Evaluation and Performance Incentives: Seeking Paths to (Relatively) Peaceful Coexistence

Michael W. Rufo, Itron Inc.,[1] Oakland, CA

ABSTRACT

This paper explores issues and tensions associated with the relationship between financial incentive mechanisms for energy efficiency program performance and evaluation of the load impacts of energy efficiency programs. We first discuss issues related to efforts to tie financial rewards or penalties to program savings or overall performance. We then select several key energy efficiency impact estimation parameters and examine them within the context of performance incentives focusing on elements of controllability, measurability, and fairness. We also discuss how, despite its many other attributes, energy efficiency is a resource with inherently difficult measurement and estimation challenges. Finally, we discuss the risk of evaluation study results being politicized, marginalized, or simply discarded when results lead to negative financial impacts for powerful stakeholders.

We make the case that aspects of performance incentive-related calls for a weakening of, or end to, application of key elements of program evaluation are short sighted. In addition, we argue that these and related challenges to evaluation must be countered to maintain and reinforce evaluation's role in improving the effectiveness of energy-related public policy and programs. We also acknowledge the understandable concerns over how best to use evaluation results within performance incentive mechanisms, given there is sometimes considerable uncertainty associated with measurement of many key parameters. We argue that performance mechanisms need to be designed in advance with an understanding of the key types and expected levels of measurement uncertainty and that doing so will result in more robust mechanisms. Finally, we provide recommendations on how evaluation and performance incentive mechanisms might be better aligned to reduce the likelihood of high stakes conflicts that weaken both.

Introduction

After three decades of booms and busts, energy efficiency is once again at the center of the energy and environment policy debate. Driven largely by the challenges posed by global warming, policymakers are increasingly turning to efficiency to deliver large reductions in energy consumption both quickly and inexpensively. Regulatory and legislative policy makers are turning to energy efficiency as a preferred resource in a growing number of jurisdictions. For example, the California Public Utilities Commission (CPUC) explicitly requires that energy efficiency savings be counted first in the resource planning loading order (CPUC 2004). Other states, such as Pennsylvania, Maryland, Connecticut, Vermont, Texas, and Nevada, have increasingly adopted formal goals for energy efficiency (ACEEE 2007). As a result, energy efficiency can be seen playing a major role in the resource plans of many utilities (Barbose, et al., 2008).

Paralleling this growing reliance on energy efficiency as a resource has been an increase in the use and investigation of incentive mechanisms that provide financial bonuses or penalties to program administrators based on their performance[2] (Cappers, et al., 2008; Jensen, 2007; PPUC 2008, Itron 2008).

---

[1] For affiliation purposes only, perspectives expressed in this paper are solely those of the author.

[2] We use the term "performance" in this paper broadly to refer to a range of elements that likely vary from jurisdiction to jurisdiction. Performance could include any number of elements or mix of element weights. Examples of performance elements include energy savings targets, demand savings targets, market transformation, cost-effectiveness maximization, net benefits maximization, cost minimization, or any number of other factors defined by any particular jurisdiction.

Most states or regulators require some level of program and portfolio evaluation to be conducted as part of such mechanisms. For example, evaluation studies are to provide estimates of gross and net impacts that drive the risk-reward incentive mechanism (RRIM) used to calculate shareholder incentives or penalties for California investor-owned utilities for program years 2006-2008. Under the California RRIM, shareholder incentives and penalties can range between +$500 million and -$500 million (CPUC 2007). Pennsylvania's new energy efficiency act (ACT 129 of 2008) requires formal PPUC evaluation of whether load reduction requirements have been achieved (the act includes penalties of $1 million to $20 million if the goals are not achieved). These types of mechanisms have attracted significant attention from utility management, as well as from ratepayer advocate groups and other stakeholders. As a result of this attention, evaluation plans and study results have come under intense scrutiny and, in some cases, criticism. This attention, and the concomitant political pressure associated with it, has resulted in some calls for an end or weakening of the use of evaluation for performance incentives - or even for an end to certain core elements of evaluation in general, such as net-to-gross estimation. Others have countered that it is these incentive mechanisms themselves that are unviable and must change or end if parties are unwilling to utilize evaluation-based results as the basis for calculation of financial rewards or penalties.

Increased investment in energy efficiency programs, along with increased use and consideration of performance incentives, has in turn led to rapid growth in evaluation funding and corresponding efforts to tie evaluation results to assessments of administrators' performance inclusive of financial consequences. The knitting together of evaluation results and performance incentives is certainly not new; it was common in a number of jurisdictions in the 1990s and continued in a few jurisdictions even prior to the current efficiency funding surge. The stakes, however, seem to have gotten higher, faster in the current environment due to the potential size of performance-based earnings and the inclusion in some jurisdictions of potential penalties (e.g., Pennsylvania), as well as the increasing size of efficiency goals (Rufo, et al., 2008).

With the stakes rising, more attention is being paid to assessing the extent to which evaluation results should be used in performance incentive calculations. Early opining tends to fall into two extreme camps, either complete reliance on ex post evaluation results for most key impact-related parameters or complete reliance on ex ante parameters (with actual measure installation counts being the only ex post element). This paper explores some of the underlying issues behind these ex ante versus ex post perspectives with the goal of encouraging policy makers and parties to consider hybrid approaches aimed at achieving more balanced, effective, and fair outcomes than either of the extreme positions.

Issues and Challenges

In this section, we discuss some of the key issues and challenges associated with real-world application of shareholder incentive mechanisms in combination with independent measurement of program and portfolio performance through evaluation.

Tying earnings to performance

Although shareholder incentives may be tied to varying levels of program performance under different earnings' mechanisms, there is a strong belief on the part of many stakeholders and many regulators that earnings must in concept and in practice be tied to program performance in some way. The key issue for our purpose in this paper is not to discuss the different ways of tying earnings to performance or dissecting the degrees of performance that may be required to trigger earnings. Rather, we take as our starting point the idea that prudent application of earnings for energy efficiency requires some level of assessment, analysis, and judgment of the administrator's actual performance. From this starting point, we then discuss some of the practical challenges that face those designing, implementing, and working within performance-based incentive mechanisms with respect to the issue of performance measurement.

Even once stakeholders accept the basic tenet that assessments of actual performance are necessary within the logic of earnings for portfolio administration, there enters quickly the question of how much of the potential earning should be tied to how much, and what types of, performance. Under some regimes, for example, California's 2006-2008 risk-reward incentive mechanism, relatively large earnings and penalties were meant to be tied to exemplary and poor performance, respectively, with hard triggers linked to quantitative measurement of ex post savings of the entire portfolio. Ex post savings as measured by CPUC managed evaluations were then to be compared to the CPUC savings goal established at the outset of the implementation cycle. In this example, earnings were to trigger at a high fraction of goal accomplishments and penalties at low levels of accomplishment. Under moderate accomplishment, neither earnings nor penalties were to occur (a so-called "deadband"). Other approaches have set earnings on a more continuous basis, beginning with moderate levels of performance. Some recent approaches utilize penalties instead of incentives, assessing them if goals are not met (e.g., Pennsylvania).

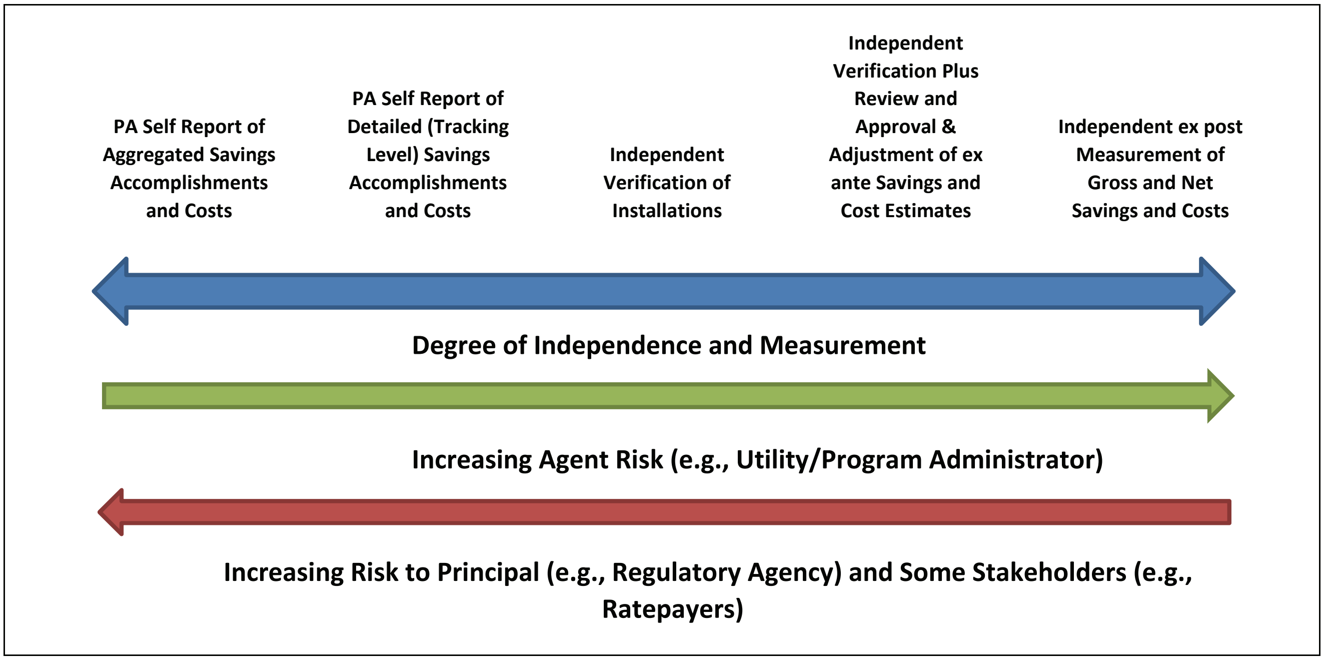
The role and magnitude of evaluation is often closely tied to the size of incentives and penalties and the relative importance of performance in a particular mechanism. Given the wide range in levels of earnings and penalties for energy efficiency performance in historic and currently planned efficiency incentive mechanisms, there has been a concomitantly wide range of approaches used for performance assessment. In Figure 1 we provide a simplified conceptual illustration of the range of performance assessment approaches, from complete administrator self reports to complete independent measurement, as related to the risks posed to principals and agents. At one end of the spectrum are aggregated program administrator self reports, with little independent review or evaluation of assumptions and outcomes used to quantify performance. Historically, this extreme tended to occur in jurisdictions that had no earnings mechanisms, penalties, or aggressive savings targets. Some jurisdictions have accepted such self reports during periods of low levels of efficiency program activity. Moving towards the middle of the continuum we have the inclusion of additional activities such as independent verification of project installations and independent and regulatory review and approval of ex ante savings assumptions. At the other end of the continuum is independent ex post measurement of savings and costs (the original approach required in the CA RRIM and an emerging approach in other jurisdictions). In theory, the interests of the *agent* (e.g., the utility or other portfolio administrator) and the *principal,* the entity regulating or hiring the agent (e.g., utility commissions, state energy agencies) are opposed with respect to this spectrum of performance assessment. The agent's risk increases with more independence and measurement while the principal's increases with less.

Readers may note that the preceding discussion and Figure 1 presume that there is significant risk of bias in ex ante estimates of performance by agents and other financially affected stakeholders (e.g., ratepayers). Almost by definition, in an ex ante only world with financial incentives, the agent has an incentive to set or promote ex ante values that are most advantageous to it financially.[3] Conversely, ratepayer advocates will typically want to set those same values conservatively to avoid the risk of overpayment from ratepayers to utility shareholders or other program administrators. The goal of the principal should be to develop unbiased estimates on program impacts and costs to balance the potentially biased tendencies of parties with direct financial stakes. Note that shifting from a mostly ex post measured to mostly, or even completely, ex ante estimated basis for calculating performance incentives does not resolve problems associated with measurement uncertainty or bias. It simply shifts the debate more intensely from the ex post to the ex ante arena. This can be problematic both because ex ante estimates may

[3] We acknowledge that program administrators may also consider other factors besides financial incentives in their development of ex ante values, for example, the desire to maintain good regulatory relations or to maximize the accuracy of reporting for load forecasting and resource procurement. How they weigh these factors as compared to their direct financial incentives will likely vary from administrator to administrator.

be biased (intentionally or unintentionally) and because there is usually inadequate time available to thoroughly vet and approve the sometimes hundreds or thousands of measure-level parameters that may be associated with a planned portfolio.

Figure1. Principal Versus Agent Risk as Related to Degree of Measurement Independence*



*"PA" refers to program administrator.

Controllability, measurability, and fairness

Another important set of concepts associated with developing incentive mechanisms for energy efficiency program administrators has to do with controllability, measurability, and fairness. It stands to reason that incentives for performance should be related as much as possible to elements of performance over which the administrator is likely to have a reasonable level of control and that the degree to which this is true is likely to be related to the overall fairness of the mechanism. Administrators may argue that there are very few performance parameters under their control due to the dynamic nature and complexity of markets, the range of technology performance in different end user applications, and the vagaries of trade ally practices and consumer decision making and behavior, as well as other factors. From this perspective, it might follow that the bulk of parameters associated with performance incentives should be tied to stipulated, ex ante assumptions. Countering this, however, is the perspective that influencing markets, targeting and directing technology applications, and changing trade ally practices and end user decision making are actually core objectives of efficiency policy and part of the administrator's mandate. An important related factor in balancing controllability and fairness is measurability. The degree to which policy makers tie performance incentives to actual versus planned performance might be related to attainment of minimum levels of certainty in measurement. Assessing measurement uncertainty is thus an important part of improving risk assessment and management and maintaining fairness.

There are many factors that can introduce risk into assessments of program administrator performance, including measure adoption, free ridership, baseline specification and performance, measure specification and performance, measure life, and measure costs among others. Several of these factors that impact the level of savings and cost effectiveness achieved by an administrator are discussed as they relate to controllability, fairness, and measurability below.

Measure Adoptions. Most assessments of program administrator performance start with a strong focus on total measure adoption. Even in cases where all per unit estimates of net performance are stipulated, the degree to which actual performance varies from plan will generally follow directly from the extent to which the number and type of gross adoptions achieved varied from plan. Virtually all jurisdictions assess total gross savings as at least one measure of performance and, in so doing, implicitly utilize counts of total adoptions as a key performance parameter. This is the case even though there can be considerable risk associated with achieving ex ante forecasts of gross adoptions. This risk is partially compensated for by the fact that administrators are often given considerable leeway within their portfolios, allowing them to increase savings from some elements if others are lagging against plan. Perhaps most importantly within the context of this paper is the fact that there is little measurement risk associated with adoption claims. Thus, despite the significant risk associated with achieving measure adoptions, most jurisdictions consider adoption levels to be well under the control of the administrator and, partly due to the low measurement risk and expectations of relatively high program administrator accuracy in self reporting measure counts,[4] generally utilize actual ex post adoptions rather than ex ante targets as the basis for assessing performance. Anything less than actual adoption counting simply does not pass a front page test. Some jurisdictions accept administrator's self reports of adoptions with no further assessment; however, many require independent *verification* of actual measure installations.[5]

Free Ridership. Free ridership rates are one of the most controversial elements of measurement and evaluation within the context of energy efficiency performance incentives. Free ridership is often considered an element over which the 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. For example, a custom incentive program in which each application is reviewed and approved individually offers much more opportunity for control over free ridership (Quantum Consulting, 2005) than does an upstream program targeted at a mass market. Once the administrator sets their program and incentive strategy in the mass market case they will generally only be able to make adjustments after they receive new information from evaluation activities or other program and market feedback mechanisms. Stipulation of ex ante net-to-gross ratios may be the fairest approach in such circumstances if regulators are satisfied that all best available information has been utilized to set net-to-gross values in the design and initial approval of the program; or if it is difficult to obtain new information in a timely enough fashion to make major alterations in program design within a program cycle.

Under custom type programs, however, the situation is usually different. The administrator is continuously reviewing applications and often has the information necessary on each individual project to determine the likelihood of free ridership and take additional actions to reduce it (e.g., by providing incentives only for projects that push beyond the customer's or market's standard practice in a particular area). Evaluators have found ample evidence of high free ridership in these types of programs (Itron, Inc., 2008), particularly in industrial process applications. Program types with high free ridership risk levels *and* high levels of administrator control over free ridership warrant ex post vigilance by principals; at least until administrators show that program tactics and management directives have resulted in reduced free ridership levels. Custom programs often represent a significant portion of overall portfolio savings claims; thus, for these types of programs, tying performance metrics to ex post estimates of free ridership can be seen as a

---

[4] Historically, measure installation verification rates have been very high for traditional downstream incentive programs. However, some programs, such as upstream CFL programs, add a greater degree of ex post performance risk to the measure installation parameter.

[5] What constitutes acceptable verification also varies, ranging from customer self reports to physical on-site inspections.

fair and often necessary approach to mitigating the incentive to otherwise purposefully ignore free ridership in individual cases.

The degree to which free ridership and associated net-to-gross ratios[6] are stipulated based upon ex ante values or determined on an ex post basis from evaluation results is one of the most hotly debated aspects of most energy efficiency performance incentive mechanisms. In our opinion, a balanced assessment of the relative importance of performance, controllability, and measurability may lead to different choices regarding whether free ridership is deemed or measured ex post for different types of programs. Hybrids in which independent evaluators participate in real-time assessment of free ridership likelihood - in custom type programs - are also possible and offer risk reduction benefits to both agents and principals (Fagan, et al., 2006).

Baselines. The extent to which actual baseline specification and energy-related performance differs from that estimated and assumed within ex ante savings calculations can have a significant effect on measure and program performance. But do program administrators have enough control over baseline conditions to make them subject to ex post adjustment as part of incentive mechanisms? Ex ante baseline specification risk is associated with determining whether efficiency estimates for baseline technologies or practices should be based on applicable codes and standards, market averages upon replacement, in situ equipment, or some combination thereof. Another dimension of baseline uncertainty has to do with the operation of the equipment, for example, hours of use and load shape.[7] Like free ridership, the degree of control that the program administrator has over the accuracy of baseline specifications varies by program type. For custom and direct install type programs, it is typical for programs to include verification of baseline conditions as a pre-requisite for provision of incentives (although pre-installation verifications often focus primarily on equipment identification and less so on operational characteristics). For prescriptive and upstream programs, there is typically little or no baseline data collected as part of the program implementation process. In either case, prior to program implementation, baselines must be estimated based on best available information. Some administrators may work hard to collect and analyze baseline information and set program baselines at expected value levels while others may not seek out such information or be tempted to set baseline efficiency levels as low as possible to maximize ex ante savings claims. Principals may want to put extra effort on review of baseline assumptions during ex ante review processes or leave some elements of baseline determination to ex post evaluation. In the case of custom and direct install programs, there are opportunities for principals to use evaluators to assess baselines during program implementation and provide early feedback if adjustments appear necessary.

Remaining and Effective Useful Life. The need to estimate effective useful lives (EUL) of energy efficiency measures is a common and accepted core practice for the estimation of lifetime energy efficiency measure savings and cost effectiveness tests. However, the concept of remaining useful life (RUL) for the existing in situ equipment replaced by the energy efficiency measure is less well understood and less utilized in energy efficiency savings estimation. The concept of RUL is also closely related to baseline and net-to-gross estimation. Some programs (usually retrofit programs) and projects utilize in situ equipment

---

[6] This subsection focused on free ridership, however, much of the discussion is relevant also to spillover. Spillover, like free ridership, is difficult to measure. In contrast to free ridership, it is a highly desirable performance outcome. Ongoing measurement and evaluation of spillover and market effects are an important part of assessing program performance. Whether to include spillover in performance incentives is a separate question that must also weigh the associated measurement uncertainty.

[7] It is typical to assume that baseline operation is identical to measure operation, but there may be important exceptions under which this assumption does not hold true. For example, hours of use may not be the same for the baseline technology and the efficiency measure due to possibly behavioral changes associated with the measure (e.g., CFLs left on for longer than incandescent lamps they replaced due to lower operating costs or knowledge that high on-off rates reduce service life).

efficiencies as the baseline for calculating efficiency measure savings. In some cases, these baselines are utilized over the entire life (EUL) of the efficiency measure. The implicit claim in such cases is that either the program induced extremely early retirement of the existing equipment (e.g., that the RUL equals the EUL) or that any alternative equipment that would have been purchased during the EUL to replace the pre-existing equipment would have been no more efficient than the pre-existing equipment. In some cases, such as for refrigerator early replacement programs, one sees explicit estimation of the RUL and utilization of this as the basis for the savings calculation, but in many cases, such as lighting retrofit programs and industrial process applications, one often sees no such adjustment. To properly assess impacts, early replacement savings should be estimated between the in situ equipment efficiency and the replacement efficient equipment over the RUL and, then, the baseline should be changed at the end of the RUL to represent the efficiency of equipment that would otherwise be purchased at the end of the RUL and, if an efficiency delta remains, only savings relative to the newer baseline should be calculated. If there is no remaining efficiency delta associated with the equipment expected to be purchased at the end of the RUL and the efficiency measure, then no savings beyond the RUL would be calculated. Note that the RUL is closely related to net-to-gross estimation in that the RUL must by definition be program-induced; otherwise the measure installation event is a replace-on-burnout or otherwise naturally occurring event on the part of the end user and there is no RUL (or justification for an in situ baseline).

Let us now consider RUL and EUL with respect to controllability and fairness in the context of measurement and performance incentives. Both EUL and RULs are difficult to measure in practice and measurement can be costly. In the case of EULs, many years must often pass before the mean or median life is approached. Tracking measures in the field over years or decades takes considerable effort and resources. More importantly within the context of performance incentives, the time lag associated with such tracking and EUL estimation is likely far too long to be acceptable as a financial risk for most administrators. In addition, there is considerable inherent uncertainty in EULs over which the administrator has little control. On the other hand, EULs can have a significant effect on performance incentives that are based on lifecycle benefits. Thus EULs should be periodically measured to at least improve ex ante estimates on an ongoing basis; however, utilizing ex post measurement of EULs in performance incentives may be impractical except perhaps in limited situations (e.g., where there is strong evidence that best available information was purposefully not used in ex ante assumptions). Estimation of RULs poses a similar type of measurement challenge and adds the additional uncertainty associated with the fact RULs cannot be directly observed. On the other hand, overestimation of RULs can lead to very significant overestimation of program lifecycle program impacts. Thus, again, even if not subject to ex post true up, increased efforts are needed to improve estimation of RULs to increase confidence in savings claims and cost effectiveness generally, but particularly in cases where administrator's performance incentives are tied to lifecycle savings impacts.

Measure Costs. Though measure costs are not directly related to program and portfolio impacts, they are still of great import to many types of performance incentive mechanisms, specifically any mechanism that uses net (or even gross) total resource cost (TRC) benefits in some way. Measure costs are a fundamental input into cost effectiveness calculations, being the denominator, with the addition of program administration costs, of benefit-cost tests. Strangely, however, measure costs and incremental measure costs have been terribly neglected as an empirical field of study. The reason for this neglect may be related to perceived difficulties with cost measurement or perhaps it has simply become a long-standing bad habit that is difficult to break. Consider the fact that over the past twenty years, cumulative expenditures on evaluation efforts aimed at estimating program impacts have probably been on the order of half a billion to a billion dollars, while over this same time period, cumulative expenditures on cost estimation are likely on the order of five to ten million dollars. Thus, expenditures on measure cost estimation likely represent about 1 percent as much even though their import is almost equal in the TRC test.

With respect to controllability and fairness, actual measure costs generally are not always directly under the administrator's control, although program designs can have significant indirect effects on such costs. Administrators do have direct control over incentive costs, which enter into some performance incentive mechanisms through the program administrator test (PAC). Incentive costs are generally easy to verify through simple auditing procedures and are usually not a controversial aspect of ex post measurement and verification efforts. Incremental measure costs can be difficult to measure and generally require formal measure cost studies to develop (e.g., using matched pair and hedonic pricing methods). Because of this and because administrator's influence over costs is mostly indirect and time lagged, incremental costs are probably best kept to stipulation of ex ante values. For measures and programs where administrators purposefully focus on pure retrofit or significant periods of early replacement, full measure costs become a more important part of cost effectiveness calculations. In addition, for some such measures and programs, administrators do sometimes have the ability to exert strong influence over costs because incentive levels are set at a very high percentage of full – not incremental - costs (e.g., direct install lighting retrofit programs). In these situations, due diligence review, measurement, and analysis of actual participant costs may be useful to measure whether program efforts have decreased or increased participant costs.[8]

Measurement uncertainty

Despite many excellent and attractive resource attributes, one characteristic of energy efficiency that has been and continues to be a challenge is the difficulty of measuring it. These measurement difficulties in turn lead to varying levels of uncertainty around estimates of energy efficiency impacts. Uncertainty around individual parameters and measures is, however, much more manageable if the underlying estimates are unbiased because errors will tend to cancel one another if portfolios are well diversified. Conversely, the benefits of diversification are limited if savings estimates are systematically biased one way or another (i.e., high or low).

Rather than focusing on the problem of eliminating or minimizing systematic bias, there is instead an unfortunate tendency of late in the efficiency industry to claim that the presence of uncertainty discredits the measurement science itself and, in a logic twist, increases the validity of stakeholder savings claims, even when those stakeholders have direct financial conflicts of interest. It is disheartening to hear these arguments in the efficiency industry given that this kind of reaction to scientific uncertainty has been discredited in the area of climate change (which, ironically, has been the most important driver of the recent surge in efficiency program investments). It is sometimes argued that, as a result of measurement uncertainty, we should stop trying to measure that which is difficult to measure, in particular, free ridership, short-term spillover and long-term market effects. But this approach will likely only hurt the perception of efficiency as a resource and approach to greenhouse gas reduction.

This is not to say that there are not legitimate concerns over measurement uncertainty that may in some cases warrant exclusion of elements of evaluation results from calculations of financial payments or penalties for program administrator's performance. However, we argue that such exclusions should be based on thoughtful examination of the tradeoffs between controllability and measurability, as well as other factors, on a case by case basis, in order to produce fair and effective performance mechanisms. Blanket arguments that measurement uncertainty is not only a basis for excluding evaluation results from financial calculations, but also a basis for ceasing or reducing measurement itself will assuredly undermine confidence in efficiency impacts over the long term. Rather, performance mechanisms need to be designed

---

[8] Admittedly, such measure cost analyses can be difficult in practice. However, as noted above, this is an area in which evaluation efforts can be significantly improved (though such advances should probably be attempted first in jurisdictions with already extensive experience in energy efficiency evaluation).

in advance with an understanding of the key types and levels of measurement uncertainty. Doing so will result in more robust financial risk and reward mechanisms that are less sensitive to underlying uncertainty levels (whether ex ante or ex post). The result will also hopefully increase – not reduce – focus on measurement and evaluation and thereby lead to further improvements and increased confidence in results.

Money, politics and science – Can the center hold?

Some observers of performance incentive mechanisms ask whether parties and regulatory commissions will have the fortitude to accept evaluation results, uncertainty and all, when doing so may have significant financial consequences. There may be a correlation between the dollar magnitude of performance incentives, the political power of the party on the down side of the evaluation results, and a decrease in the likelihood that the results will be used as intended. That is, the bigger the financial stakes and greater political power of the affected party, the less likely the evaluation results are to be utilized if they go against the interests of the party with strong political power. This is perhaps obvious but well worth examining and continuing to monitor for, if true, it begs the question of whether evaluations should be designed into performance incentive mechanisms in the first place.[9] Of course, conversely, admission of this tendency would beg the question as to what value a performance incentive has at all if the environment is such that the evaluation results are likely doomed to being politicized if they come out on the *wrong* side. It could also lead to cynicism and cast a shadow of doubt over the veracity of efficiency program claims at a time when policy makers are interested in increasing investment and reliance on this important resource. Maintaining the integrity of the evaluation science for the purposes of tracking and reporting actual impacts to policy makers, resource planners, and forecasters may be more important in the long haul than using evaluation results mechanically for short-term performance incentives. That might lead to removal of evaluation from the process of directly calculating financial rewards and penalties in exchange for its continued and favored use as the basis of reported impacts for long-term tracking, load forecast adjustments, resource procurement, and program planning.[10]

Recommendations

Both evaluation and performance incentive mechanisms have important roles to play with respect to increasing and improving energy efficiency program impacts. Recommendations for improving coexistence between these two policy tools are provided below.

Consider hybrid performance incentive mechanisms that balance the extremes of using all ex ante versus all ex post measurement with respect to assessing actual performance and application of incentives or penalties. An objective of this paper has been to emphasize that neither complete ex ante estimation nor complete ex post measurement is likely to be optimally fair, efficient, or practical with respect to performance-based incentives or penalties. Instead, some combination of these approaches is more likely to appropriately balance the competing perspectives of parties, agents, and principals. This can be accomplished through thorough assessment of degrees of controllability, accountability, and measurability - on a case-by-case or type-by-type basis - for the most important underlying impact and cost estimates of key measures and programs.

---

[9] The degree to which the oversight environment is politicized by powerful interests is obviously case by case and may vary widely across jurisdictions and overtime within them.

[10] Though better than an avoidance of evaluation altogether, this is likely a second best outcome as the keeping of separate books – one for performance payments and one for everything else – may not reflect well on the resource or political processes around it, and will invariably be difficult and confusing to maintain over time.

Consider performance incentive mechanisms that can use directional results or ranges. Because there is often significant uncertainty around many of the underlying parameters needed to estimate energy efficiency impacts,[11] performance incentive mechanisms need to be designed with this in mind. Though baseline and evaluation methods will likely continue to improve (if there is a corresponding willingness to invest in these activities), there will still likely remain at least moderate uncertainty around efficiency impacts.[12] Thus, performance incentive mechanisms should try to reduce their sensitivity to narrow point estimates (i.e., avoid approaches that have large financial consequences based on small to moderate changes and uncertainties in key impact parameters).

Consider mechanisms that have continuous rather than discrete performance criteria. Related to the recommendation above, one of the ways that performance mechanisms can be designed to be less sensitive to uncertainties in point estimates is to utilize continuous rather than discrete performance criteria.[13]

Incorporate periodic ex ante true-ups using best available information during long program cycles. Over the past few years, many energy efficiency program cycles have increased from one year to multiple years. Although there are many advantages of a multiyear program cycle, one disadvantage that can sometimes occur is that evaluation results become very lagged if they are held until the end of the cycle. Waiting until the cycle is complete is necessary to make a full accounting of total accomplishments. At the same time, evaluations and regulatory processes should be designed to provide as much in the way of early and interim results as possible so that programs and portfolios can adjust to any changes in best available information. Developing and using timely interim evaluation and related research to improve planning estimates of key parameters such as net-to-gross and unit energy savings for key measures and programs can lead to significant adjustments and improvements in final performance. In contrast, purposefully ignoring interim results can increase the likelihood of having highly contentious and acrimonious disagreements over performance at the end of cycle, a time when debate over potential design adjustments is rendered moot.

Adjust goals in parallel with evaluations of portfolio impacts when appropriate. Another approach that should be considered in policy environments that tie performance incentives to specific savings goals is to use evaluation results to make any necessary adjustments in the goals in parallel with use of such results to develop ex post savings estimates. Keeping goals and ex post savings estimates in step temporally can reduce underlying inconsistencies and thus increase fairness.

When using ex ante values as the basis for planning or some elements of performance assessment, ensure an expected value orientation. Regardless of whether a given parameter for a particular measure or program is stipulated or measured ex post for purposes of performance incentives, ex ante estimates should be developed from an expected value orientation. Systematic biases across parameters, that is, a deliberate tendency toward purposefully optimistic or conservative assumptions,

---

[11] Note that the more diversified efficiency portfolios are and the more unbiased the underlying individual measure and program parameters, the more robust will be the estimates of overall portfolio impacts, since many of the underlying errors will tend to cancel if they are random. Conversely, the more portfolios are heavily weighted to a small set of measures or programs and have ex ante parameter estimates that have systematic bias (i.e., bias correlated across measures and programs), the more likely that portfolios impact estimates will have large uncertainty bands.

[12] Some may mistakenly see this emphasis on residual uncertainty as a kind of admission of failure on the part of the evaluation industry or that it makes evaluation work somehow less than scientific; however, this is an erroneous conclusion. Uncertainty is a part of virtually all scientific endeavors. It is the ignoring or denial of uncertainty, or deliberate biasing of estimates in response to it, that is unscientific.

[13] Taking the CA RRIM as an example (CPUC 2007), this would mean eliminating the discrete cliffs in the mechanism that occur at 65% and 85% of goal accomplishment (65% being the point and which penalties occur and 85% the point at which rewards begin, in between being the so-called dead band).

should be assiduously avoided. An expected value approach is vital and robust in that inevitable errors in measure- and program-level parameters are more likely to cancel in aggregate.

Use ex post evaluation-based estimates of savings for long-term tracking and reporting of efficiency program accomplishments. A concern that arises when performance payments are tied more to ex ante estimates than to ex post evaluation results is which set of impact estimates should be used for long-term tracking and report of accomplishments. As discussed in this paper, considerations of controllability, measurability, and politics may lead to the use of ex ante estimates for some or even perhaps many parameters, measures, and programs for the purposes of financial payments or a portion thereof; however, such practical compromises should not obfuscate the fact that the ex post, evaluation-based estimates of program impacts should be the basis of official record keeping and tracking. This is critical to the further advancement of efficiency as a resource in which planners, forecasters, and policy makers can have confidence. Anything less risks continued wholesale discounting of reported impacts.

Separate the science from the politics. Energy efficiency evaluations are not privileged in their being subjected to intense scrutiny and criticism from parties financially affected by their findings. There are innumerable examples of science and moneyed interests coming into extreme conflict in almost every field of human endeavor. Nor should efficiency-related evaluations be specially protected from such scrutiny and critical review, rather the opposite. If evaluations are to be respected as science, then they need to be treated with all of the review and expectations of repeatability that comes with it. Review and honest criticism are integral to the scientific process; however, politicizing or censuring results is not.[14] Maintaining an environment that fosters scientific inquiry is paramount to the long-term credibility of the efficiency enterprise.

[14] The political pressure to censure science is a political tendency of which to be vigilant (Revkin, 2006).

References

ACEEE. 2007. *The State Energy Efficiency Scorecard for 2006*, Report Number EO75, American Council for an Energy Efficient Economy. June.

Barbose, Galen, Ryan Wiser, Amol Phadke, and Charles Goldman. 2008. *Reading the Tea Leaves: How Utilities in the West Are Managing Carbon Regulatory Risk in the Resource Plans*. LBNL-44E. Berkeley, Calif.: Lawrence Berkeley National Laboratory.

Cappers, P., C. Goldman, M. Chait, G. Edgar, J. Schlegel, W. Shirley. 2008. "Quantitative Financial Analysis of Alternative Energy Efficiency Shareholder Incentive Mechanisms", Proceedings of the American Council for an Energy Efficient Economy (ACEEE) 2008 Summer Study on Energy Efficiency in Buildings. Washington, D.C. August.

CPUC. 2007. Interim Opinion on Phase 1 Issues: Shareholder Risk/Reward Incentive Mechanism for Energy Efficiency Programs, September 20.

California Public Utility Commission (CPUC). 2004. Decision 04-09-060, Interim Opinion: Energy Savings Goals for Program Year 2006 and Beyond. Rulemaking 01-08-028, Order Instituting Rulemaking to Examine the Commission's Future Energy Efficiency Policies, Administration and Programs. September.

C. Blumstein. 2009. "Program Evaluation and Incentives for Administrators of Energy-Efficiency Programs: Can Evaluation Solve the Principal/Agent Problem?" Center for the Study of Energy Markets (CSEM) Working Paper, CSEM WP 184. March.

Fagan, J., K. Bradley, and M. Rufo. 2006. "They Said It Couldn't Be Done: One Program's Approach to Minimization of Free Ridership," proceedings of the 2006 National Energy Services Conference.

Itron, Inc. 2008. *Portfolio Best Practices Report*, National Energy Efficiency Best Practices Study, prepared for California Best Practices Project Advisory Committee. July. ***www.eebestpractices.com***

Jensen, V. 2007. Aligning Utility Incentives with Investment in Energy Efficiency: A Product of the National Action Plan for Energy Efficiency. Fairfax. VA.: ICF International, Inc.

Pennsylvania Public Utility Commission (PPUC). 2008. *Act 129 of 2008.*

Itron Inc. 2008. 2004-2005 Statewide Nonresidential Standard Performance Contract Program Measurement and Evaluation Study, Final Report, September.

Quantum Consulting, Inc. 2005. Energy Efficiency Best Practices Study - Large Nonresidential Comprehensive Incentives Programs, prepared for California Best Practices Project Advisory Committee. July. ***www.eebestpractices.com***

Revkin, A. 2006. "Climate Expert Says NASA Tried to Silence Him," New York Times, January 29.

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.