Gross Is Gross and Net Is Net: Simple, Right?

Richard Ridge Ph.D., Ridge & Associates, San Francisco, Calif.
Michael Baker, SBW Consulting, Bellevue, Wash.
Nick Hall, TecMarket Works, Oregon, Wis.
Ralph Prahl, Prahl & Associates, University Park, Fla.
William Saxonis, New York State Department of Public Service, Albany, N.Y.

ABSTRACT

The identification of the correct baseline is essential for accurately estimating the gross impacts of any measure. For example, for early replacement, the appropriate baseline is the energy use of the old equipment that was replaced. For replacement on burnout (ROB) or an addition, there are various baselines that could be used including applicable local, state and federal energy codes and standards (code and standards or C&S) with a compliance adjustment as necessary and the market average, current practice, or industry standard practice (hereafter referred to as current practice) to represent the energy use of equipment purchased on average by consumers in the market. Once gross impacts are estimated, the next step typically is to determine what portion of the gross impacts is caused by the program. In many cases, this is done by estimating a net-to-gross ratio (NTGR) and multiplying it by the gross impacts to yield net impacts.

However, some evaluators are now arguing that using current practice as the baseline for estimating gross impacts and then adjusting these savings using a NTGR is a mistake since the gross savings are in many ways closer to net than gross. While the authors agree that to refer to the difference between annual energy use associated with current practice and that of the rebated measure is not purely net savings, they disagree about which solutions to recommend. They note that what counts as credible evidence in a given jurisdiction may help to decide which way to proceed.

Introduction

Let's begin with the definitions of gross and net savings from the *California Energy Efficiency Evaluation Protocols: Technical, Methodological, and Reporting Requirements for Evaluation Professionals* (Hall et al. 2006).

> Gross Load Impact: The change in energy consumption and/or demand that results directly from program-related actions taken by participants in a DSM program, regardless of why they participated.
>
> Net Load Impact: The total change in load that is attributable to the utility DSM program.

For gross savings, the establishment of the correct baseline is critical. For example, for a participant who installed insulation in their home, the gross savings are defined as the difference between the annual energy use of the home without insulation and the annual energy use with insulation. Or, for an air conditioner, which is replaced when it fails, it is typically calculated as the difference between the annual energy use associated with an air conditioner that complies with applicable local, state and federal energy C&S and the annual energy use of the replacement air conditioner.

Establishing the Counterfactual

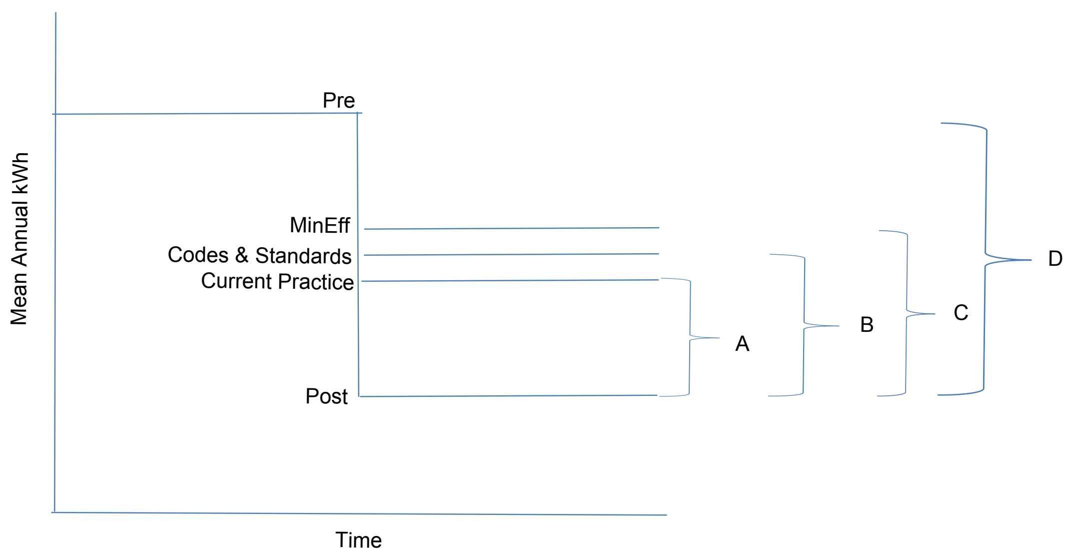
Of course, the ultimate aim of any evaluation of an energy efficiency program is to estimate the net savings due to the program intervention. Over the years, various research designs have been proposed to estimate net savings (Campbell and Stanley 1963; Mohr 1995). These designs all represent an effort to construct a counterfactual, i.e., what would have happened to the participants had they not participated in the program. The extent to which a given research design is able to establish a causal relationship between the intervention and the targeted outcome is referred to as internal validity. Internal validity is only relevant in studies that try to establish a causal relationship. It's not relevant in most observational or descriptive studies, for instance. But for studies that assess the effects of social programs or interventions, internal validity is perhaps the primary consideration. In those contexts, you would like to be able to conclude that your program or treatment made a difference -- it improved test scores or reduced energy use. But there may be lots of reasons, other than your program, why energy use might have changed. The key question in internal validity is whether observed changes can be attributed to your program or intervention (i.e., the cause) and not to other possible causes (sometimes described as "alternative explanations" for the outcome).

Various designs have been used to control for the threats to internal validity. In some cases, true experimental designs have been used which employ random assignment of customers to treatment and control groups. However, since such a design is rarely possible for energy efficiency programs, evaluators have often employed one of the many quasi-experimental designs that rely on existing groups of participants and nonparticipants such as the non-equivalent control group design. The customers in the control group in the true experimental design or the nonparticipants in a quasi-experimental design provide insights into what participants would have done in the absence of the program, the counterfactual.

For cases in which experimental or quasi-experimental designs are not possible, evaluators have relied on non-experimental methods consistent with the realist position (Weiss 1998; Weiss & Rein 1972; Cronbach 1982) to construct the counterfactual. An example of this approach is referred to as the self-report approach (SRA) and involves asking one or more key participant decision-makers a series of closed and open-ended questions about their motivations for installing the efficiency equipment and the efficiency of the equipment that they would have installed in the absence of the program. In addition, questions are typically asked to establish the temporal precedence of the program and to rule out rival explanations for the installation (Weiss 1998; Scriven 1976 and 2009; Shadish 1991; Wholey et al. 1994; Yin 1994; Mohr 1995; Rogers et al. 2000; Donaldson, Christie, & Mark 2008). The result is an index of program influence, called the net-to-gross ratio (NTGR) that reflects the percent of the gross savings that are attributable to the program.

Definitional Concerns

One of the more common approaches in the evaluation of energy efficiency programs is to first estimate the gross savings and then to adjust these savings using a net-to-gross ratio. However, recently, some confusion has arisen regarding the definition of gross savings and the application of the NTGR. The problem arises in situations in which billing analysis of participant energy use is inappropriate (e.g., pre-installation energy use is not the correct baseline as is the case of additions where no baseline is observable) or infeasible (e.g., when the signal-to-noise ratio is too small). In such cases, the problem is linked to how one defines the baseline for calculating gross savings. There are several ways to define the baseline depending on the installation situation (e.g., early replacement, replacement on burnout (ROB), and addition). Figure 1 illustrates four baselines that are commonly used to estimate gross savings. It also illustrates the four definitions of gross savings that might arise from comparing each baseline to the observed post-period energy use.



where

Pre=	Pre-installation mean annual energy of a representative sample of participants
MinEff=	Minimum efficiency available in the market
Code & Standards=	Efficiency level that is compliant with codes and standards (C&S)
Current Practice=	Mean annual kWh for a representative sample of measures purchased in the market, also known as common practice, current practice, or industry standard practice.
Post=	Post-installation mean annual kWh for the same random sample of participants

Figure 1. Four Baseline Scenarios

For many evaluators, the next step is to multiply these gross savings (A, B, C and D) by a NTGR in order to obtain net savings. Later in this paper, we will illustrate the problems with estimating savings using current practice as the baseline and then adjusting it with a NTGR. Table 1 presents each measure category, the appropriate baseline, and the associated gross savings consistent with Figure 1.

Table 1. Measure Category, by Baseline and Gross Savings

Measure Category	**Baseline**	**Gross Savings**
1. Any measure that is replaced before the end of its useful life or any non-energy using measure that is added to a site that reduces energy use. This category includes any energy using measure as well as non-energy using measures such as air conditions, motors, insulation, time clocks, VSDs, cool roofs, etc.	Pre	D
2. An energy using measure that is replaced on burnout or added that has no applicable energy C&S. This includes such measures such as air compressors.	Current Practice or MinEff	A or C
3. An energy using measure that is replaced on burnout or added that has an applicable C&S. This includes such measures as air conditioners.	C&S	B
4. New Construction or renovation	C&S	B

It is important to note that Figure 1 and Table 1 represent something of a simplification of actual baseline practices for ease of exposition. In actual practice, there are many gradations in the way baselines are established in the energy efficiency industry. For example, sometimes the way baselines are established for replacement on burnout situations may be closer to using the modal out-of-program practice (i.e., which single practice or measure do non-participants most often adopt) rather than the mean for the entire population. Similarly, for new construction, the C&S as baseline is often adjusted to reflect over- or under-compliance. We return to these variations later in the paper.

For additions of equipment such as time clocks and insulation or cases of early replacement, using pre as the baseline to estimate gross savings seems to make sense. The main focus of the remainder of this paper is on the confusion regarding the use of current practice and minimum efficient option as two possible baselines and how to interpret the savings they produce.

The Problem of Current Practice

The problem is that using current practice as the baseline produces savings that appear to be more consistent with the standard definition of net savings not gross savings. That is, if the participants had not participated in the program, they would have installed measures consistent with what similar customers in the larger market would on average have installed. As a result, if current practice is used as the baseline, an argument might be made that no further adjustments for free-ridership are needed. The goal of using of current practice as the baseline is the same as the traditional quasi-experimental designs in that they are attempting to determine the counterfactual in order to estimate net energy impacts. Theoretically, if one does a very good job of defining energy use based on current practice and calculates the difference between that energy use and the energy use of participants and then does a very good job of estimating a multiple regression model involving both participants and non-participants, one should get the same answer in terms of net energy savings. Some parties therefore argue that applying a NTGR to either of these two estimates represents an additional and unnecessary adjustment in savings.

There is some variation with respect to how different jurisdictions or regional bodies have addressed this issue. At least four jurisdictions that we know of consider the savings using current practice as the baseline to be net or close enough to net. The four are the Regional Technical Forum (RTF), the Northwest Energy Efficiency Alliance (NEEA), Indiana, and Delaware. RTF's *Guidelines for the Development and Maintenance of RTF Savings Estimation Methods* (RTF 2012) states:

> The current practice baseline defines directly the conditions that would prevail in the absence of the program (the counterfactual), as dictated by codes and standards or the current practices of the market. The most important conflict would arise if savings were estimated against a current practice baseline and then those savings were further adjusted by a net-to-gross ratio, where the net-to-gross ratio was the probability that the measure would have been delivered in the absence of program influence. (p. 2)

For measures for which there is an applicable C&S, the RTF assumes that customers in the market on average are meeting C&S and that the annual energy associated with the efficiency C&S represents current practice, i.e., the counterfactual baseline. However, if compliance is suspected to be less than or greater than 100 percent, then compliance studies can be conducted to provide a more accurate estimate of current practice. More will be said later about the C&S baseline and how other jurisdictions have interpreted the savings that result.

The *Evaluation Protocols for NEEA Commercial Sector Advice Initiatives* (Baker and Ridge 2011) states:

> NEEA's concept of causality involves trying to first establish the counterfactual—*what would have happened in the absence of the initiative*. A customer who installs an efficient measure will experience an annual reduction in energy use that is the difference between the annual energy use of the old measure and the new efficient measure. However, only the difference in the annual energy use of the new measure (promoted by the NEEA initiative) and the measure that they would have installed (in the absence of the NEEA initiative) can be claimed to have been caused by NEEA.
>
> At any point in time, customers are making decisions on equipment purchases, design features, or operational practices. The average efficiency that results from these decisions constitutes an estimate of what would have happened in the absence of NEEA's initiatives. This is referred to as the current-practice baseline in the *RTF Guidelines* and represents the counterfactual. The difference between the efficient equipment that NEEA promotes through its initiatives and the counterfactual (which varies by measure)

> constitutes the savings that NEEA has caused. Any additional adjustments, such as the application of a net-to-gross ratio, are unnecessary. (p. 13)

Two additional states, Indiana and Delaware, have also defined net savings in the same manner. In Indiana, the use of the standard market practice to estimate net savings is established within the State's *Evaluation Framework* (TecMarket Works 2012):

> The standard market practice (SMP) approach is a way to set energy impact analysis baselines so that the baseline already incorporates the influence of freeriders. In this approach a freerider assessment is not needed because the use of a standard market practice baseline is already what the market is doing without the program's direct influence. The SMP baseline is typically set at the mean of the level of energy efficiency being installed across the market being targeted by the program. (p. 55)

In Delaware, the State's *Evaluation Framework* also specifies the use of standard market practice approaches to identify net savings (Opinion Dynamics Corporation 2012):

> Because free riders are expected to take part in Delaware programs, a Net-to-Gross analysis will be completed for all programs in which free riders are expected, unless the evaluation approaches use experimental or quasi-experimental designs or set energy impact baseline conditions at standard market practice levels that lead directly to the estimation of net savings. (p. 24)

Some jurisdictions recognize that for many types of custom projects, there may be no common industry practice. In such cases, the project's baseline can be and often is unique to a particular project rather than a measure efficiency specification that can be broadly applied across multiple projects. For this baseline, one would have to determine what the customer is most likely to have done had s/he not installed the incentivized measure. Given this statement of the counterfactual, the difference between this baseline annual energy use and that of the efficient measure could also be interpreted as net savings. A recent evaluation of NYSERDA's New Construction Program used this approach, which they referred to as modeled partial net (MPN) (Megdal & Associates Impact Evaluation Team 2012). The MPN net savings reflect the savings that are attributable to the Program for each project. The MPN compares the as-operating energy use, which reflects the efficiency of the program measures and any inside spillover[1] that occurred at the project, to the project-specific baseline, which includes the free rider effects and is determined through energy analysis based on participant interviews.

Other jurisdictions, such as California, have taken a different position. The Energy Division of the California Public Utilities Commission has defined (Itron and KEMA 2012) what they call industry standard practice (ISP) as . . . typically adopted industry-specific efficiency levels that would be expected to be utilized absent the program" (p. 13-3). The difference between the annual energy use associated with ISP and the efficient measure is defined as gross savings, exactly the opposite of the other four jurisdictions.

[1] Inside spillover occurs when, due to the project, additional actions are taken to reduce energy use at the same site, but these actions are not included as program savings.

What Is An Evaluator To Do?

The authors of this paper are in agreement that ignoring the issue of current practice raised in this paper is a serious mistake, but disagree on the solution. Some are in agreement with NEEA, the RTF, Indiana and Delaware that, while not perfect, the difference between current practice and the efficient measure/building is sufficiently net and that to spend limited evaluation funds pursuing a more refined answer would produce an estimate that would still contain a fair amount of uncertainty. Others feel that the energy use associated with current practice and the rebated measures are so contaminated as to be unreliable and argue that one solution would be to use the minimum efficient option available in the market as the baseline.

Concerns

Both positions on how to address current practice have a number of problems worth examining more closely. First, all recognize that there are problems with using current practice. The nonparticipant energy use baseline might be lower (use less energy) than it normally would have due to the fact that for many years some utilities have been running energy efficiency programs some of which are primarily designed as market transformation programs. As a result, these programs may well have generated nonparticipant spillover, some of which is permanent (i.e., market effects), which was not captured in prior evaluations. As a result, the difference between the energy use of the efficient measures purchased by participants and the energy use of the measures adopted by customers in the larger market will be biased downwards. Note that if the market is relatively small as in pump-off controllers used in oil fields, using market average can contain a fair number of past participants (e.g., end users, installers and distributors) who have already been influenced by the program. The effect is to lower more dramatically the annual energy use of the measures that constitute the current practice baseline resulting in an estimate of gross impacts that is even more severely biased downwards.

The methods and costs to establish current practice can also vary widely and affect the reliability of the estimates. In determining what constitutes common practice, the assessment needs to focus on the equipment choices and installation configurations would have normally been adopted in the absence of the program. Evidence for the current practice baseline can include discussions with designers and/or vendors familiar with the process affected by the measure (e.g., interviews with wastewater treatment plant engineers to determine whether VFDs are common practice on wastewater aerators). It could also include a review of available shipment or sales data or interviews of similar nonparticipating customers. Each approach has its own limitations. For example, shipment or sales data can be very expensive to obtain and the data may not be representative since some manufacturers, distributors and retailers might not be willing to release their data. Or, surveys of nonparticipants who have recently purchase a particular measure must be representative, achieve a reasonable level of confidence and precision, address non-response bias, and reliably associate each claimed purchase with a model number and unit energy consumption. The sources of error and how they propagate through the calculations are not trivial.

To complicate matters further, there is the fact that participants self-select (Rossi, Lipsey and Freeman 2004) into any social action program, including energy efficiency programs. That is, those who decide to participate in a program may be predisposed to purchasing the efficient equipment prior to participation. Unless this is accounted for in the calculation, the net savings will be overestimated[2]. Unfortunately, given the calculation, there are no reliable methods to statistically correct for this bias.

Thus, both points of comparison contain some measurement error, the magnitude of which would be difficult if not impossible to estimate and control for. Furthermore, for such an approach to work requires that the estimate of current practice be valid and reliable, which can be rather expensive if one attempts to do this for multiple measure or measure groups within a given program.

The adoption of the minimum efficient option available in the market has its own problems. Estimating this baseline faces the same problems discussed earlier in estimating current practice. However, there is another potentially significant source of error since a NTGR must be estimated to convert these gross savings into net savings. Such methods as the self-report approach (SRA) and nested logit models have often been used to estimate NTGRs. The SRA involves asking one or more key participant decision-makers a series of closed and open-ended questions about their motivations for installing the efficiency equipment, about whether they would have installed the same energy efficient equipment in the absence of the program, to establish the temporal precedence of the program, as well as questions that attempt to rule out rival explanations for the installation (Weiss 1972; Scriven 1976; Shadish 1991; Wholey et al. 1994; Yin 1994; Mohr 1995; Rogers et al. 2000; Donaldson, Christie, & Mark 2008). While there are various methods available to mitigate various sources of error associated with the SRA (Ridge, Fagan and Willems 2009), some evaluators are skeptical that some keys sources of error can be adequately controlled (Peters and McRae 2008). One could use the nested logic approach to estimate a NTGR. This involves estimating two models, one that predicts the probability of participation in the program and one that predicts the probability of installing an efficient measure. Such an approach can be very data intensive, and if under-budgeted, cannot be counted upon to provide a reliable estimate of the NTGR. Primarily, it is the verification of the nonparticipant efficiency levels that both drive up the budget and introduce error.

Conclusions

While the authors agree that to refer to the difference between annual energy use associated with current practice and that of the rebated measure is not truly net savings, they disagree about how to proceed. One group is satisfied that it is sufficiently close to net to be accepted as net. Another group is worried that both points of comparison are so contaminated with measurement error that the difference between the two is so ambiguous as to be essentially useless.

Ultimately, what one does is affected by the policies in a given jurisdiction regarding the required level of accuracy, confidence and precision, i.e., what counts as credible evidence will vary. We have identified four jurisdictions and regional bodies that have accepted this difference as net or at least a sufficiently reliable estimate of net. Once they are alerted to this problem,

[2] As suggested earlier, there are gradations in the way baseline assumptions are established in actual practice. It is possible that some of these gradations may help to mitigate the problems associated with interpreting the difference between participant and current practice without raising other difficulties. For example, from time to time the authors have seen the concept of "current practice" be operationalized as the modal action taken by non-participants. While the modal action taken by non-participants is itself a measure of central tendency, one might argue that it is less likely than the mean of the entire population to result in a meaningful adjustment in the direction of net savings.

other jurisdictions may agree with these four, or they may decide that the only alternative is to use MinEff baseline to estimate gross and that existing methods of estimating NTGRs are sufficiently reliable to adjust these gross savings. While different jurisdictions may develop different solutions to this problem, we argue that they can no longer ignore it since the potential for getting it wrong is enormous.

Something that could be done very quickly would be to survey additional jurisdictions and regional bodies to assess whether and under what conditions they use current practice as the baseline and how they interpret the results. It would also be useful to know the extent to which the various interpretations of these savings have become controversial and how they are proposing to settle the issue.

In the meantime, for those who have yet made up their minds, more research could be done to assess the magnitude of the problem and propose other solutions. For example, one could estimate net impacts using experimental or quasi-experimental designs and compare the resulting estimates of net savings to the savings using the current practice baseline. Note that for this to be a fair and, therefore, useful exercise, both approaches must be methodologically rigorous. If the results are similar, then one might be more willing to consider the difference in energy use between current practice and the efficient measure as net savings. On the other hand, if the results using the current practice baseline are much larger, then one might be more inclined to consider these results as gross or at least somewhere between net and gross. Of course, replicating the results of this experiment multiple times would reduce the uncertainty even more. An investment in such research would be well worth it if it could settle this important and emerging problem.

References

Baker, Michael and Richard Ridge. 2011. *Evaluation Protocols For NEEA Commercial Sector Advice Initiatives.* Prepared for the Northwest Energy Efficiency Alliance.

Campbell, Donald T. and Julian Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Boston, MA: Houghton Mifflin Company.

Cronbach, L. J. 1982. *Designing Evaluation and Social Action Programs*. San Francisco: Jossey-Bass.

Donaldson, Stewart I, Christina A. Christie and Melvin Mark (Eds.) 2009. *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Hall, Nick, Johna Roth, Carmen Best, Sharyn Barata, Pete Jacobs, Ken Keating, Ph.D., Steve Kromer, Lori Megdal Ph.D., Jane Peters, Ph.D., Richard Ridge, Ph.D., Francis Trottier, and Ed Vine, Ph.D. 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.

Megdal & Associates Impact Evaluation Team. 2012. *New Construction Program (NCP) Impact Evaluation Report for Program Years 2007 – 2008: Final Report*. Prepared for The New York State Energy Research and Development Authority.

Mohr, Lawrence B. 1995. *Impact Analysis for Program Evaluation*. Thousand Oaks, CA: Sage Publications, Inc.

Opinion Dynamics Corporation, TecMarket Works, Building Metrics, and Nexant. 2012. *Delaware Evaluation Framework*. Prepared for Department of Natural Resources and Environmental Control.

Peters, Jane and Marjorie McRae. 2008. "Free-Ridership Measurement If Out of Sync with Program Logic . . . or, We've Got the Structure Built, but What's Its Foundations?" In the *Proceedings of the 2008 ACEEE Summer Study on Energy Efficiency in Buildings*.

Randazzo, Katherine and Richard Ridge. 1996. "How Many Mills Ratios Does It Take to Estimate Net Impacts?" *In the Proceedings of the 1996 ACEEE Summer Study on Energy Efficiency in Buildings*.

Regional Technical Forum. 2012. *Guidelines for the Development and Maintenance of RTF Savings Estimation Methods.*

Ridge, Richard, Phillipus Willems, Jennifer Fagan and Katherine Randazzo. 2009. "The Origins of the Misunderstood and Occasionally Maligned Self-Report Approach to Estimating the Net-To-Gross Ratio". *In the Proceedings of the International Energy Program Evaluation Conference in August.*

Rogers, Patricia J., Timothy A. Hacsi, Anthony Petrosino, and Tracy A. Huebner (Eds.) 2000. *Program Theory in Evaluation: Challenges and Opportunities*. San Francisco, CA: Jossey-Bass Publishers.

Rossi, Peter H., Mark W. Lipsey and Howard E. Freeman. 2004. *Evaluation: A Systematic Approach*. Thousand Oaks, CA: SAGE Publications.

SBW & Ridge & Associates. 2011. *Evaluation Protocols for NEEA Commercial Sector Advice Initiatives.* Prepared for the Northwest Energy Efficiency Alliance.

Scriven, Michael. 1976. *Maximizing the Power of Causal Explanations: The Modus Operandi Method*. In G.V. Glass (Ed.), *Evaluation Studies Review Annual* (Vol. 1, pp.101-118). Beverly Hills, CA: Sage Publications.

Scriven, Michael. 2009. Demythologizing Causation and Evidence. In Stuart I. Donaldson, Christina A. Christie, and Melvin Mark (Eds.). *What Counts as Credible Evidence in Applied Research and Evaluation Practice?* Los Angeles, CA: SAGE Publications.

Shadish, Jr., William R. and Thomas D. Cook, and Laura C. Leviton. 1991. *Foundations of Program Evaluation*. Newbury Park, CA: Sage Publications, Inc.

TecMarket Works, The Cadmus Group, Opinion Dynamics Corporation, Integral Analytics, Building Metrics, and Energy Efficient Homes Midwest with support from Vickie Benson and Maria Larson. 2012. *Indiana Evaluation Framework*. Prepared for the Indiana Demand Side Management Coordination Committee.

Weiss, R. S. and M.Rein. 1972. The Evaluation of Broad-Aim Programs: Difficulties in Experimental design and an Alternative. In C. H. Weiss (ed.) *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn and Bacon.

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.

Wholey, Joseph S., Harry P. Hatry and Kathryn E. Newcomer. 1994. *Handbook of Practical Program Evaluation*. San Francisco, CA: Jossey-Bass, Inc.

Yin, Robert K. 1994. *Case Study Research: Design and Methods*. Thousand Oaks, CA: SAGE Publications.

Itron, Inc. and KEMA, Inc. 2012. *Evaluation of the 2010-2012 Custom Impact Energy Efficiency Programs for the California IOUs: Procedures for Site-Specific Impact Analysis Under Work Order WO033.* Prepared for California Public Utilities Commission (CPUC).