaluation studies were focused exclusively on measuring the direct effects of a specific program (Rufo and Bester, 1991). However, it was not long before analysts took a broader view of the potential program impacts, not only on program participants, but also on non-participants and on the entire market for a particular efficiency-related product or service (Eto, Prahl, Schlegel, 1996). Additional program impacts not captured by the initial net-of-free-riders (NOFR) framework included *program-induced* efficiency actions taken by participants for measures not included or claimed directly in the program (referred to generally as "participant spillover"), *program-induced* efficiency adoptions from non-participants ("non-participant" spillover), and program-induced changes in market characteristics that are causally related to efficiency adoption (e.g., reductions in energy efficiency incremental measure costs,

[3] For example, a program run for a long period of time might lead to market effects that cause ongoing and self-sustaining increases in efficiency adoption (that is, market transformation). These changes could include effects like reduction in the incremental costs of efficiency measures, increases in consumer awareness of and satisfaction with key efficiency measures, improvements in efficiency product features, incorporation of efficient equipment purchasing into corporate policies, etc.

increases in high efficiency product availability, increases in awareness of efficient products, etc.), which are often referred to as "market effects". Consequently, the operating definition of the "net-to-gross" ratio was expanded by many practitioners and jurisdictions to include spillover and market effects.[4]

Starting around the mid-1990s a new wave of studies was launched to estimate these additional impacts separately. While there have been a number of very good and informative studies of spillover and market effects conducted over the years, the studies have been challenged by the same kinds of difficulties associated with free ridership estimation *but even more so* due to the fact that these additional impacts occur, by definition, outside of the direct program participation experience (Prahl, 2013). Thus, spillover and market effects attribution tends to be even more difficult than evaluation focused on the direct program-related actions of program participants. Different jurisdictions have taken different approaches to dealing with these challenges at different points in time. In some jurisdictions, NTG continued to be purposefully defined for most regulatory purposes (e.g., attainment of efficiency goals, cost-effectiveness analysis, and shareholder/performance incentives) to exclude spillover and market effects, that is, NTG = NOFR. In other jurisdictions, spillover and market effects have been included in NTG but with sometimes strong disagreements over the magnitude and reliability of these estimates. The debate on whether the research on spillover and market effects is sufficiently robust and quantifiable to incorporate into official metrics of program performance continues. Some jurisdictions, like California, have tried to strike a balance by adopting a modest NTG "adder" for spillover and market effects, at an overall portfolio level, based on an assessment of multiple spillover and market effects studies and consideration of the particular characteristics of their programs.

To Stay or Change the Course? Given the significant change in the underlying landscape of energy efficiency programs and markets, it should not be too surprising that practitioners, program administrators, regulators, policy makers, third party implementers, and other stakeholders have been engaged in serious, and often unsatisfactory, debates about the relative accuracy, viability, and even existential importance of estimating net impacts. Over the past five to 10 years, the debate in the US has featured calls to substantially modify or abandon estimation of free ridership and NTG altogether. Some analysts have expressed strong concerns in several areas, generally focusing on their belief that: 1) estimating free ridership is intractable, 2) that current methods and applications produce overestimates of free ridership, and 3) that these overestimates and related overconfidence in non-program efficiency drivers lead to pre-mature abandonment of efficiency programs (Peters and McRae, 2008; Mahone, 2011; and Friedmann, 2011). The papers referenced provide useful perspectives and good discussions of the authors' concerns and readers are encouraged to review them in their entirety. Rather than restate these authors' positions in detail (due to page constraints), a few of their most significant criticisms are taken up with the goal of informing recommendations at the end of this paper.

To begin with, NTG Concern 1 above is a mix of both a legitimate concern at the vexing nature of counterfactual-related evaluation and, in this author's view, an unwarranted overreaction given that few critics in this camp have put forward a meaningful *alternative* to assessing program effectiveness. Instead, they seem to leave one with the suggestion that *no* program influence evaluation is preferable to difficult program influence evaluation. We return to this point below.

With respect to Concern 2 above, Peters and McRae, 2008 make some important points based on findings from broader social science research (e.g., social desirability bias and cognitive dissonance theory) that may, in some cases, lead to *upward* bias in free ridership from application of self-report methods to

[4] Since early net impacts studies generally focused only on estimation of free riders, net impacts were gross impacts minus impacts from free riders and NTG = Net of Free Riders (NOFR). By the mid-1990s, more and more studies and policies defined NTG as inclusive of spillover, e.g., NTG = NOFR + Spillover + Market Effects.

individual, residential end users. Their criticism is weaker, in this author's view, with respect to non-residential end user decision making and includes insufficient acknowledgment of the potential *downward* bias in free ridership due to program participants' ascribing influence to programs because they would like to see incentive programs continued to enable their future use of such funds.

Mahone also argues that some NTG methods can result in a predictable, systematic upward bias in free ridership (downward bias in NTG). Mahone stops short of advocating a pullback in NTG studies and focuses instead on trying to illustrate where potential overestimation of FR may occur. He provides a number of recommendations for policy makers and evaluators to consider when conducting and assessing the results from NTG studies. In one of his examples, Mahone indicates that if a non-participant "control" group (in a *quasi*-experimental not *pure* experimental design application) is used to estimate FR in a program group, and if the non-participant group has made some efficiency adoptions as a result of spillover from the program activity, then, if not measured and properly accounted for, the result will be an overestimate of free ridership in the program group. This concern is legitimate with respect to contemporaneous[5] spillover and has contributed to limitations in the use of within jurisdiction non-participants as control groups for NTG studies.

The issue Mahone raises with respect to spillover and market effects that result from the cumulative effects of programs *prior* to the program cycle being evaluated is more complex.[6] In the latter case, the long-term, non-participant spillover attributable to program-induced market effects will downwardly bias an attempted estimate at long-term program effects if the current program participants were being used as the basis for such an estimate (which would be an incorrect design for a long-term spillover and market effects study). However, these long term effects would not bias an estimate of the effect of the *current* program on the *current* participants. That is to say that there are at least two indicators of program influence for programs that have been around in a similar form for multiple years: long term and short term. Even when programs have long-term effects, it is still useful to assess whether the current, short-term effects are still significant on the margin. In this way, the NOFR metric is a kind of "what have you done for me lately" estimate which, though not necessarily the whole story, is an important part of the story and critical for ongoing improvement and assessment of the program. At the same time, it is very important that policy makers and those responsible for funding and authorization decisions understand that some of those estimated to be free riders in today's programs may be spillover from the effects of past programs.

Peters and McRae, along with Friedmann, recommend replacement of traditional free ridership and NTG evaluation methods in favor of measurement of efficiency adoptions and impacts in the population of interest over time. While longitudinal measurement of total efficiency improvement from all influences (i.e., inclusive of voluntary programs, codes and standards, other government programs, and market forces) is a critically important part of the overall research agenda, it is hard to see how this approach provides the necessary and adequate short- and mid-term feedback on individual program effectiveness essential to program improvement. Whole market assessments of "all in"[7] efficiency improvements should be viewed as a complementary, but not substitutable research activity. Program evaluations provide program specific

[5] By contemporaneous spillover we mean non-participant spillover that occurs as the result of the program cycle that is being evaluated. This is in contrast to non-participant spillover, or market effects, that are associated with the cumulative effects of indirect program influence from prior, multiple program cycles.

[6] And these longer term effects are likely to be much larger than contemporaneous spillover as they are typically due to sometimes powerful market effects such as decreases in the incremental costs of efficiency measures, increases in efficiency product availability, improvements in efficiency product and services features, and the like.

[7] Efficiency improvements occurring in the entire market, not just among program participants, that is, due to *all* influences, both program and non-program, on total market adoption of high efficiency products and practices.

feedback that is vital to continuous program improvement. Long term studies of efficiency in aggregate are critical to assessments of the cumulative effects of all policies, programs, and market forces combined, but typically have less ability to differentiate the degree of influence of each element and provide limited feedback on specific program designs.

Peters and McRae, as well as Friedmann, appear to argue that FR and NTG studies are so far off track that they do more harm than good and should essentially cease. In our view, that would be a very unfortunate outcome of this line of criticism, opening the door for every program implementer and advocate to be able to merely assert that a particular program is maximally effective and worthy of funding. Removing independent assessment of program influence would decrease the effectiveness of efficiency portfolios by limiting program adaptation, hindering program-related assessment needed for portfolio optimization, and diminishing the credibility of reported impacts. Measurement challenges should not lead to measurement avoidance. Instead, evaluation, planning, and policy must continue to evolve to better address short-term and long-term issues associated with estimating program attribution.

NTG-related Concern 3 is that the research is net biased toward underestimating NTG and that this leads to pre-mature termination of programs. This is the primary concern expressed in Mahone's paper; however, no examples of where and when this has occurred are provided. Peters and McRae, as well as Friedman, also express the concern that NTG studies as practiced in the US can lead to programs being prematurely terminated, but again, no specific case studies are provided. In fact, it can be argued that the opposite is just as often true; namely, that programs, or funding for certain measures, are continued, or go unmodified, for long after NTG research indicates they are only marginally effective. For example, research strongly indicated that first-generation programmable thermostats were standard practice with new furnaces in California, yet incentives for this equipment were continued for many years afterward.

In a similar vein, Peters and McRae state that "program design occurs with a view to minimize a number – free riders" and Friedmann states that NTG inhibits innovation. However, a causal link - from the supposed stifling effects of a biased evaluation framework to limited program innovation - is not convincingly made. There is equal, if not stronger, anecdotal evidence that many programs seem to operate in a vacuum without regard to maximization of program influence and, at times, with indifference to free ridership. It is unclear why evaluation would so assuredly inhibit innovation, as asserted by these authors, when improving program effectiveness is the goal of such research. Program designers and implementers have a variety of actions at their disposal in response to findings of limited or decreasing program influence. They can shift program targeting to more efficient measures and different market segments, they can change incentive structures, they can adapt their marketing and outreach approaches, and they can use pilots and controlled trials to test new approaches, among other actions. In addition, when program innovation does occur, and is successful, it is more likely to manifest in higher rates of efficiency adoption, even under imperfect measurement approaches. Finally, in the debate over the limitations and potential biases of methods to estimate free ridership and NTG, there seems to be an under valuing of the broader purpose of NTG studies beyond "the number"; for example, the provision of findings and recommendations for improving program features and outcomes.

Nonetheless, the recent debate over FR and NTG has been healthy and useful to all concerned in challenging assumptions and methods in a difficult area with important outcomes at stake. The debate is also important because it has sought to bring important concerns and caveats to policy makers and regulators with the goal of broadening their perspectives beyond basing decisions solely on the numeric outcomes of NTG studies. Concern over ever encroaching climate change has added additional pressure, underpinning the view that it is more important to take aggressive action than it is to overly parse and assess that action for the purpose of incremental improvement (i.e., via NTG studies). This point of view can be argued both ways though, with the counter view that, in a climate- and economically-challenged world, each dollar of investment has an opportunity cost and must be put to its fullest, most effective use to reduce GHGs.

Cost Effectiveness Tests

Overview. Cost effectiveness (C-E) tests, also often referred to as benefit-cost (B-C) tests, have been a formal, foundational part of energy efficiency program planning, implementation, evaluation, and policy making from the earliest days of programs. C-E analysis has been a critical component of the justification for energy efficiency programs in the US for 30 years or more. From the beginning, it was argued, using C-E tests, that energy efficiency was more a cost effective investment than alternative power generation investments.

Early practitioners sought to draw on the literature and lessons learned from cost effectiveness analyses developed for social programs (e.g., welfare economics) and from financial analyses used for utility resource planning. One of the earliest and most influential frameworks developed was the California Standard Practice Manual (C-SPM), which was first released in the early 1980s (C-SPM, 2001). A number of other C-E manuals and guidebooks have been published in the US since the initial C-SPM. These other references generally seek to provide refinements, clarifications, and perspectives on the various tests, as well as particular jurisdictions decisions on how to use the tests for their own regulatory, planning, and evaluation purposes. The C-SPM sought to address a wide range of perspectives on energy efficiency (and, more broadly, demand-side management) programs through the use of multiple tests, each of which characterized a particular stakeholder point of view. Five perspectives were characterized in the tests shown in Figure 1.

Figure 1. High-Level Summary of Energy Efficiency Cost Effectiveness Tests

Test	Benefits	Costs
Total Resource Cost Test (TRC)	Generation, transmission and distribution savings Environmental externalities	Generation, transmission and distribution Program costs paid by the administrator Participant incremental measure costs
Participant Cost Test (PCT)	Bill reductions Incentives Non-energy benefits	Bill increases Participant incremental measure costs Non-energy costs
Program Administrator Cost Test (PAC)	Generation, transmission and distribution savings	Generation, transmission and distribution Program costs paid by the administrator Incentives
Ratepayer Impact Measure Test (RIM)	Generation, transmission and distribution savings Revenue gain	Generation, transmission and distribution Revenue loss Program costs paid by the administrator Incentives
Societal Cost Test	Generation, transmission and distribution savings Environmental externalities Non-energy benefits Participants avoided equipment costs (fuel switching only)	Generation, transmission and distribution Program costs paid by the administrator Participant incremental measure costs Non-energy costs

Given the range of different perspectives, this was a reasonable approach that has held up fairly well over time. Although several of the tests have had their fair share of criticism, the range of tests, and breadth of what they cover *collectively*, has been shown to be relatively robust. There was some effort in the early 2000s to develop an additional test, the Public Purpose Test (PPT). This test was explored for application to market transformation types of programs. Although the test did not gain traction, a thorough and useful framework for cost effectiveness analyses and market transformation resulted from the effort (Sebold 2001).

A significant portion of the debate in the US over the various C-E tests has been focused on: 1) jurisdictional decisions to prioritize or exclusively use a single, particular test, 2) the accuracy of the inputs

to the tests, and 3) the appropriateness of the implementation of tests. To be sure, there has also been some criticism of the theoretical and analytical structure of the tests themselves; however, in this author's opinion, much of the structural criticism has been shown to be a form of one of the three issues listed above. Given the extensive literature and high awareness of the tests, as well as the page limitations of this paper, rather than further discuss the structure and intentions of the tests, we will move directly to the major debates and points of view that have been expressed by various stakeholders and authors.

The Cost Effectiveness Debate – Then and Now. There have been two periods of "great" debate over energy efficiency cost effectiveness tests. The first period corresponded with the wave of demand-side management programs that began in the early 1990s and came to significant decline due to electric industry restructuring in the late 1990s. During this period the debate pitted free market economists opposed to intervention in demand-side energy markets against economists and advocates in support of efficiency program interventions. This was largely a debate focused on the RIM test (favored by free market advocates and utilities strongly preferential to increasing power generation) versus the TRC test (favored by efficiency advocates). Though each side had its jurisdictional supporters, the pro-intervention side was fairly successful in convincing the majority of jurisdictions to use the TRC and to initiate and increase funding for EE programs based on the argument and evidence that cost effective efficiency opportunities were being left on the table by the market. Some jurisdictions did adopt the RIM test as primary (generally in the southern states, including Florida). During this time, efficiency supporters put significant effort into promoting the TRC test as superior to the RIM test. Some efficiency supporters also advocated for the societal test at this time, arguing that focusing only on direct energy and capacity avoided costs was too narrow a frame and that the environmental and energy security benefits of EE were significant and should be reflected in C-E results through incorporation of environmental externality adders to the TRC or via a full blown societal test. The TRC gathered significant support as a reduced scope societal test that incorporated a blend of both the utility and participant's perspectives and, in some jurisdictions, inclusion of some environmental externalities.

In the second period of debate, which began five or so years ago, efficiency advocates are now somewhat divided among those who continue to support the TRC and those that now recommend the PAC (Neme and Kushler, 2010). In addition, some analysts have renewed the argument for a move to a full societal test (Woolf, et al., 2012). The RIM test has not figured much in the latest round of debates except that some jurisdictions are beginning to focus more on the potential rate impacts of efficiency programs than they had previously (Haeri and Khawaja, 2013). Only one state of 41 included in a recent US survey used RIM as the primary cost effectiveness test (ACEEE, 2013).

Haeri and Khawaja (2013) trace the renewed interest in assessing the appropriateness of the various cost-effectiveness tests to recent perceptions that program cost-effectiveness ratios are decreasing in the US in some key markets. Ironically, in this author's opinion, these decreases in C-E ratios are partially attributable to the success of long running EE programs and codes and standards, due to the fact that resulting increases in efficiency levels (and correspondingly lower levels of end use consumption) have made it more difficult to capture cost effective savings in some key measure groups and market segments.[8] In some jurisdictions, C-E ratios have also declined due to decreases in gas prices over the decade.

There are two related thrusts in the current debate, one that has focused on criticisms of the TRC and the other that has focused on promotion of the PAC to replace the TRC as the primary test to be utilized for program screening and value assessment. A number of authors have presented a variety of TRC critiques, several of which focus on inputs that could cut across multiple tests, these include: 1) setting the discount rate lower than the traditional utility cost of capital benchmark (principally to account for lower risk and

[8] One indicator of this change can be seen in California where overall, portfolio wide, TRC ratios tended to be in the range of 2.5 in the 1990s but are now on the order of 1.0 – 1.5.

greater environmental benefits of EE resources), 2) removing the 20-year cap on the length of the effective useful life (EUL) of efficiency measures (in some jurisdictions), 3) including non-energy benefits (NEBs), 4) improving the estimation of incremental measure costs and excluding costs that are not EE related, and 5) questioning the relevance of the incremental cost element of the TRC.

The first critique would apply to all tests that use the utility cost of capital, instead of a social discount rate, to present value future benefits from avoided energy, capacity, distribution, and environment externality costs (NRC, 2010; Hall, et al, 2008). Use of very low or negative social discount rates has been proposed since the inception of EE programs; however, it has not gained much traction until recently. During the early years of EE programs, energy and capacity were the key drivers for cost effectiveness and there were limited benefits of avoided environmental externality costs included in TRC analyses. Given the growing challenge and concern over climate change and its associated costs, there may be increased receptivity to reducing discount rates based on the strong case that can be made that future emissions reductions should be as valuable, if not more valuable, then current reductions, and the related argument for intergenerational equity with respect to the environmental impacts of energy production and consumption. One compromise approach could be to continue to discount the capital and direct energy costs using cost of capital proxies but utilize a low discount rate for the environmental externality costs of carbon.

Criticism 2 above is, in this author's opinion, reasonable when justified on a case specific basis. The 20 year cap for EULs that is common practice in the US is likely a historic artifact of early practice (circa 1980s/early 1990s) in which the combination of high discount rates,[9] limited computational tools,[10] and high early program C-E results, led to a de facto capping of EULs at 20 years in some jurisdictions. Computational limits are certainly no longer a factor and there is strong empirical evidence that some measures' average EULs should be longer than 20 years (e.g., windows and other long lasting structural measures). Coupled with the importance of reducing greenhouse gas emissions well into the future, removing the 20 cap in favor of case by case EUL determination makes sense.

Criticism 3, regarding concern over exclusion of non-energy benefits, is complex and one's position on it tends to follow from whether one views the TRC as an energy resource planning test or as a stripped down societal test. Some critics argue that non-energy benefits such as noise reduction (e.g., via double pane windows), comfort (e.g., via reduced infiltration), productivity (e.g., via improved lighting quality), and the like, should be included in the TRC. Opponents argue that many of those kinds of non-energy benefits are private benefits that occur *only* to the participant and that many such benefits are unrelated to the energy system or even, in some cases, to any "public" good that warrants programmatic support. In addition, this author notes that it is curious and seemingly asymmetric that proponents of NEBs rarely mention non-energy *costs* (NECs),[11] though some jurisdictions do make the distinction and focus on the integration of the two with the result framed as non-energy *impacts* (NEIs). An area where these two camps have reached some agreement is with respect to the idea that, instead of focusing on including participants' non-energy *benefits*

[9] When discount rates are high, the NPV benefits of extending EULs beyond 20 years are minimal, under lower discount rates, they become more material.

[10] Early computers, upon which the first integrated resource planning models were built, had limited computational power. Running programs out 20 years, and extending each program year out to a maximum 20 years of EUL benefits, required a 40 year computational framework. This was the limit of most models and considered adequate at the time.

[11] Non-energy costs are very related to market transformation barriers (see Eto, Prahl, Schlegel, 1996) and can include a variety of outcomes related to efficiency measures that result in poor performance across characteristics of importance to end users, for example, poor light quality and on/off responsiveness from early CFLs, decreased indoor air quality if a home is too tight, higher perceived operations and maintenance costs, lower perceived reliability, etc. In fact, one could argue that, for measures with short energy paybacks, it is when perceived NECs exceed NEBs (e.g., net negative NEIs) that program interventions are justified, i.e., when market barriers are high.

in the TRC, one can more easily focus on ensuring that the participants' *incremental* costs used in the TRC test include only those costs associated with the incremental efficiency of the product. That is, ensuring that any higher costs for product features that provide NEBs but not energy savings are excluded from the TRC.[12] Some would say that this has always been the intended definition of incremental costs as applied to use in the TRC test. There are some cases, however, where full costs are used in the TRC when incremental costs are more correct. This may be due to the fact that there has been much less research on estimating incremental efficiency measure costs as compared to estimating savings (Itron 2014).

Criticism 4 is essentially captured by this last point, namely, that costs associated with non-energy benefits should be excluded from the cost side of the equation.

Criticism 5 focuses on arguments put forth that the participants' costs should be considered irrelevant to C-E screening for energy efficiency and that the TRC should be replaced as the primary test by the PAC test. This advocacy is driven by two lines of thinking. One being that, if NEBs are excluded, then participant costs should be excluded as well, as excluding those costs is easier than trying to estimate a pure incremental cost exclusive of NEBs. This view also holds that estimating incremental costs is difficult in practice. An alternative view is that incremental costs are actually easier to estimate than energy impacts for efficiency since there is not a counterfactual problem. The challenge of incremental costs estimation in this author's view is more one of lack of effort, experience, and expertise. It is likely that 100 times more resources have been devoted to estimating energy impacts than incremental costs during the first three decades of efficiency programs in the US. There are a number of measure cost studies that provide strong evidence that incremental costs can be reliably estimated even with budgets far below those set for savings estimation (Itron 2014; Ting, 2013).

The second line of thinking is that energy efficiency should be treated identically to power generation resources in utility financial decision making. For example, in less regulated or deregulated wholesale markets in the US, electric and gas utilities procure power at a market price and this market price may or may not be related to the costs of generating the power by the provider. The purchasing utility is not concerned with the generators' costs; they are only concerned with the price at which they are acquiring the power. In this argument, PAC proponents maintain that, similarly, utilities procuring energy efficiency resources need not be concerned with the costs paid by the end user for their efficiency measures. While this may be attractive in its simplicity and apparent evenhandedness, it goes against the broader historic *raison d'etre* for energy efficiency programs in the context of energy resource planning. Efficiency programs are often justified as part of energy planning processes that purposefully take into account a wide variety of non-price resource characteristics, including environmental externality costs and resource diversification.

In other words, modern resource planning has not been based exclusively on utility out of pocket payments, it has taken into account a variety of characteristics and total costs of resources. The same should hold true within energy efficiency cost effectiveness analysis. Incremental costs are an important piece of information that should be used to help prioritize among options. In addition, the TRC tends to be much more closely correlated to the Participant Test than is the PAC test, and provides more insight into the level of program incentives, if any, a particular measure may require. Incremental costs can also be a first order proxy indicator for embedded energy, which is rarely, if ever, addressed in efficiency analyses in the US.

Insofar as energy efficiency programs are, by definition, market interventions they should be informed by a broad set of considerations and measure characteristics, including incremental cost. For

[12] For example, if a home owner chooses to replace the windows in their home because they are old, leaky, do not adequately reduce noise, and do not provide enough energy savings, and an efficiency program provides a financial incentive that covers less than the full cost of the replacement, it would not be appropriate to use the entire cost of the window replacement to assess the cost effectiveness of the program. Instead, the cost associated with the home owner's willingness to pay a portion of the cost due to their desire to acquire the non-energy benefits should be removed from the full cost.

example, end users might love a particular measure (Measure A) with a high incremental cost based on NEBs and be willing to adopt it with a modest rebate; however, the savings in aggregate, and TRC, might be low as compared to another measure (Measure B) with lower incremental costs, higher aggregate savings, and a higher TRC. Consequently, marginal program dollars would likely be better spent on Measure B. In addition, the incremental cost information for Measure A and B is useful to calculating the Participant Test, customer payback, and ROI, all of which contribute to estimation of likely free ridership and for assessments of the relative importance of cost versus non-cost market barriers that warrant intervention. A PAC-only view can mask situations in which an incentive is a very low or insignificant share of the participant's cost, a circumstance that sometimes, though not always, is indicative of low program influence (high free ridership).

In summary, for jurisdictions that want to rely solely on one test, the weight of the evidence and argumentation continues to point to the TRC, inclusive of a *full accounting* of energy-based environmental externalities (especially GHG costs), as the most appropriate test for screening and prioritizing energy efficiency programs. That said, the TRC should not be a hard cap for all programs at all times, nor should it be used exclusively. The original intent of the C-SPM was to use multiple tests to shed light on resources from multiple points of view. For market transformation (MT) programs in particular, the TRC should not be a hard cap since, by definition, the intent of MT efforts is to address barriers, including cost, that inhibit EE adoption in the short term but that have the potential to become very cost effective in the mid- and long-terms. For MT programs, a combination of the TRC and PAC should be used, along with market transformation oriented tests like the PPT (Sebold 2001) when feasible.

Lessons Learned and Recommendations

Debate in the US over program attribution and cost effectiveness has had a long and sometimes rancorous history; however, the process and results have been healthy. The following recommendations are provided to provide additional perspective on moving forward from these debates to activities that will help ensure that critically important efficiency investments provide their maximum value.

Conduct program attribution studies to assess near-term *and* long term effects, and include the full range of indicators – free ridership, spillover and market effects. Despite the challenges to program attribution studies, the alternative of not evaluating program influence is decidedly worse. NTG studies must continuously improve and build off of critical feedback. Program attribution research should ideally go beyond net-of-free ridership (NOFR) and include spillover and market effects studies.

View NOFR as an indicator of *marginal* program efficacy, not long term program influence. The NOFR form of NTG, though lacking with respect to long-term market impacts, has a particular value with respect to assessment of the marginal efficacy of programs, even, and especially, for those programs that have been around for many years. When there is strong evidence that some portion of current free ridership is a result of prior years' efforts that should be communicated, even if it cannot be quantified.

Make clear that FR and NTG estimates are uncertain and provide appropriate caveats. Estimated NTGs are often used as inputs to C-E tests, estimation of goal achievement, and other quantitative policy applications. While such applications necessarily require quantification, they should not be presented in a way that overly focuses on the accuracy of point estimates out to the second or third decimal point. Evaluation studies and applications of NTG should educate users on the uncertainty of the estimates and provide caveats based on the specific limitations of the particular research conducted.

Frame NTG estimates as *directional* (e.g., high, moderately high, moderately low, and low) and use qualitative results of NTG and market research to assess program effectiveness and design enhancements. Framing NTG estimates in terms of their qualitative, directional results is likely to be more appropriate, useful, and influential than overly emphasizing point estimates. To address the uncertainty

associated with NTG estimates, scenario analyses should be used across a *reasonable* range of values, while remaining grounded in the evaluation work conducted.

Recognize that some degree of free ridership is in most cases unavoidable, do not expect complete elimination, or unreasonably lower levels, of free ridership. Program features and requirements, along with measure and market segment targeting, can be used to reduce free ridership; however, they are unlikely to avoid such participants entirely. Program participation can build good will with end users that are market leaders who expect to be able to draw on program funds to which they have contributed through their bill payments. Programs should make efforts to continue to push even these leaders to increasingly higher levels of efficiency rather than only settle for paying them for efficiency projects they would have done anyway.

Consider using a portfolio-level adder to account for spillover and market effects. Quantifying spillover and market effects is very difficult, especially for individual programs over the short term. Such studies can have large data requirements and long time horizons. There is evidence over a wide range of studies that, in aggregate, energy efficiency portfolios generate some degree of net positive impacts beyond those accounted for directly in programs. Rather than conducting extensive market effect studies over multiple years for every program, applying an adder for spillover and market effects to net-of-free ridership estimates can help to address potential downward bias with respect to longer term program effects. Adders can be developed based on careful review of existing spillover and market effects studies with appropriate adjustments for local program features and program size as compared to market size.

Conduct broader, longitudinal market saturation and characterization studies, to assess "all in" efficiency and market change. Don't let the challenges of program attribution get in the way of long-term market analysis. Comprehensive quantification of the saturation, efficiency, and consumption of key market sectors provides the basis for assessing progress toward overall energy policy and GHG reduction goals across policies and programs. Conducting such studies periodically (longitudinally) is necessary for "all in" long term assessment that program-specific NTG studies cannot provide on their own.

Calculate cost-effectiveness using a range of tests, placing the greatest weight or primacy on the TRC test (with environmental externality costs). A TRC test with well estimated *incremental* costs and a full accounting of environmental externalities is more informative for energy resource planning and management than a PAC test. Given their potential magnitude, it is especially important that avoided carbon costs be included with the TRC test. The TRC inclusive of avoided environmental externality costs captures much of the value of a societal test but is more easily implemented. That said, not all programs should be required to pass the TRC at all times. Some market transformation programs will not be cost-effective initially as they are designed to accelerate the market changes necessary to increase cost effectiveness over time. The PAC test provides an additional screen for market transformation programs that do not pass the TRC test. Even programs that do not pass the PAC test may warrant consideration based on how likely, how quickly, and for how long, they might become cost effective.

Assess and quantify non-energy impacts (NEIs) based on their relative import to end users and society. The relative import of NEIs can range from insignificant to primary. Rather than trying to quantify these precisely for the purpose of including in a C-E test, which is extremely difficult in practice, use NEI research to help develop and prioritize program interventions and to focus and inform measure cost research. If NEI research indicates that there are significant NECs that are inhibiting adoption, for example, lower quality end use service, or high operations and maintenance costs, then program activities should include elements to improve the service product quality and reduce O&M costs. If NEI research indicates that there are significant NEBs of a measure with significant incremental measure costs, additional research may be needed to tease out what portion of the incremental costs are associated with the increased efficiency of the measures as compared to non-energy features and attributes.

References

ACEEE. 2013. *A National Survey of State Policies and Practices for the Evaluation of Ratepayer-Funded Energy Efficiency Programs,* American Council for an Energy Efficiency Econony. Report Number U122. February.

C-SPM. 2001. *California Standard Practice Manual: Economic Analysis of Demand-Side Programs and Projects.* October, 2001.

ECEEE, 2013. "Understanding the Energy Efficiency Directive, Steering Through the Maze #6: A Guide from ECEEE". European Council for an Energy Efficient Economy.

Eto, J., R. Prahl, and J. Schlegel. 1996. *A Scoping Study on Energy-Efficiency Market Transformation*, Prepared for the California Demand-Side Measurment Advisory Committeee: Project 2091T.

Friedmann, R. 2011. "A Fresh Look at Evaluation to Support Energy Efficiency in the 21st Century." *Proceedings of the International Energy Program Evaluation Conference*, Boston, MA. August.

Haeri, H. and M. S. Khawaja, 2013. "Valuing Energy Efficiency: The Search for a Better Yardstick," *Public Utilities Fortnightly*, July: 28-36.

Hall, N., R. Ridge, G. Peach, S. Khawaja, J. Mapp, B. Smith, R. Morgan, P. Horowitz, G. Edgar, J. Luboff, and B. Evans. 2008. "Reaching our Energy Efficiency Potential and our Greenhouse Gas Objectives: Are Changes to our Policies and Cost Effectiveness Tests Needed?" *Proceedings 2008 AESP National Conference*, Association of Energy Services Professionals, Phoenix. AZ

Itron, Inc. 2014. *Energy Efficiency Measure Cost Study*, prepared by Itron Inc. for the California Public Utilities Commission. June.

Itron, Inc. 2008. *National Energy Efficiency Program Best Practices Study*, prepared by Itron Inc. for California's Investor-Owned Utilities and the California Public Utilities Commission. ***www.eebestpractices.com***

Mahone, D. 2011. "Free-Ridership as a Way to Kill Programs - How Evaluation Policies Can Frustrate Efficiency Goals." *Proceedings of the International Energy Program Evaluation Conference*, Boston, MA. August.

National Research Council Study (NRC). 2010. *Hidden Costs of Energy: Unpriced Consequences of Energy Production and Use*, Committee on Health, Environmental, Environmental, and Other External Costs and Benefits of Energy Productions and Consumption.

Neme, C. and M. Kushler. 2010. "Is it Time to Ditch the TRC? Examining Concerns with Current Practice in Benefit-Cost Analysis," *Proceedings of the American Council for an Energy Efficient Economy 2010 Summer Study on Energy Efficiency in Buildings*. Washington, D.C.: American Council for an Energy-Efficient Economy.

Peters, J. and M. McRae. 2008. "Free-Ridership Measurement Is Out of Sync with Program Logic…or, We've Got the Structure Built, but What's Its Foundation?" *Proceedings of the American Council for an Energy Efficient Economy 2008 Summer Study on Energy Efficiency in Buildings*. Washington, D.C. August.

Prahl, R., R. Ridge, N. Hall, and W. Saxonis. 2013. "The Estimation of Spillover: EM&V's Orphan Gets a Home." *Proceedings of the International Energy Program Evaluation Conference*, Chicago, IL. August.

Rufo, M. 2009. "Evaluation and Performance Incentives: Seeking Paths to (Relatively) Peaceful Coexistence." *Proceedings of the International Energy Program Evaluation Conference*, Portland, OR. August.

Rufo, M., M. Ting, M. Messenger, and M. Wheeler. 2008. "Energy Efficiency as the First Resource: Opportunities, Challenges, and Beating the Next Bust," *Proceedings of the American Council for an Energy Efficient Economy (ACEEE) 2008 Summer Study on Energy Efficiency in Buildings*. Washington, D.C. August.

Rufo, M. and N. Bester. 1989. "An Investigation of Commercial and Industrial Utility Demand Side Management Program Impacts." *Proceedings of the International Energy Program Evaluation Conference,* Chicago, IL. August.

Skumatz, L. A. and E. Vine. 2010. "A National Review of Best Practices and Issues in Attribution and Net-to-Gross: Results of the SERA/CIEE White Paper." *Proceedings of the American Council for an Energy Efficient Economy 2010 Summer Study on Energy Efficiency in Buildings*. Washington, D.C.: American Council for an Energy-Efficient Economy.

Sebold, F., A. Fields, S. Feldman, M. Goldberg, K. Keating, and J. Peters. 2001. *A Framework for Planning and Assessing Publicly Funded Energy Efficiency*: Study ID PG&E-SW040, prepared for Pacific Gas & Electric Company.

Stern, F. and D. Vantzis. 2014. "Protocols for Evaluating Energy Efficiency – Both Sides of the Atlantic," *Proceedings of the International Energy Policies & Programmes Evaluation Conference IEPPEC),* Berlin, Germany. September.

Ting, M., M. Rufo, M. Messenger, J. Loper. 2013. Measure Costs: "The Forgotten Child of Energy Efficiency." Proceedings of the European Council for an Energy Efficiency Economy's 2013 Summer Study.

Woolf, T., W. Steinhurst, E. Malone, K. Takahashi, 2012. "Energy Efficiency Cost-Effectiveness Screening: How to Properly Account for 'Other Program Impacts' and Environmental Compliance Costs." *Prepared by Synapse Energy Economics, Inc. for the Regulatory Assistance Project and the Vermont Housing Conservation Board.*