

Evaluation, Measurement, and Verification (EM&V) for Behavior-Based Energy Efficiency Programs: Issues and Recommendations

*Annika Todd, Elizabeth Stuart, and Charles Goldman, Lawrence Berkeley National Laboratory¹
Steven Schiller, Schiller Consulting*

ABSTRACT

Behavior-based programs have been identified as a major potential source of energy savings and are increasingly being adopted by energy efficiency program administrators nationwide. However, because these programs often lack a technology-based unit for which savings can be modeled or deemed (such as a CFL), documentation of energy savings requires approaches which are common in the experimental sciences, but not in the efficiency industry. In this paper (based on a report created for the State and Local Energy Efficiency Action Network), we describe and discuss eleven key evaluation issues and recommend analysis approaches that can be used to define whether a behavior-based program evaluation of energy savings can be considered rigorous and internally valid. In particular, we recommend using randomized controlled trials (RCTs) for behavior-based efficiency programs, which result in robust, unbiased program savings impact estimates. We also discuss issues relating to the double counting of savings, and the importance of third party evaluators. With respect to external validity, we discuss conditions under which impact estimates from behavior-based programs can be applied to different populations in future years. We recommend that a control group that is representative of all of the different participating populations should be maintained for every year in which program energy estimates are being used to claim savings. Finally, we discuss challenges in moving towards estimating savings based on a calibrated predictive analytic model, which could be used to produce deemed savings estimates at some point in the future.

Introduction

Historically, residential energy efficiency programs have typically used strategies such as rebates, other financial incentives (loans), and technical assistance/information (e.g. audits) to motivate consumers to install technologies and high efficiency measures in their homes. During the last several years there has been increasing interest in broadening residential energy efficiency program portfolios to include behavior-based programs that utilize strategies intended to affect consumer energy use related behaviors in order to achieve energy and/or peak demand savings. These programs typically include outreach, education, competition, rewards, benchmarking and/or feedback elements. In some cases, this new generation of programs takes advantage of technological advances to both capture energy data at a higher temporal and spatial resolution than ever before, and to communicate the energy data to households in creative new ways that leverage social science-based motivational techniques.

The trend of incorporating behavior-based programs into the portfolio of energy efficiency programs stems from a desire to capture all cost-effective energy efficiency resources as some pilot behavior-based programs have been shown to be cost effective as compared to

¹This paper is based on: State and Local Energy Efficiency Action Network 2012, available at behavioranalytics.lbl.gov.

supply-side alternatives (Allcott and Mullainathan 2010). Some of the obstacles to their widespread adoption relate to whether these programs can be evaluated in a rigorous way. In this paper we provide recommendations for eleven key evaluation issues and analysis factors that specifically address internal validity and also touch briefly on external validity. We make recommendations for each factor using a star-based ranking system, highlighting best (5-star) and better (3- or 4-star) approaches. These recommended evaluation methods need to be rigorous because large-scale behavior-based energy efficiency programs are a relatively new strategy in the energy efficiency industry and savings per average household are small.

Recommendations for Internal Validity of Estimated Savings

Internal validity of savings estimates refers to whether the impact we estimate for the given population during the initial time period was caused by the program as opposed to other factors. Methods and best practices for ensuring internal validity are well established, and are currently being used for a number of behavior-based programs in order to claim savings.

Evaluation Design: Validity of the Randomized Controlled Trial Method

The true energy savings from an energy efficiency program cannot be measured (because we can never know how much energy households in the program would have saved had they not been in the program). Therefore, energy savings must be estimated by measuring the difference between the energy use of the households participating in the program (the “treatment group”) relative to the energy use of a comparison group of households that we consider similar to those in the participant households (the “control group”) during the same period of time. The difference between the energy use of the households in the treatment and the control group can be attributed to three sources: (1) the true impact of the program; (2) the pre-existing differences between households in the treatment and control group, which is called “bias” or “selection bias”; and (3) inherent randomness. A good estimate of energy savings is one which eliminates or minimizes the second source (i.e., is unbiased) and the third source (i.e., is precise) so that the estimate is as close as possible to the true savings.

The way in which a control group is constructed and compared to the treatment group in order to estimate the program savings impacts (i.e., the “evaluation design”), is the most important factor in creating estimates of program impacts that are unbiased and internally valid. We recommend using randomized controlled trials (RCT) that will result in robust, statistically unbiased estimates of program energy savings. For energy efficiency programs, an RCT randomly assigns households into a treatment or control group, creating a control group that is statistically identical to the treatment group. This eliminates selection bias and allows evaluators to calculate an unbiased estimate of savings (Angrist and Pischke 2008; Duflo, Glennerster, and Kremer 2007, Imbens and Wooldridge 2009; LaLonde 1986). Using an RCT is a key initial step in ensuring the validity of estimates of program savings for behavior-based efficiency programs.

RCTs address two specific types of selection biases. First, concerns about *free-riders* (i.e., the households in the program that would have taken actions to save energy even in the absence of the program) are completely eliminated because the treatment and control groups each contain the same number of free-riders through the process of random assignment to the treatment or control groups. This is one of the main benefits of an RCT over other evaluation methods. Second, *participant spillover*, in which participants engage in additional energy efficiency actions outside of the program as a result of the program, is also automatically







captured by a RCT design for energy use that is measured within a household. An RCT design produces an estimate of net energy savings and thus also addresses *rebound effects* or *take-back* during the study period, which can occur if consumers increase energy use as a result of a new device’s improved efficiency.

If RCTs are not feasible, we suggest using approaches such as “regression discontinuity” (which compares households on both sides of an eligibility criterion), “variation in adoption” with a test of assumptions (which compares households that will decide to opt-in soon to households that already opted-in), or a “propensity score matching” approach (which compares households that opt-in to households that didn’t opt-in but were predicted to be likely to opt-in and have similar observable characteristics). Because they use control groups that are not randomly assigned, these are called “quasi-experimental” methods, and are less robust and possibly biased as compared to RCTs (e.g., quasi-experimental methods can over- or under-state energy savings by 200% or more; Allcott 2011).

We do not recommend other quasi-experimental methods for behavior-based efficiency programs including “non-propensity score matching” (which compares households that opt-in to households that didn’t opt-in and have some similar observable characteristics), and “pre-post comparison” (which compares households after they were enrolled to the same households before they were enrolled) because these methods use control groups that are less likely to be similar to the treatment groups and are therefore more biased.

We rank these methods according to their level of bias; more stars is a better method and indicates less bias; while fewer stars indicate that the method yields estimates of program savings impacts that are likely to be more biased. We also provide a real-world example of a program that implemented a RCT.

Table 1. Evaluation Design: Recommendation²

	Randomized Controlled Trial results in unbiased estimates of savings
	Regression Discontinuity results in estimates of savings likely to be unbiased
	Variation in Adoption could result in biased estimates of savings
	Propensity Score Matching could result in biased estimates of savings
	Non-Propensity Score Matching could result in biased estimates of savings
	Pre-Post Comparison could result in very biased estimates of savings

Length of Study and Baseline Period

Program savings should be estimated by taking the difference between the change in energy use (i.e., the energy used before the program less the energy use after the program is implemented) by the households in the treatment group and the change in energy use by the households in the control group during the study period. In order to estimate the energy saved by households in both groups, their energy use during the program should be compared to their baseline energy use in the time period immediately prior to the program’s implementation.

² Note that these rankings assume that the evaluation design was performed correctly.

Relatively longer study periods and baseline data periods are likely to lead to greater precision of the estimated program impact because patterns of household energy use often varies by season. Thus, it is strongly advised that at least one full year (the twelve continuous months immediately prior to the program start date) of historical energy use data be available for each customer – both for those in the treatment and in the control group – so that the baseline energy use reflects seasonal effects. If an RCT design is used, evaluations that collect less than one year of historical data will still yield unbiased estimates of energy savings. For non-RCT evaluation methods, failure to collect twelve months of historical data can result in biased estimates of energy savings that are inaccurate and thus not advised. See Table 2 for our recommendations.

Table 2. Length of Baseline Period: Recommendation

If RCT:	If Quasi-Experimental:	
★★★★★	★★★★★	Twelve months or more of historical data collected ³
★★★★☆	★Not Advisable★	Less than a complete twelve months of historical data collected
★★★★☆	★Not Advisable★	No historical data collected

Avoiding Potential Conflicts of Interest

Evaluations of behavior-based efficiency programs should be managed in a way that produces the least potential for a conflict of interest to arise regarding the validity of savings estimates. This is particularly important if the evaluation being undertaken is intended to inform cost recovery or payment of incentives. In other situations (e.g., when a program administrator is testing program marketing or design concepts or conducting a small pilot that involves technology demonstration and application), independent third party evaluators may not be necessary to test preliminary program theories.

In particular, we recommend that the assignment of households to control and treatment groups (whether randomly assigned or matched) is performed by a third party. We do this because we believe that the temptation of making slight changes in the control or treatment groups in order to increase the reported savings levels is too great (i.e., the savings could be severely biased by adding or subtracting a carefully selected group of households such as those with high energy usage, with college bound teenagers, or those that have been proven in prior evaluations to show high or low energy savings). Table 3 lists our additional recommendations for avoiding potential or perceived conflicts of interest in estimating program impacts.

³ If efficiency programs are designed to reduce usage only during a specific season (e.g. summer) then only data from that season is necessary.

Table 3. Avoiding Potential Conflicts of Interest: Recommendation

	An independent, third-party evaluator ⁴ transparently defines and implements:
★★★★★	<ul style="list-style-type: none">• Program evaluation• Assignment of households to control and treatment groups• Data selection and cleaning, including identification and treatment of missing values, outliers, and account closures, and the normalization of billing cycle days
★Not Advisable★	Program implementer or sponsor implements any of the above

Analysis Model

The analysis model is the set of algorithms used to estimate energy savings through engineering and/or econometric techniques such as regression analysis. Three basic analysis model specification options potentially affect the accuracy and precision of savings estimates: (1) whether the model uses panel data (many energy data points over time) or data that is aggregated over time; (2) whether the model compares energy usage or *the change* in energy usage; and (3) if the model includes extra control or interaction variables or not. If an RCT evaluation design is utilized, then all of the models will yield savings estimates that are unbiased if they comply with the recommendations in this paper (with the exception of models that include interaction variables), although some are likely to be more precise than others. If quasi-experimental evaluation methods are used, then some model specifications will likely result in savings estimates that are less biased than others and some models will be more precise than others. See Table 4 for our recommendations.

⁴ Each jurisdiction may define its own criteria for “third-party” and “independent” but generally it is considered to be an entity without a financial or related interest in the outcome of the evaluation in terms of how much energy was saved.

Table 4. Analysis Model Specification Options: Recommendation

If RCT:	If Quasi-Experimental:	
★★★★★	★★★★★	Panel Data Model with Fixed Effects (comparing change in use), with or without Control Variables, with a primary analysis that does not include Interaction Variables ⁵
★★★★★	★★★★★	Time Aggregated Data Model, with or without Control Variables, with a primary analysis that does not include Interaction Variables
★★★★★	★★★★★	Model comparing use (not <i>change</i> in use), with a primary analysis that does not include Interaction Variables
★ Not Advisable	★ Not Advisable	Any Model with a primary analysis that includes Interaction Variables

Cluster Robust Standard Errors

Any panel data model (e.g., monthly data points for the pre- and post-program periods) must use standard errors that are *cluster robust at the unit of randomization*. Failure to do so results in biased measures of precision that appear to be much more precise than they are in reality. Clustering standard errors mean that the analysis accounts for the fact that 12 months of energy use data from one household is not the same as one month of energy use data from 12 households. For an example in which the precision is inflated by more than double, see Bertrand, Duflo and Mullainathan (2004). The unit of randomization is the level at which households were randomly allocated into a control or treatment group. Typically, the level of randomization is the household; thus, standard errors should be clustered at the household level (see Table 5).

Table 5. Clustered Standard Errors: Recommendation

★★★★★	Cluster Robust Standard Errors or Time Aggregated Data
★ Not Advisable	Non-Cluster Robust Standard Errors with non-Time Aggregated Data




Equivalency Check

An important part of the analysis is validating that the control and treatment groups are equivalent. This is because the degree to which a savings estimate is unbiased depends on the similarity of the groups. Validating is done by testing whether households in the treatment group have characteristics that are statistically similar to those in the control group (also called a

⁵Control variables help explain the patterns of energy use unrelated to the program, whereas interaction variables provide insights as to the relationships between the program and other factors. Interaction variables should not be included in the primary analysis that assesses the overall program impact, but could be included in secondary analyses. If necessary from a financial or regulatory standpoint, the primary analysis could include time-based, dummy interaction variables in addition to a model that does not include interaction variables.

balanced treatment/control check or a randomization check if the method is an RCT). Evaluators should use professional judgment to decide what characteristics need to be tested. Possible tests include monthly or yearly pre-program energy use, load profiles, distribution of pre-program energy use, geographic location, dwelling characteristics, demographic characteristics (e.g., age, income), psychographic characteristics (e.g., opinions) and any other baseline variables that may affect the household’s response to the program for which data are available. This should be done whether the program is designed as an RCT or a quasi-experiment (see Table 6).⁶

Table 6. Equivalency Check: Recommendation

	An equivalency check is performed with household energy usage profiles as well as demographic, geographic, and other household characteristics
	An equivalency check is performed with household energy usage profiles
	An equivalency check is not performed

Statistical Significance

An estimate of program impact savings should not be accepted if it is not precise enough. Stated another way, the savings estimates are too risky to accept if there is too big a chance that the true program savings do not satisfy the required threshold level (e.g., a risk that the savings are not greater than zero or that they are not sufficient to support a cost effectiveness screening requirement). To ensure a level of precision that is considered acceptable in behavioral sciences research, a *null hypothesis* (i.e., a required threshold such as the level or percent of energy savings needed for the benefits of the program to be considered cost effective) should be established, and the program savings estimate should be considered acceptable (and the null hypothesis should be rejected) if the estimate is statistically significant at the 5% level or lower (i.e., provides 95% confidence). For example, if the desired test is whether or not a program’s energy savings is greater than zero, the null hypothesis would be that the energy savings are *not* greater than zero. Then if the savings estimate is statistically significant at 5% (or lower), it essentially means that there is only a 5% (or less) chance that the savings are not greater than zero (and a 95% chance that the savings are greater than zero). Because it is less than 5%, the savings estimate should be considered acceptable.⁷

⁶ With RCTs, one option to ensure that the randomization is balanced is to perform multiple randomizations (e.g., 1000), do an equivalency check for each one, and then choose the randomization that is the most balanced (e.g., that has the smallest maximum t-statistic out of all of the compared baseline covariates).

⁷ Note that these recommendations apply to the measurements of savings, and have nothing to do with the sample size selection requirement typically referenced in energy efficiency evaluations. Sample size selection is usually required to have 10-20% precision with 80-90% confidence and may be referred to as “90/10” or “80/20”. A 5% level of statistical significance does not mean “95/5”.

Table 7. Statistical Significance: Recommendation



An estimate that is statistically significant at 5% should be accepted and the null hypothesis (or required threshold) should be transparently defined.



An estimate that is statistically significant at >5% should not be accepted

Excluding Data from Households that Opt-out or Drop Out

Typical ways that study populations are segmented are by excluding those who opt-out of the program or those who closed their accounts. Households that opt-out should never be excluded from the dataset; they should be included as part of the treatment group to avoid selection bias. If households that opt-out of a program are excluded, then the treatment group no longer contains the same types of households as the control group (because these people can't be excluded from the control group; see Table 8). The report on which this paper is based has additional information on calculating an unbiased estimate of the effect of the program on those that did not opt out.

Table 8. Excluding Data from Households that Opt-out or Drop Out: Recommendation



Only data from households that closed accounts are excluded*; households that opt-out of the treatment or control group are included in the analysis (although the program impact estimate may be transformed to represent the impact for households that did not opt-out, as long as it is transparently indicated).



Data from households that closed their accounts are included*



Households that opt-out are excluded from the analysis

**If there is a compelling reason to include households that closed their accounts and an analysis is undertaken correctly to deal with unbalanced data sets, then it may be advisable.*

Accounting for Potential Double Counting of Savings

In many states, behavior-based efficiency programs are offered in an environment where the administrator already has many other residential efficiency programs. Thus, the evaluation questions are often framed in terms of how much additional savings are gained from behavior-based programs and at what program cost. In this environment, there is the possibility that more than one program could claim savings from installation of the same measure; thus program administrators, evaluators and regulatory staff need to address issues related to potential “double counting” of savings (e.g., a CFL rebate program, an education program, and a behavior-based program might all claim savings for installation of CFLs; see Table 9 for recommendations).

One of the advantages of a behavior-based efficiency program that is evaluated with a treatment and control group is that it provides a method for at least partial accounting for this phenomenon. This is because “double counted” savings are equal to the amount of savings

Table 9. Accounting for Potential Double Counting of Savings: Recommendation

If RCT:	If Quasi-Experimental:	
★★★★★★★☆☆	★★★★★★★☆☆	<p>Double counted savings:</p> <ul style="list-style-type: none"> •Are rigorously estimated for programs in which efficiency measures can be tracked to specific households; and •Do not exist or a compellingly rigorous estimation approach was used for programs in which efficiency measures cannot be tracked; and •The measurement period (e.g., accounting for seasonal load impacts), and the effective useful lifetime of installed measures (when lifetime savings are reported) are taken into account; and •Program costs are appropriately allocated along with double counted savings
★★★★★☆☆☆☆	★★★★★☆☆☆☆	<p>Double counted savings:</p> <ul style="list-style-type: none"> •Are rigorously estimated for programs in which efficiency measures can be tracked to specific households; and •Attempt to be estimated for programs in which efficiency measures cannot be tracked; and •The measurement period (e.g., accounting for seasonal load impacts), and the effective useful lifetime of installed measures (when lifetime savings are reported) are taken into account; and •Program costs are appropriately allocated along with double counted savings
★★★☆☆☆☆☆☆	★★★☆☆☆☆☆☆	<p>Double counted savings:</p> <ul style="list-style-type: none"> •Are rigorously estimated for programs in which efficiency measures can be tracked to specific households; and •The measurement period (e.g., accounting for seasonal load impacts), and the effective useful lifetime of installed measures (when lifetime savings are reported) are taken into account; and •Program costs are appropriately allocated along with double counted savings
☆☆☆☆☆☆☆☆	☆☆☆☆☆☆☆☆	<p>Double counted savings are not documented</p>

claimed by the ‘other program(s)’ for households in the treatment group minus the amount of savings claimed by the ‘other program(s)’ for households in the control group. For example, assuming that the control and treatment groups have the same number of households, if customers in the treatment group used 100 more refrigerator rebates than customers in the control group and each high-efficiency refrigerator is estimated to save 50 kWh per year, then the incremental “double counted” savings are 5,000 kWh/year. For programs in which

participation can be tracked to a specific household (e.g., installation of insulation by a contractor), this can be directly determined. For programs in which efficiency measures cannot be tracked to specific households (e.g., upstream CFL rebates), customer surveys can be used to estimate measures installed by customers through that upstream efficiency program.

When estimating the amount of double counted savings, it is also important to take into account differences between programs in the measurement period (e.g., seasonal load impacts), and the effective useful lifetime of installed measures (when lifetime savings are reported); (Wilhelm and Agnew 2012).

For planning purposes or as part of cost-effectiveness screening or as part of the contract between a program administrator and implementation contractor/vendor, it may also be necessary to address issues related to attribution of savings to specific programs (i.e., which program – behavior-based efficiency program or another existing efficiency program – induced the customer to install these measures). There are several ways in which the “double counted” savings may be allocated. Any approach should include appropriately allocating program costs in addition to the savings (e.g., factors such as program expenditure, incentives, administrative costs, consumer costs, measure life, and program strategy). Because the program and evaluation design of the behavior-based program utilizes a treatment and control group, we can infer that the “double counted” savings were caused by the behavior-based program and that it is therefore reasonable to assign at least half of the “double-counted” savings to the behavior-based efficiency program (while also appropriately assigning program costs).⁸ However, we refrain from recommending any particular assignment of savings between programs, but rather we recommend the transparent identification of the magnitude of the double counted savings when participation in the other programs can be tracked to a specific household.⁹

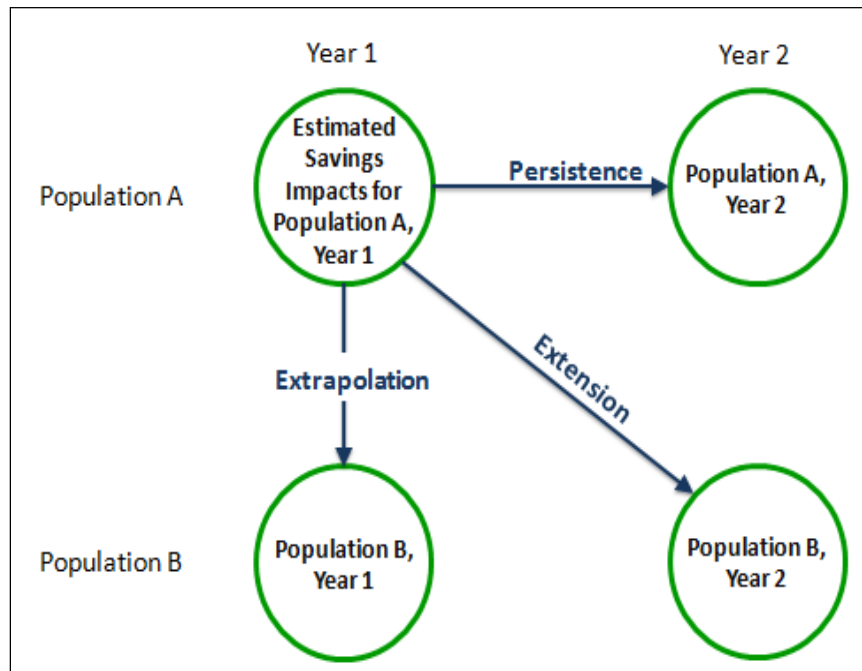
Recommendations for External Validity: Applying Impact Estimates to Different Populations and Future Years

Program implementers will often be interested in expanding a program both over time and to other populations. In this section we assess whether the estimated savings for the initial program can be generalized and applied to the new populations and future years (commonly referred to as *external validity*). We examine whether a valid program savings impact estimate for a given population (population A) in year 1 of a behavior-based energy efficiency program can be (1) extrapolated to population B that also participates in the program in year 1; (2) used to estimate savings in future years (e.g. second and/or third year) for the given population (i.e., persistence of savings); or (3) applied and extended to a new population B in a future year (e.g., a pilot program is rolled out to more households in year 2). Figure 1 illustrates this concept of external validity. In contrast to methods for ensuring internal validity, methods for applying behavior-based program savings estimates to new populations and future years in an accurate way that ensures external validity are not well established (Allcott and Mullainathan 2012).

⁸ If households in the treatment group claim more rebates than those in the control group, then it must be true that the behavior-based program is causing those extra rebates (i.e., the behavior-based program is a “necessary” condition). Because the rebate program is not implemented with a treatment and control group, we don’t know if the rebate program is also causing the extra rebates (i.e., the rebate program may or may not be a “necessary” condition; these treatment households may have purchased the energy efficient equipment with or without the rebate).

⁹ The double counted savings could be entirely allocated to behavior-based programs, and the respective incentive, administrative costs, consumer costs, and implementation costs of the marginal installed measures could be attributed the behavioral program. Or the double counted savings might be entirely allocated to the non-behavior-based programs, while behavior-based programs are compensated for respective marketing costs.

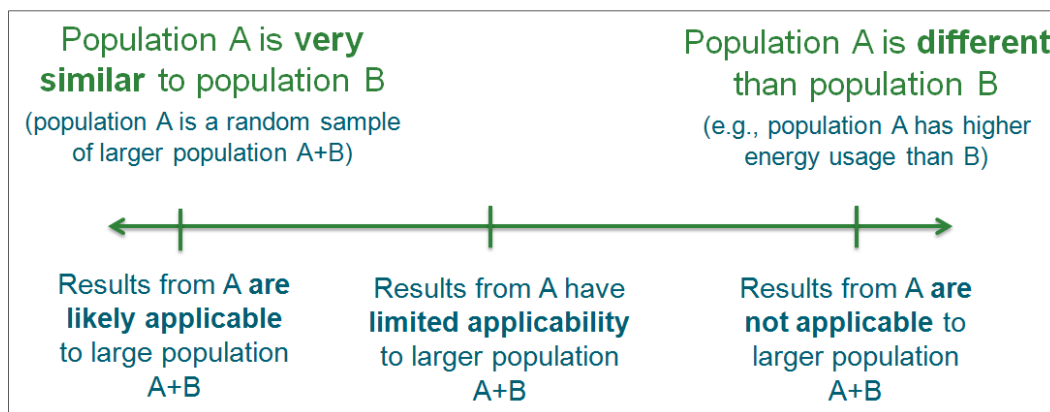
Figure 1. External Validity



Extrapolation to a Different Population in Year 1 of the Program

This section explores the possibility of extrapolating a valid savings impact estimate for one population to another population during the same year. This is a situation in which there are two populations (A and B) that a program administrator wishes to enroll in the program. Population A has a treatment and control group and their energy use data are measured in order to estimate a program savings impact. However, perhaps due to budgetary constraints, population B's energy use is not measured and evaluated, and all households in population B receive the program (i.e., the entire population is a treatment group). The question of interest is under what conditions the program savings impacts for population A can be extrapolated and applied to population B. In this situation, the external validity of the estimate depends on the similarity of population A to population B (see Figure 2). We recommend the following.

Figure 2. Applicability of Program Impact Estimates From One Population To Another



Recommendation: If there are two program participant populations, A and B, and A's energy savings are evaluated but B's are not, then the program impact estimates for A can only be extrapolated to B in the case that A was a random sample from population A + B and the same enrollment method was used (e.g., opt-in or opt-out).¹⁰

Persistence of Savings

An important issue for many stakeholders is whether energy savings from behavior-based programs continue over time (i.e., whether they persist beyond the initial program year). There are at least two different situations for which evaluators may assess persistence of savings: (1) the program provides periodic or continuous intervention (e.g., information and/or feedback) and customers may or may not continue to respond as they did initially and thus savings may erode, or potentially increase, during the program period; and (2) the program stops providing the intervention and thus savings may persist or erode in the absence of the intervention.

Because the subject programs are based primarily on changing energy behaviors or practices, there is concern that savings may not last or persist for many years or may be less predictable over time as compared to savings from installation of high-efficiency equipment and appliances. There is very limited evidence from just a few behavior efficiency programs that document savings for programs after the first year. This is at least in part because residential behavior-based programs have either only been offered or evaluated for the initial year. Thus, there is not enough evidence to draw any definitive conclusions (Skumatz 2009). This is complicated by the fact that persistence may not be uniform across different designs and types of behavior-based efficiency programs; it may depend on specific program elements, such as marketing channels (e.g., internet, letters, face to face), timing or consistency of feedback (e.g., monthly or real-time feedback), the type of customer segment that is targeted, or other factors. Persistence of savings may also be influenced by external conditions that change from year to

¹⁰ For planning purposes and cost-effectiveness screening, it may be appropriate to determine the degree to which population A is similar to B in order to establish and project savings estimates.

year (e.g., economic or weather conditions, other concurrent energy efficiency programs, the political climate, and popular culture).

From a planning perspective, it would be useful to know how the savings impacts of a typical behavior-based program change over time, which could be considered in cost-effectiveness screening. From an impact evaluation perspective, it would be useful to know whether a program must be evaluated after every year or whether the results from the first year (or first few years') can reasonably be extrapolated to future years (e.g., if year 1 had 2% savings, can we assume that savings for years 2 and 3 will also be 2%?).¹¹

More evidence on persistence of energy savings in behavior-based programs will become available as the new generation of programs mature and are evaluated over the next several years. Once a program has been running for several years and valid program impact estimates have been calculated in each of those years, it may be possible to evaluate the program every two or three years. However, at this time, we do not recommend applying results from one year directly to another year and foregoing evaluation entirely. Note that this implies that a control group must be maintained for every year in which program impact estimates are calculated and that the program treatment group therefore cannot be expanded to every household in a given population (see Table 10).¹²

Table 10. Persistence of Savings: Recommendation



A control group is maintained for every year in which program impacts are estimated, and the program is evaluated ex-post every year initially and every few years after the program has been running for several years



Program impact estimates are directly applied from the first year(s) of the program to future years without measuring and analyzing energy use data

Applying Savings Estimates to a New Population of Participants in Future Years

If a pilot behavior-based efficiency program is successful, program administrators may want to extend the program to additional populations over time. In this case, it may be important to assess whether the initial program's impact estimates can be applied to the expanded program. There are two contexts for which the validity of the estimates may be relevant: (1) program planning or cost-effectiveness screening; and (2) claiming energy savings credits after the program is implemented.

For planning purposes, the degree to which the initial population is similar to the new population and future years are similar to initial years determines the extent to which the initial savings estimates can be regarded as an ex ante savings estimate and extrapolated to this new situation. For the purpose of claiming savings credits as discussed in the previous section, we do not recommend directly applying program savings impact estimates from an initial program to an expanded program with a new population in a future year. Instead, we recommend the following.

¹¹ Saving impact estimates for an initial year could also be extrapolated to future years based on an assumed decay function that would reflect the fact that some/many customers will not continue energy efficient behaviors or practices in the absence of an ongoing program (e.g., initial year savings of 2% are assumed to decrease by 10-15% in subsequent years through year 4).

¹² However, the control group does not have to be half of the population, it could be far less. It is only necessary to keep a control group that is sufficiently large to yield statistical significance of the savings estimate (taking into account closed accounts and other attrition). If the control group is found to be larger than needed to yield statistical significance, then some households in the control group could be offered the program.

- **Recommendation: If the program is expanded to new program participant populations,** a control group that is representative of all of the different participating populations should be created and maintained for every population in the expanded program for every year in which program energy savings estimates are calculated.

Recommendations for the Future: Using Predictive Models to Estimate Savings

In theory, it is possible that a predictive model could be created that allows program estimates to be extrapolated to future years and new populations without actually measuring the savings estimates in those years. That is, it is possible that behavior-based programs could move towards estimating savings based on a calibrated analytic model, which could be used to produce a deemed savings estimate. However, we are not yet at this point and thus more behavior-based programs will need to be implemented and rigorously evaluated over multiple-year periods before we can assess whether predictive models can be developed that produce accurate and reliable estimates of deemed savings for these types of programs.

Rather than prescribe a method for creating a predictive model, we recommend a set of criteria that any predictive model must meet in order to be credible. These criteria focus on the reliability of a predictive model used for claiming energy savings credits for an implemented program that has not been evaluated with measured data (although predictive models could also be used for planning and cost-effectiveness screening purposes). These criteria are:

- **Internal conditions.** Ideally, internal program conditions (i.e., those controllable by the program administrator such as implementation methods) should remain the same in the predicted years as they were in the measured years.
- **External Conditions.** The model should account for external conditions (i.e., those uncontrollable by the program administrator) that may change over time (e.g., economic conditions, weather conditions, social norms, costs and availability of efficiency products and services, other efficiency strategies (e.g. new appliance standard), and other conditions that may affect the energy behavior of households).
- **Model Validation.** The model should be validated with many years of actual data by making a prediction ex-ante that is verified ex-post every few years.
- **Risk Adjustment.** From a policymaker or regulator's perspective, it is likely to be more risky to accept estimates of savings based on a predictive model than on measured data. One option is to adjust and reduce savings estimates produced by the predictive model to account for uncertainties.

Conclusion

This report provides guidance and recommendations on methodologies that can be used for estimating energy savings impacts resulting from residential behavior-based efficiency programs. Regulators, program administrators, and stakeholders can have a high degree of confidence in the validity of energy savings estimates from behavior-based programs if the evaluation methods that are recommended in this report are followed. In particular, we recommend using randomized controlled trials (RCTs) for behavior-based efficiency programs,

which result in robust, unbiased program savings impact estimates. We recommend that the RCT is maintained for every year in which program energy estimates are being used to claim savings. We also highlight the importance of mitigating the potential double counting of savings and of using third party evaluators. Finally, we discuss challenges in moving towards estimating savings based on a calibrated predictive analytic model.

Acknowledgements

The work described in this report was funded by the U.S. Department of Energy's Office of Energy Efficiency and Renewable Energy, Weatherization and Intergovernmental Program and the Permitting, Siting and Analysis Division of the Office of Electricity Delivery and Energy Reliability under Contract No. DE-AC02-05CH11231. It was prepared by Lawrence Berkeley National Laboratory for the State and Local Energy Efficiency Action Network's (SEE Action) Customer Information and Behavior (CIB) Working Group and the Evaluation, Measurement, and Verification Working (EM&V) Group. It was created with direction and comment by the CIB and EM&V Working groups and was vetted by several technical experts. It does not reflect the views, policies, or otherwise of the federal government.

References

- Allcott, H. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics*.
- Allcott, H., and S. Mullainathan. 2010. "Behavior and Energy Policy." *Science* 327.
- Allcott, H., and S. Mullainathan. 2012. "External Validity and Partner Selection Bias." NBER
- Angrist, J. D, and J. S Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton Univ Pr.
- Bertrand, M., E. Duflo, and S. Mullainathan. 2004. "How Much Should We Trust Differences-in-Differences Estimates?*" *Quarterly Journal of Economics* 119 (1): 249–275.
- Dougherty, A., A. Dwelley, R. Henschel, and R. Hastings. 2011. *Moving Beyond Econometrics to Examine the Behavioral Changes Behind Impacts*. IEPEC Conference Paper.
- Duflo, E., R. Glennerster, and M. Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." *Handbook of Development Economics* 4: 3895–3962.
- EPRI, Palo Alto, CA. 2010. *Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols*.
- Harding, M, and A. Hsiaw. 2011. "Goal Setting and Energy Efficiency." Working Paper.
- Imbens, G. M, and J. M Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47 (1): 5–86.
- LaLonde, R. J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review*: 604–620.

- Opinion Dynamics Corporation. 2011. Massachusetts Cross-Cutting Behavioral Program Evaluation. Waltham, MA.
- Skumatz, Lisa. 2009. Lessons Learned and Next Steps in Energy Efficiency Measurement and Attribution. Berkeley, CA. California Institute for Energy and Environment.
- Smith, Brian, and Michael Sullivan. 2011. Assessing Energy Savings Attributable to Home Energy Reports. International Energy Program Evaluation Conference.
- State and Local Energy Efficiency Action Network. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations*. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. <http://behavioranalytics.lbl.gov>
- Sullivan, M. 2009. "Using Experiments to Foster Innovation and Improve the Effectiveness of Energy Efficiency Programs." California Institute for Energy and Environment.
- Vine, E., M. Sullivan, L. Lutzenhiser, C. Blumstein, and B. Miller. 2011. Experimentation and the Evaluation of Energy Efficiency Programs: Will the Twain Meet? Boston, MA: International Energy Program Evaluation Conference.
- Wilhelm, Bobette, and Ken Agnew. 2012. Addressing Double Counting for the Home Energy Reports Program. Puget Sound Energy's Conservation Resource Advisory Group. <https://conduitnw.org/Pages/File.aspx?rid=786>.