The Estimation of Spillover: EM&V's Orphan Gets a Home

Ralph Prahl, Prahl & Associates, University Park, FL
Richard Ridge, Ridge & Associates, San Francisco, CA
Nick Hall, TecMarket Works, Oregon, WI
William Saxonis, New York Department of Public Service, Albany, NY

ABSTRACT

Spillover has been defined as the energy savings associated with energy efficiency measures (EEMs) adopted by consumers who were influenced by an energy efficiency program, but without direct financial or technical assistance from the program. Spillover includes additional actions taken by a program participant as well as actions undertaken by non-participants who have been influenced by the program. While this paper focuses on spillover, we note that observed spillover that is sustainable and reflective of structural changes in the market is referred to as market effects. Spillover savings have always been a relatively small share of the direct savings from any given energy efficiency (EE) program and have, as a result been given limited attention by the EM&V community. It has been considered more like an added bonus. However, in recent years, the importance of spillover has been increasing in jurisdictions where the total resource cost test (TRC) results have been decreasing making even small amounts of spillover the focus of greater attention. In some jurisdictions, the spillover share has become substantial, rising in some cases, well over 50 percent of program net savings. The problem is that, while there are many sets of EM&V guidelines, none focus directly on developing reliable estimates of spillover. This has also led to a situation in which spillover is estimated using a wide variety of methods, some of which are not particularly reliable.

In response to both of these phenomena, a team of contractors assisting the New York Department of Public Service (DPS) with its evaluation oversight activities was asked to develop spillover guidelines for New York State. This paper relates the background, discusses common limitations in spillover methods, summarizes the provisions of the guidelines intended to address these limitations, and discusses the potential for their application in other states.

INTRODUCTION

In June 2008, the New York Department of Public Service (DPS) released an order establishing the New York Energy Efficiency Portfolio Standard (EEPS), one of the nation's more ambitious energy efficiency initiatives. The order establishing EEPS significantly increased total energy efficiency funding in New York and expanded the range of program administrators to include the investor owned utilities, which had not managed comprehensive energy efficiency program portfolios in about 10 years. The EEPS Order also recognized that the significant increase in energy efficiency funding called for a commensurate increase in the comprehensiveness and rigor of evaluation efforts. The Order thus increased evaluation funding from 2 percent to 5% of program budgets, and simultaneously established a statewide Evaluation Advisory Group (EAG) to advise the Commission and Department of Public Service staff on evaluation related-issues. Further, the Order directed DPS staff to issue evaluation guidelines to inform evaluation planning activities and bring uniformity to evaluation practices and reporting. NYDPS staff released an initial set of these required guidelines in August 2008. Lastly, recognizing that increased involvement in the oversight of evaluation activities would require increased technical resources, the Order authorized the hiring of a consultant to provide technical assistance to DPS staff and the EAG.

In 2009, following the release of a competitive RFP, a team of consultants led by TecMarket Works was selected to provide the technical assistance envisioned by the Order. Since that time, the DPS contractor team has been actively involved in assisting DPS staff by reviewing and commenting on evaluation plans, reports and other deliverables; providing advice on statewide evaluation policy issues; helping to prepare guidelines for program administrators on specific evaluation topics; and working informally with program administrators and their evaluation contractors.

Estimation of spillover savings, the energy savings associated with energy efficiency measures adopted by consumers who were influenced by an energy efficiency program without direct financial or technical assistance, has long been a key evaluation activity in New York. Spillover savings estimates have often proven to be of significant magnitude relative to net in-program savings. Because of this history, the DPS contractor team has from its inception devoted significant effort to reviewing and commenting on the methods being used to estimate spillover. By 2011, the DPS staff and its contractor team had developed concerns about what they viewed as certain recurring methodological limitations in the estimation of spillover by program administrators – limitations that seemed to be fairly common in the industry as a whole. In 2012, the DPS staff thus asked its contractor team to prepare a set of guidelines for the estimation of spillover savings, which were to be incorporated into the overall guidelines first developed in 2008. In November 2012, the new spillover guidelines were incorporated into the overall evaluation guidelines, and have since begun to influence the methods used to estimate spillover in New York (New York Department of Public Service, November 2012).

The focus of this paper is on the substantive methodological issues that led to the development of the New York spillover guidelines, and on the methodological requirements introduced by the guidelines in order to help address these issues. Our primary intent is not to rehash specific events and outcomes in New York in any detail, but rather to help advance the state of the art of the measurement of spillover nationwide. A key premise of the paper is that many of the methodological issues observed in studies commissioned by the New York program administrators are common elsewhere in the country, and that the methodological requirements that were adopted to help address these issues may thus be of interest to other states, program administrators, and evaluation practitioners. A second key premise is that reliable estimation of spillover is important and worth the effort, both to avoid understating program impacts and to lay the groundwork for the documentation of market effects.[1]

We begin with a discussion of methodological limitations that the authors have observed both in New York and in many other states. We then summarize the provisions of the New York spillover guidelines that were intended to help address these limitations, and provide a link for readers interested in reviewing the guidelines in their entirety. Finally, we close with a brief discussion of the potential relevance of the New York spillover guidelines to other states.

COMMON METHODOLOGICAL LIMITATIONS IN SPILLOVER ANALYSES

Any discussion of methods for the estimation of spillover must begin with the simple recognition that it is extremely difficult to do. Establishing causality for energy-efficiency measures (EEMs) taken by end-users in the course of their participation in programs is challenging enough, and has for many years been surrounded by a certain degree of controversy. Establishing causality for EEMs adopted outside of the program must deal with many additional challenges, including identifying what measures were adopted, estimating gross unit savings despite often having limited first-hand evidence of the measures, and establishing the role of the program as part of a broad array of forces bearing on end-user behaviors. Given these challenges, it seems clear that the perfect is the enemy of the good when it comes to estimating spillover, and we must accept that there are some hard limits to the reliability of the answers that social

[1] The conceptual relationship between spillover and market effects is discussed later in this paper.

science and engineering methods can provide. Nonetheless, the authors' experience has been that there are a number of industry practices in the estimation of spillover that could arguably be improved upon. The following is a discussion of some of these practices.

1. *Exclusive reliance on self-reports*. Much has been written in the industry about the limitations of self-reports as a method for the estimation of free riding, and the need to marshal other methods. However, the estimation of spillover has, if anything, been even more heavily reliant on self-reports than has the estimation of free riding, and the threats to validity surrounding this approach are probably greater for spillover. For free riding, the use of multiple methods with triangulation of the results has become fairly common. However, spillover is often treated as a secondary issue, and as a result there has often proven to be insufficient funding to deploy multiple methods (if there is even sufficient funding to execute one method effectively). The result is that the vast majority of spillover analyses rely on self-reports (either from end-users or from supply-side actors) as the sole method. This lack of methodological diversification tends to limit our ability to develop a reliable, cumulative understanding of the issue. At the same time, the market actors who are queried about spillover are often much less engaged in energy efficiency programs and issues than those who are queried about free riding, leading to potential salience issues.
2. *Reliance on survey questions regarding program influence that have the potential to inadvertently lead the respondent*. When self-reports are used to assess spillover, it is fairly common to launch directly into questions about the magnitude of the effect of the program on respondents' actions, without first asking more general or threshold questions about the range of other influences and whether the program had any effect at all. From a survey design perspective, such a direct approach has several drawbacks. To ask about the influence of the program without first asking about other influences on the respondent's actions may focus the respondent unduly on the effects of a single driver within a complex decision-making environment, potentially leading them to overestimate the effects of that driver. Similarly, to ask about the magnitude of the influence of the program without first asking whether the program has an influence at all may put the respondent in a mindset of assuming that there must be some effect, and it is only a matter of quantifying it. For these reasons, the authors would argue that self-report batteries on the causation of spillover should generally begin by asking the respondent in an open-ended manner about the range of causes of out-of-program adoption of EEMs; then proceed to ask whether the program has any influence; and only then, and only if the response to the preceding threshold question is positive, to ask about the magnitude of the influence.
3. *Lack of analysis of the specific causal mechanisms thought to be leading to spillover.* Spillover claims will generally be more convincing if they are accompanied by an explanation as to the specific causal mechanism by which the program is causing out-of-program adoptions. For example, are end-users being educated and thus seeking out EEMs in other circumstances? Are supply-side actors changing their business practices in a manner that promotes the measures being offered in the program? Is the mechanism something so simple as participating end-users being induced by the program incentive to adopt a measure but then simply neglecting to file for the incentive? In the absence of exploration of such causal mechanisms, the reviewer is left to wonder why the program should be viewed as likely to produce spillover, and the other weaknesses discussed above and below become all the more worrisome. Ideally, the program theory will clearly specify the causal mechanisms through which spillover is expected to be produced, and the evaluation will focus on assessing whether the specific hypothesized causal mechanisms are in fact occurring. However, in the absence of a clear program theory, researchers can still develop hypotheses about causal mechanisms through interviews with program staff and review of the literature on similar programs elsewhere, and allow these hypotheses to drive the study design.

Finally, even if the development of such up-front hypotheses is impractical, it should generally be feasible for self-report batteries on spillover issues to probe into the specific causal mechanisms behind the effects reported by the respondent. Spillover analyses that make none of these efforts to establish the causal mechanisms underlying claimed spillover savings should generally be viewed as unpersuasive.

4. *On the supply-side, leverage points, or a very small number of respondents driving the results.* Supply-side market actors such as contractors, distributors, and design professionals tend to be difficult to reach, particularly when they have not been directly involved in the program. Further, it is often necessary to use professional rather than survey house staff to interview them, in order for the research effort to display sufficient credibility to elicit their cooperation. Both of these factors tend to lead evaluators to limit supple-side sample sizes in order to manage costs. At the same time, supply-side factors also tend to vary greatly both in their size and in the level of program influence they report, meaning that the overall spillover results from any one sample are often dominated by a small subset of large respondents. The net effect of these two factors (small sample sizes driven by cost containment measures and the tendency for results to be dominated by a small subset of large respondents) is that spillover results based on supply-side interviews are often driven by a handful of large and influential respondents. Statisticians call such cases leverage points. Leverage points tend to cause great uncertainty, even above and beyond what might be estimated based on a straightforward calculation of sampling error. Sampling and analysis methods to limit the potential for leverage points to skew the results have long been in use in the measurement of gross savings and free riding, but in the experience of the authors, the use of such methods in spillover analyses based on supply-side interviews is rare. As a result, leverage points are often a major issue source of uncertainty in such analyses.
5. *Unconfirmed assumptions that gross savings from each unit/end-user/project are the same for spillover as for participants.* For a variety of reasons, per-unit gross savings from spillover measures are often difficult to estimate. It is challenging to identify specific spillover measures in the population, and once such measures have been identified, there are often insufficient funds and/or survey time to characterize the measure in sufficient detail to come up with a reliable estimate of gross savings. A common shortcut in dealing with this issue is to assume that the average savings from each spillover measure or project are the same outside of the program as they are inside the program. It is easy to understand why evaluators resort to this assumption. Unfortunately, however, there are good reasons to expect the assumption to result not only in imprecision, but also in a predictable upward bias in the savings estimate. All other things being equal, one would expect an end-user contemplating a relatively small project to be less likely to find his or her way into a program than an end-user contemplating a large project, as the payoff for incurring the transaction costs of participating in the program is larger. For example, a homeowner contemplating insulating his or her walls and attic may be more likely to seek out a program than one contemplating only installing sillbox insulation. Not surprisingly, in the authors' experience, when systematic comparisons are made between the same measures installed inside and outside of the program, those installed outside of the program are typically found to produce smaller average savings.
6. *Reliance on outdated studies.* Because comprehensive spillover studies are expensive to perform, often they are performed only infrequently, and thus used for many years. Yet markets, and the degree of program influence on market participants, can evolve quite rapidly. Spillover results can thus easily become outdated.
7. *Double counting.* The analysis of spillover is rife with the potential for double counting. One reason is the complexity of the paths through which products may travel in a market. Products generally flow in a downward fashion from manufacturers, to distributors, to contractors or retailers, to end-users. Good spillover studies often look at multiple links in the distribution chain. When this is the

case, analytic measures must be taken to ensure that the same spillover measure is not being counted multiple times. Another potential driver of double counting is methodological diversification. Because spillover is difficult to capture, good studies often use multiple methods to capture it. If care is not taken, these multiple methods can also result in the same measure being claimed repeatedly. A good example is a study reviewed by the authors, in which a billing analysis was used to determine participant savings, and a survey to estimate additional savings due to spillover measures installed by participants. By its nature, a billing analysis will tend to capture the effects of all EEMs installed by the participant. This study was thus likely to be counting participant spillover (at least that achieved in the short-term) both in the billing analysis and in the survey, leading to some degree of double counting.[2]

8. *Underinvestment in the analysis of spillover, relative to the level of investment in estimating gross savings and/or free riding.* This last weakness is not so much a distinct issue from the previous eight, as a frequent cause of them. In general, we would argue that the level of investment in the measurement of any one impact evaluation parameter should be roughly proportional to the relative magnitude of the estimated impacts and of the uncertainty introduced by that parameter. If the parameter is well understood, small in magnitude, shows little variation, or affects only a small portion of the end-user population, it generally does not merit a large allocation of resources. If it is poorly understood, large in magnitude, shows much variation, or affects a large portion of the population, it merits a much larger investment of resources. In our experience, the magnitude of spillover observed tends to vary widely across programs and markets. Not infrequently, it is fairly clear that the effects of spillover are likely to be limited, and thus a rigorous effort to measure those effects is not called for. However, we have often encountered studies in which there was every reason to expect spillover to be a both a major factor and a highly uncertain one, yet the total resources invested in measuring it came to a fraction of those invested in measuring gross savings and/or free riding. Because gross savings evaluation can involve on-site and engineering methods that are quite expensive on a per-unit basis, what is most common in our experience is disparity between the level of investment in gross savings vs. spillover. We have reviewed studies in which spillover ultimately accounted for more than half of total reported net savings, yet the amount of money spent estimating it was perhaps 5% of that spent estimating gross savings, even when many of the methodological weaknesses discussed above were present. Underinvestment in the estimation of spillover must thus be regarded as a weakness in its own right, and one that frequently underlies other weaknesses.

THE NEW YORK SPILLOVER GUIDELINES

It was in recognition of the common methodological problem areas discussed above that DPS staff became interested, in the Summer of 2012, in developing statewide guidelines for the estimation of spillover effects. The guidelines were developed by the DPS contractor team under the direction of DPS staff, and went through several rounds of review by the EAG before being finalized in late 2012. Due to space limitations we do not reproduce the guidelines in their entirety here, but rather summarize a few of their key components, and, for interested readers, provide a link to the guidelines themselves.[3] Components of the guidelines that are discussed below include: (1) definitions; (2) the relationship between spillover and market effects; (3) critical issues that must be resolved by evaluators before undertaking a spillover analysis; (4) methodological requirements; (5) requirements regarding levels of precision and confidence; and (6) the

[2] Conversely, when a billing analysis deploys a comparison group, if non-participant spillover is present in the population, it may well end up being inadvertently subtracted from participant savings.

[3] http://www3.dps.ny.gov/W/PSCWeb.nsf/All/766A83DCE56ECA35852576DA006D79A7?OpenDocument

avoidance of double counting. Additional components of the guidelines that are not discussed in this paper include calculation of the spillover rate and deemed savings approaches.

Definition

The guidelines begin by defining spillover as:

> . . . the energy savings associated with energy efficient equipment installed by consumers who were influenced by an energy efficiency program, but without direct financial or technical assistance from the program. Spillover includes additional actions taken by a program participant as well as actions undertaken by non-participants who have been influenced by the program.

The guidelines allow, but do not require, further disaggregation of spillover into three commonly recognized sub-categories: inside spillover (ISO), which occurs when additional program-induced actions are taken at the participating site; outside spillover (OSO), which occurs when an actor participating in the program initiates additional actions that reduce energy use at other sites that are not participating in the program;[4] and non-participant spillover (NPSO), which occurs when actors not participating in the program are induced to take action.[5]

Spillover and Market Effects

An important issue on which the guidelines are largely silent is the conceptual relationship between spillover and market effects. This reflects two facts. First, evaluations in New York have routinely addressed spillover, but few to date have explicitly sought to estimate market effects. Second, the authors believed that there was not a clear industry consensus on the relationship between spillover and market effects, and that attempting to address both in the same guidelines could therefore end up simply complicating efforts to establish standards for the estimation of spillover. Thus, instead of seeking to resolve the issue of the relationship between these two constructs, the guidelines focus on defining spillover and providing methodological guidance regarding how to measure it. However, the guidelines do recognize that market effects studies are one valid approach to estimating spillover, and include references to a few key guides regarding market effects methods.

While the issue is not addressed in the New York spillover guidelines, the authors would suggest the following conceptual relationship between spillover, market effects, and market transformation: market effects are best viewed as spillover savings that reflect significant program-induced changes in the structure or functioning of energy efficiency markets. Market transformation is market effects that are substantial and relatively lasting. Thus defined, in terms of the breadth of program effects that qualify for inclusion, market

[4] It is worth noting that one implication of these definitions is that how a piece of savings is classified may depend in part on the objectives of the program and what outcomes the program has chosen to track. As a key example, program influence achieved through the provision of technical information is clearly a legitimate source of savings, but, depending on the specifics of the situation, could end up being classified either as in-program savings, participant spillover, or non-participant spillover. If the provision of information is considered sufficiently central to the program objectives for the program to directly track this outcome, then information-induced measures may be classified as in-program savings. If information-induced measures are not tracked but are adopted by participants who also adopted rebated measures, and thus entered the tracking system, then they may end up being classified as participant spillover. If untracked information-induced measures are adopted by end-users who did not also adopt a measure for which they received a rebate, then they may be classified as non-participant spillover. While all of this suggests that the precise meaning of these terms can be somewhat specific to the situation, the guidelines are intended to provide methodological guidance that is resilient in the face of such distinctions.

[5] While not called out in the guidelines, other categorizations of spillover are possible. One useful distinction may be the time frame involved (e.g., near-term, mid-term, long-term). Another important distinction, as discussed in the next section, is whether the spillover involves significant structural changes in the market, thus constituting a market effect.

effects are a subset of spillover, and market transformation is a subset of market effects. This proposed relationship reflects the fact that not all spillover effects reflect significant program-induced changes in the market. For example, when an end-user is induced by a rebate to adopt an EEM but forgets to file for the rebate, this constitutes spillover, but does not qualify as a market effect.

This formulation of the conceptual relationship between spillover and market effects has important practical implications. One implication is that, because market effects are one specific type of spillover, for the most part the methodological issues affecting the estimation of spillover are also relevant to the estimation of market effects. However, because market effects studies must seek to pinpoint lasting structural changes in the market, they face special methodological requirements above and beyond those faced by other types of spillover studies. A third implication is that documenting spillover is not necessarily sufficient to document market effects or market transformation. Practitioners seeking to measure these latter outcomes must generally adhere to the best practices established for the estimation of both spillover *and* market effects. A fourth and final implication is that, because some but not all spillover effects are lasting, it is important that evaluations consider how long the specific types of spillover observed are likely to last.

Key Decisions for Evaluators

Next, the guidelines lay out a number of critical decisions that an evaluator must make before deciding whether and how to estimate spillover, including the following:

1. Will the evaluation address participant spillover, nonparticipant spillover, or both?
2. Does the size of the expected savings warrant the expenditure of evaluation funds needed to estimate these savings at an appropriate level of reliability?
3. Will spillover be estimated based on data collected from end users, those upstream from end users (e.g., vendors, installers, manufacturers, etc.), or both?
4. What is the level of aggregation? Although participant spillover is always estimated at the program level, if an evaluator is attempting to estimate nonparticipant spillover, will the evaluator estimate it at the program level or the market level? One potential reason for estimating nonparticipant spillover at the market level is that, in some circumstances, reliably teasing out the spillover savings attributable to one specific program among many may be nearly impossible due to the difficulty nonparticipants may have in attributing any of their installations to a specific program. In such a case, evaluators can choose to conduct market effects studies which include naturally occurring adoptions, program-rebated adoptions, participant and nonparticipant spillover, other program effects that cannot be reliably attributed to a specific program (e.g., upstream lighting programs and the effects of the portfolio of programs on such things as increases in the allocation of shelving space to efficient measures), and other non-program effects due to such factors as DOE Energy Star, programs funded by the American Recovery and Reinvestment Act (ARRA) and the gradual non-program induced evolution of the market in terms of attitudes, knowledge and behavior regarding energy efficiency. The net savings resulting from market effects studies must be included in the portfolio-level benefits-costs analyses.
5. If an evaluator decides to conduct a market effects study, then he or she must decide whether the study should be focused on the region targeted by a given PA, multiple regions or even the entire state.

Once these questions are answered, the evaluator can proceed to determining the specific methods to be used to estimate spillover.

Methodological Requirements

Central to the guidelines is the provision for two alternative levels of methodological rigor in the estimation of spillover, reflecting the fact that there are large variations (both across programs and across program administrators) in the likely magnitude of spillover savings and available resources. The two levels of rigor are referred to as standard and enhanced treatment. Each level of rigor addresses a range of methodological issues, including the overall methodological approach, the estimation of per-unit gross savings, estimation of program influence for end-users and upstream actors, and documentation of causal mechanisms. The specific requirements for each level of rigor are shown in Table 1.

Table 1. Level of Methodological Rigor for Estimating Spillover Savings and Program Influence

	Standard Rigor	**Enhanced Rigor**
Overall Methodological Approach	May rely solely on self-reports from end-users and upstream market actors to support estimates of gross savings or program influence.	Basic self-reports from end-users and upstream market actors typically not sufficient as sole method to support estimates of gross savings or program influence.
Estimation of average gross savings for spillover measures for end users (participants and/or nonparticipants).	Simplifying assumptions may be made, such as average gross unit savings being the same for spillover measures as for in-program measures.	Average gross unit savings for spillover measures must be documented empirically, based on a combination of self-reports and/or on-site visits.
Estimation of gross savings from upstream actors (participants and/or nonparticipants).	Self-reports generally sufficient.	Researchers must attempt to confirm self-reports using methods such as changes in sales, stocking or shipment data, review of planned or completed project or permits, or on-sites.
Estimation of program influence for end users (participants and/or nonparticipants).	Basic self-reports generally sufficient.	Enhanced self-reports generally sufficient[6].
Estimation of program influence for upstream actors (participants and/or nonparticipants).	Basic self-reports generally sufficient.	Either additional methods such as quasi-experimental design, econometric analysis, or Delphi panels[7] should be deployed or a case should be made that such methods are either not viable or not cost-effective.
Documentation of causal mechanisms	Recommended but not required.	Required, using methods such as self-reports from end-users or market actors regarding the manner in which the program influenced their behavior, and/or theory-driven evaluation practices. [8]

[6] Basic self-reports typically involve interviewing one participant decision-maker or market actor. Enhanced self-reports on the other hand typically involve more intensive data collection and analysis in the estimation of the net-to-gross ratios. For example, it can include collecting data from more than one participant decision-maker as well as from others such as relevant vendors, retailers, installers, architectural and engineering firms, and manufacturers. It can also include the consideration of past purchases and other qualitative data gleaned from open-ended questions.

[7] Delphi panels can be useful as long as members are provided sufficient market-level empirical data to inform their deliberations. Delphi panels should not be confused with brainstorming.

[8] Documentation of causal mechanisms can include verification of the key cause and effect relationships as illustrated in the program logic model and described in the program theory. Weiss (1997, 1998) suggests that a theory-driven evaluation can substitute for classical experimental study using random assignment. She suggests that if predicted steps between an activity and an outcome can be confirmed in implementation, this matching of the theory to observed outcomes will lend a strong argument for causality: "If the evaluation can show a series of micro-steps that lead from inputs to outcomes, then causal attribution for all practical purposes seems to be within reach" (Weiss 1997, 43).

Program administrators are asked to propose whether a given spillover analysis should receive standard or enhanced treatment. DPS staff reviews PA proposals from program administrators and makes a determination based on the value of the data balanced against the cost of the research. Factors to be considered in assigning a level of rigor include: (1) past results for the same PA program; (2) program theory or market operations theory; (3) national research literature for similar programs; (4) size of the program; (5) size and complexity of the market; and (6) nature of the technology(ies) promoted by the program.

Levels of Confidence and Precision

Several of the common methodological weaknesses discussed earlier in this paper stem, either directly or indirectly, from sample size limitations. The authors of the spillover guidelines were thus interested in including provisions that would help to encourage practitioners to develop samples that would provide reasonable levels of statistical precision, and limit the effects of leverage points.[9] After a fair amount of discussion, the approach that was ultimately adopted toward this end was to expand the confidence and precision standards that already existed in New York to include net savings. New York had long had a minimum standard of 90/10 confidence and precision for the measurement of gross savings. As part of the new spillover guidelines, practitioners must now also target this same level of confidence and precision for overall net savings at the program level. Here, overall net savings includes both in-program net savings and any reported net spillover savings.

If reported savings results include spillover savings, there is no required level of confidence and precision specifically for the individual components of net savings from in-program measures and net savings from spillover.[10] However, PAs are still accountable for achieving 90/10 for overall program net savings. The standard error of overall program-level net savings can be calculated by combining the achieved levels of confidence and precision for the net savings from in-program measures and for spillover savings using standard propagation of error formulas (Taylor, 2006; TecMarket, 2004).[11]

The intent of this provision was to give practitioners an incentive to pay attention to the effects of sampling error in the measurement of spillover, while avoiding the establishment of specific requirements for spillover samples that would not take into account the wide range of situations likely to be encountered. The underlying reasoning was that confidence and precision for overall net savings is typically a combined function of that for gross savings, free riding, and spillover. Practitioners could thus allocate their resources strategically to optimize the precision of this overall outcome, taking into account the likely relative magnitude and variability of each individual component. For example, if spillover had historically been very small for a particular program, it would make sense to allocate only limited resources to those samples

[9] This is not to downplay the importance of methodological issues regarding the fundamental reliability of self-reports, which are a measurement error issue. However, regardless of how well one does in designing survey instruments, it will never be possible to reliably quantify spillover effects if the sample is too small and/or poorly designed to yield decent precision and an absence of leverage points.

[10] While there are no precision *requirements* for the individual components of net savings, the precision actually achieved for each of these components must be reported at the 90% level of confidence, in order to help facilitate assessment of the reliability of the results.

[11] This is generally true as long as each of the individual components making up the total net savings estimate (e.g., gross savings, free riding, spillover, etc.) has been estimated based on independent random samples and methods that allow for the calculation of standard errors. However, there are legitimate circumstances under which the sample designs and methods for one or more components do not meet these requirements. One example is a market effects study in which total net program impacts are estimated using a preponderance of evidence approach. Another example is a case in which one or more components are deemed. A third example is a case in which multiple methods are used to estimate net impacts or the net-to-gross ratio, and a Delphi analysis is used to integrate the results. If *none* of the individual components meet these requirements, then clearly the issue of precision does not apply. If some components meet these requirements but others do not, then the program administrator is required to take clear note of this fact and propose an approach to ensuring that the components of the study that do meet these requirements are performed in a manner that gives due attention to limiting the effects of sampling error.

intended solely to support the estimation of spillover. Contrarily, if spillover had historically been large relative to gross savings and free riding, it would make sense to allocate more resources to this component. [12]

Double Counting

Lastly, the guidelines note the potential for double counting of spillover savings, emphasizing the potential specifically for double counting across program administrators. This is a particularly important aspect of double counting for New York due to the presence of a large number of program administrators with overlapping territories. PAs are asked to propose specific methods for avoiding double counting of both participant and nonparticipant spillover. Determining how the estimated spillover savings should be allocated among different programs within a given PA's portfolio and/or across PA portfolios can be based on such factors as the size of the program budgets, program theories and logic models that demonstrate the causal mechanisms that are expected to produce spillover, and the results of theory-driven evaluations.

CONCLUSIONS AND RECOMMENDATIONS

The application of the New York spillover guidelines is still in its infancy, and the specific effects the guidelines will have on evaluation practice in New York State are not yet well understood.[13] Further, the broader evaluation guidelines of which the spillover guidelines are a component are intended to be a living document, updated fairly often as experience with them accumulates. Lastly, the guidelines were written specifically to apply to the New York program administrators, who tend to be relatively large, and thus to have relatively large evaluation budgets. For these and other reasons, the authors do not advocate the New York spillover guidelines in their current form as being in any way the final word regarding nationwide methods for the evaluation of spillover.

Nonetheless, the guidelines may be useful for other states to adopt, whether in whole or in part. The common methodological weaknesses that the spillover guidelines were intended to address are widespread in the energy efficiency evaluation industry, and the authors would argue that addressing them in a systematic manner can be expected to improve spillover estimation practices over the long run. Specific aspects of the New York spillover guidelines that may be particularly broadly applicable include:

- Shifting the focus of statistical precision standards (where these exist) to overall net savings, with the use of propagation of error techniques to aggregate the effects of sampling error for individual components of net savings.
- Recognizing that the widespread assumption that spillover projects result in the same per-unit savings as in-program projects may result in upward bias.
- Acknowledging that the likely magnitude of spillover savings varies widely across programs, markets and program administrators, and that methodological standards for spillover estimation must take these variations into account.
- Recognizing the need to diversify spillover estimation methods to go beyond self-reporting.

[12] Requiring 90/10 confidence and precision for overall net savings is a demanding standard. Typically, meeting it will require a good statistical precision for each of the underlying components, including gross savings, free ridings, and spillover. Further, requiring 90/10 confidence and precision for overall net savings will effectively increase the confidence and precision that must be targeted from gross savings beyond the 90/10 level. (Adding the sampling error associated with free riding and spillover increases uncertainty, so if gross savings is estimated at only 90/10 confidence and precision, combining the gross savings estimate with sampling-based estimates of free riding and/or spillover will generally cause confidence and precision for overall (i.e., program-level) net savings to fall below 90/10.) All of this suggests that the effects of the 90/10 standard for overall net savings bear watching, and may be challenging to meet for smaller program administrators.

[13] However, for an early account of one spillover study designed to be responsive to the guidelines, see Wirtshafter et al., 2013.

- In self-reporting batteries, acknowledging the importance of avoiding question wording that has the potential to inadvertently lead the respondent toward attributing influence to the program.
- Recognizing the importance of documenting the specific causal mechanisms through which spillover savings are being generated, in order to enhance the credibility of spillover savings claims.

REFERENCES

California Public Utilities Commission: Energy Division and the Master Evaluation Contractor Team. (2007). *Guidelines for Estimating Net-To-Gross Ratios Using the Self-Report Approaches*.

Donaldson, Stewart I. (2007). *Program Theory-Driven Evaluation Science: Strategies and Applications*. New York: Psychology Press.

Eto, Joseph, Ralph Prahl and Jeff Schlegel. (1996). *A Scoping Study on Energy-Efficiency Market Transformation by California Utility DSM Programs*. Prepared for The California Demand-Side Measurement Advisory Committee: Project 2091T

Frederick D. Sebold, Alan Fields, Shel Feldman, Miriam Goldberg, Ken Keating and Jane Peters. (2001). *A Framework for Planning and Assessing Publicly Funded Energy Efficiency:Study ID PG&E-SW040.* Prepared for the Pacific Gas & Electric Company.

New York Department of Public Service. November 2012. "*Evaluation Plan Guidance for EEPS Program Administrators*." Update #3. Appendix F. Albany NY.

New York State Public Service Commission. June 23, 2008. Case 07-M-0548, "*Order Establishing Energy Efficiency Portfolio Standard and Approving Programs."* Albany NY.

Saxonis, William,"Free Ridership and Spillover: A Regulatory Dilemma," 2007 International Energy Program Evaluation Conference Proceedings, Chicago IL, 2007.

Taylor, John R. (1997). *An Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. Sausalito, California: University Science Books.

TecMarket Works Team. (2004). *The California Evaluation Framework.* Prepared for the California Public Utilities Commission and the Project Advisory Group, Framework.

The TecMarket Works Team. (2006). *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals*. Directed by the CPUC's Energy Division, and with guidance from Joint Staff.

Weiss, Carol H. (1997).Theory-Based Evaluation: Past, Present, and Future. In Debra J. Rog and Deborah Fournier (eds.) *Progress and Future Directions in Evaluation: Perspectives on Theory, Practice, and Methods*. San Francisco: Jossey-Bass Publishers.

Weiss, Carol H. (1998). *Evaluation*. Upper Saddle River, New Jersey: Prentice Hall.

Wirtshafter, Robert M, Jennifer Fagan, Bobbi Tannenbaum and Gregory French. 2013. "A Case Study of How a Market Transforming Program Claims Spillover." 2013 International Energy Program Evaluation Conference Proceedings, Chicago IL, 2013..