
Table of Contents for Document (DRAFT)

1. Measure Description
2. Applicability Conditions of Protocol
3. Savings Concepts
 - 3.1 Definitions
 - 3.2 Experimental Research Designs
 - 3.3 Basic Features
 - 3.4 Common Designs
4. Savings Estimation
 - 4.1 Sample Design
 - 4.2 Data Requirements
 - 4.3 Analysis Methods
 - 4.4 Energy Efficiency Program Uplift and Double Counting of Savings
5. Reporting
6. References

DRAFT

1 Measure Description

Residential behavior-based (BB) programs use strategies grounded in the behavioral social sciences to influence household energy use. The programs often target multiple energy end uses and encourage energy savings, demand savings, or both. Specific strategies may include providing households with real-time or delayed feedback about their energy use; supplying energy-efficiency education and tips; rewarding households for reducing their energy use; comparing households to their peers; and establishing games, tournaments, and competitions.¹ Savings from BB programs are usually a small percentage of energy use, typically less than 5%.²

Utilities introduced the first large-scale residential BB programs in 2008. Since then, dozens of utilities have offered these programs to their customers.³ While program designs differ, many share these features:

- They are implemented using an experimental design, with eligible homes randomly assigned to treatment or control groups;
- They are large-scale by energy-efficiency program standards, targeting thousands of utility customers;
- They provide customers with an analysis of their historical consumption, energy-savings tips, and energy-efficiency comparisons to neighbor homes, either in personalized home reports or through a web portal, as well as offering incentives for savings energy; and
- They are typically implemented by outside vendors⁴.

While utilities will continue to implement residential BB programs as large-scale, randomized control trials (RCTs), some are now experimenting with alternative program designs that are smaller-scale, involve new communication channels such as the web, social media, and text messaging, or that employ novel strategies for encouraging behavior change (e.g., Facebook competitions).⁵ These programs will create new evaluation challenges and may require different evaluation methods than those currently employed.

¹ See Ignelzi et al. (2013) for a classification and descriptions of different BB intervention strategies.

² See Alcott (2011), Davis (2011), or Rosenberg, Agnew, and Gaffney (2013) for savings estimates from residential BB programs.

³ See the 2013 Consortium for Energy Efficiency (CEE) database for a listing of utility behavior programs; it is available for download: <http://library.cee1.org/content/2013-behavior-program-summary-public-version>.

⁴ Vendors that offer residential behavior-based programs include Aclara, C3 Energy, Opower, and Simple Energy.

⁵ The 2013 CEE database includes descriptions of many residential behavior-based programs with alternative designs such community-focused programs, college dormitory programs, K-12 school programs, and programs relying on social media.

2 Applicability Conditions of Protocol

This protocol recommends the use of randomized control trials (RCTs) or randomized encouragement designs (REDs) for evaluating BB programs that satisfy the following conditions:⁶

- Residential utility customers are the target;
- Energy or demand savings are the objective;
- An appropriately-sized analysis sample can be constructed; and
- Accurate energy-use measurements for sampled units are available.

The next section of this protocol carefully defines and explains these evaluation methods. A significant body of evidence indicates that experimental approaches work, that is, they result in unbiased and robust estimates of program energy and demand savings.

This protocol applies only to residential BB programs. In theory, evaluators can apply the experimental methods recommended in this protocol to nonresidential BB programs as well, and there are examples of them using such methods.⁷ But utilities have offered relatively few behavior-based programs to nonresidential customers thus far. As result, there is much less knowledge about the efficacy of evaluation methods in the nonresidential sector. As more evidence accumulates, NREL could expand this protocol to include nonresidential programs.

This protocol also addresses best practices for estimating both energy and demand savings. There are not significant conceptual differences between measuring energy savings and measuring demand savings; thus, evaluators can apply the algorithms in this protocol for calculating BB program savings to either. The protocol does not directly address the evaluation of other BB program objectives, such as increasing utility customer satisfaction, educating customers about their energy use, or increasing awareness of energy efficiency.

This protocol also requires that the analysis sample be sufficiently large. The analysis sample must be large enough to detect the expected savings with a high degree of confidence. Because most BB programs result in small percentage savings, the number of sampled units required to detect savings must be large. This protocol does not address program evaluations when working with insufficient sample sizes.

Finally, this protocol requires that it is possible to clearly identify and measure the energy use of participants or households affected by the program (both for the treatment and control groups). Typically, the unit of analysis is the household; in this case, it must be possible to identify treatment group households and measure individual household energy use. However, depending

⁶ As discussed in Considering Resource Constraints in the Introduction of this UMP report, small utilities (as defined under the U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

⁷ For example, PG&E offers a Business Energy Reports Program, which it implemented as a field experiment. See http://beccconference.org/wp-content/uploads/2013/12/BECC_PGE_BER_11-19-13_seelig.pdf

on the BB program, the units of analysis may not be households. For example, for a BB program that generates an energy competition between thousands of housing floors at a university, the unit of analysis may be floors; in this case, the protocol requires that the energy-use measurement of individual floors be available.

The characteristics of BB programs that *do not* determine the applicability of the evaluation protocol include:

- Whether the program is opt in or opt out;⁸
- The specific behavior-modification theory or strategy; and
- The channel(s) through which program information is communicated.

This protocol does not recommend quasi-experimental methods to evaluate behavior-based programs covered by this protocol. Evaluators have employed quasi-experimental methods to evaluate behavior-based programs, but more knowledge about their efficacy is needed.⁹ As more evidence accumulates, NREL may update this protocol as necessary.

Also, the protocol does not recommend evaluation methods for behavior-based programs *not covered* by this protocol. Though this protocol does not apply to all BB programs, evaluators may still be able to use experimental methods to develop savings estimates. Currently, however, there is not enough evidence whether evaluators can uniformly adopt these methods. For information about alternative approaches to estimating savings from BB programs not covered by this protocol, consult EPRI (2010), SEE Action (2012), or EPRI and LBNL (forthcoming).

2.1 Examples of Protocol Applicability

The following are examples of residential BB programs for which the evaluation protocol applies:

Example 1: A utility sends energy reports encouraging conservation steps to thousands of randomly selected residential customers.

Example 2: Several hundred residential customers enroll in a Wi-Fi-enabled thermostat pilot program offered by the utility.

⁸ In opt-in programs, customers enroll or select to participate. In opt-out programs, the utility enrolls the customer, and the customer remains in the program until they opt out. An example opt-in program is having a utility web portal with home energy use information and energy-efficiency tips that residential customers can use if they choose. An example opt-out program is sending energy reports to utility selected customers. Application of experimental research designs to opt-in programs is more complicated because random assignment of customers to the treatment or control group occurs after interested customers have opted in to the program.

⁹ Allcott (2011) shows that a within-subject design using a pre-post comparison of monthly energy use of households receiving energy reports overestimates savings in comparison to difference-in-differences estimation using both treatment and control group subjects. Harding and Hsiaw (2012) use variation in timing of participation for program participants and difference-in-differences estimation to estimate savings from utility customer setting of energy-savings goals.

Example 3: A utility invites thousands of residential customers to use its web portal to track their energy use in real time, set goals for energy saving, get ideas about how to reduce their energy use, and get points/rewards for saving energy.

The following are examples of programs for which the protocol does not apply:

Example 4: A utility uses a mass-media advertising campaign relying on radio and billboards to encourage residential customers to conserve energy.

Example 5: A utility initiates a social media campaign (e.g., using Facebook or Twitter) to encourage energy conservation.

Example 6: A utility runs a pilot program testing the savings from in-home energy-use displays, and enrolls too few customers to detect the expected savings.

Example 7: A utility runs a BB program in a large college dormitory to change student attitudes about energy use. The utility randomly assigns some rooms to the treatment group. The dorm is master-metered.

The protocol does not apply to Example 4 or Example 5 because the evaluator cannot identify who received the messaging. The protocol does not apply to Example 6 because the number of customers in the pilot is too small to accurately detect energy savings. The protocol does not apply to Example 7 because energy-use data is not available for the specific rooms in the treatment and control groups.

3 Savings Concepts

This protocol applies to residential BB programs that satisfy the conditions described in the Applicability Conditions of Protocol section. The recommended approach for estimating energy or demand savings is using RCT or RED with a regression analysis of energy use from the periods before and during the program for both treatment and control subjects in the analysis sample. The objective of employing a RCT or RED is to achieve an unbiased estimate of the program treatment effect on the analysis population. This section defines some key concepts and describes the specific evaluation methods.

3.1 Definitions

Control Group. In an experiment, the control group comprises subjects (utility customers) who do not receive the program intervention or treatment.

Experimental Design. Experimental designs are methods of program evaluation that rely on observing energy use and randomly assigning program treatments or interventions in a controlled process.

External Validity. Savings estimates are externally valid if evaluators can apply them to different populations or different time periods from those studied.

Internal Validity. Savings estimates are internally valid if the savings estimator is expected to equal the causal effect of the program on consumption.

Opt-In Program. Utilities use **opt-in** BB programs if it is necessary that the customer agree to participate, and the utility cannot administer treatment without consent.

Opt-Out Program. Utilities use **opt-out** BB programs if it is unnecessary that customers agree to participate. The utility can administer treatment without consent, and customers remain enrolled until they opt out by asking the utility to stop the treatment.

Quasi-Experimental Design. Quasi-experimental designs are methods of program evaluation that rely on a comparison group that is not obtained via random assignment. Such designs observe energy use and determine program treatments or interventions based on factors that may be partly random but not controlled.

Randomized Control Trial. A RCT is an experimental design that yields an unbiased estimate of savings. Evaluators randomly assign subjects from a study population to a treatment group or a control group. Subjects in a treatment group receive one of the program treatments (there may be multiple treatments), while subjects in the control group do not receive a treatment. The RCT ensures that whether subjects receive the treatment is uncorrelated with their energy-use characteristics, and that evaluators can attribute any difference in energy use between the groups to the treatment.

Randomized Encouragement Design. A RED is an experimental design in which evaluators randomly assign subjects to a treatment group that receives *encouragement* to participate in a program or to a control group that does not receive encouragement. Evaluators cannot restrict program participation to subjects in the treatment group. The RED yields an unbiased estimate of

the effects of encouraging energy-efficient behaviors and the affect on customers who participate because of the encouragement.¹⁰

Treatment. The treatment is an intervention administered through the BB program to subjects in the treatment group. Depending on the research design, the treatment may be a program intervention or encouragement to receive an intervention.

Treatment Effect. The treatment effect is the effect of the BB program intervention(s) on energy use for a specific population and time period.

Treatment Group. The treatment group includes subjects who receive the treatment.

3.2 Experimental Research Designs

This section outlines experimental methods for evaluating BB programs. The most important benefit of a RCT or RED is that, if carried out correctly, the experiment results in savings estimates with internal validity. Internal validity refers to whether the estimate of savings in the study population is unbiased. A result is internally valid if the evaluator expects the value of the estimator to equal the savings caused by the program intervention. The principal threat to internal validity in BB program evaluation derives from potential selection bias in who receives a program intervention. RCTs and REDs yield unbiased savings estimates because they ensure that receiving the program intervention is uncorrelated with the energy-use characteristics of subjects.

In addition, experimental research designs may yield savings estimates that are applicable to other populations or time periods, making them externally valid. Whether savings have external validity will depend on the specific research design, the study population, and other program properties.

A benefit of experimental methods is their versatility: evaluators can apply them to a wide range of BB programs regardless if they are opt-in or opt-out programs. Evaluators can apply experimental methods to any program where the objective is to create energy or demand savings; evaluators can construct an appropriately-sized analysis sample; and accurate measurements of the energy use of sampled units are available. Finally, experimental methods generally yield highly robust savings estimates that are not model-dependent, that is, they do not depend on the specification of the model used for estimation.

The choice of whether to use a RCT or RED to evaluate program savings should depend on several factors, including whether it is a opt-in or opt-out program and the utility's tolerance for subjecting customers to the requirements of an experiment. For example, using a RCT for an opt-in program might require delaying or denying participation for some customers. A utility may prefer to use a RED to accommodate all the customers who want to participate.

¹⁰ If the effect of program participation is the same for compliers, those who participate because of the encouragement, as for others, those who would have participated without encouragement ("always-taker") and those who do not participate ("never takers"), the RED yields an unbiased estimate of the population treatment effect.

Finally, it requires up-front planning to implement a RCT or RED design. Program evaluation must be an integral part of the program planning process; this need will be evident in the experimental research design descriptions described in the following section.

3.3 Basic Features

This section outlines several types of RCT research designs, which are simple but extremely powerful research tools. The core feature of RCT is the random assignment of study subjects (e.g., utility customers, floors of a college dormitory) to a treatment group that receives or experiences an intervention or to a control group that does not receive the intervention.

This section outlines some common features of RCTs and discusses specific cases.

Common Features of All RCT Designs

The steps for employing a RCT design include the following:

1. **Random Assignment:** To avoid the appearance of a conflict of interest and to ensure the integrity of the RCT, this protocol highly recommends that a qualified, independent third-party randomly assign subjects to treatment and control groups. Also, to preserve the integrity of the experiment, customers must not have a choice about their assignment.
2. **Sample Size:** The numbers of subjects necessary for treatment and control groups depends on the type of experimental design (e.g., REDs and opt-out RCTs generally require more customers). The number of subjects assigned to the treatment versus control groups does not have to be equal, but each should be large enough to detect the hypothesized program effect with sufficient probability.¹¹

Evaluators can determine the number of subjects required with a statistical power analysis. This results in minimum sample sizes for the treatment and control groups as a function of the hypothesized program effect, the coefficient of variation of energy use, the specific analysis approach that will be used (e.g., simple differences of means, a repeated measure analysis), and tolerances for Type I and Type II statistical errors.¹² Most statistical software including SAS, STATA, and R now include packages for performing statistical power analyses. It is not uncommon for BB programs with expected savings of less than 5% to require thousands of subjects in the treatment and control groups.¹³

An important component of the random assignment process is to verify that the treatment and control groups are statistically equivalent or balanced in their observed covariates. At

¹¹ The number of subjects in the treatment group may also depend on the size of the program savings goal.

¹² A Type I error occurs when a researcher rejects a null hypothesis that is true. Statistical confidence equals 1 minus the probability of a Type I error. A Type II error occurs when a researcher accepts a null hypothesis that is false. Many researchers agree that the probability of a 5% Type I error and a 20% Type II error is acceptable. See List, Sadoff, and Wagner (2010).

¹³ EPRI (2010) illustrates that, all else equal, repeated measure designs, which exploit multiple observations of energy use per subject both before and after program intervention, require smaller analysis sample sizes than other types of designs.

a minimum, evaluators should check for statistically significant differences in the average energy use and in the distribution of energy use between treatment and control homes before the intervention. If differences exist, evaluators must use random assignment.

3. **Administer the Treatment:** The next step is to administer the intervention to the treatment group while withholding it from subjects in the control group. To avoid a Hawthorne effect, in which subjects change their energy use in response to observation, control group subjects should receive minimal information about the study. Depending on the subject of research and type of intervention, the utility may administer treatment once or repeatedly and for different durations. However, the treatment period should be long enough for evaluators to observe any effects of the intervention.
4. **Data Collection:** The fourth step of employing a RCT design is to collect the energy use data for all subjects from both the treatment and control groups. It is very important to collect data from all of the subjects and not only from those who actually chose to participate or only from those who did not drop out of the study/experiment.

It is preferable for evaluators to collect multiple energy-use measurements from both before and after the start of the treatment. Such data will enable the evaluator to control for time-invariant differences in the average energy use between the treatment and control groups in order to obtain more precise savings estimates. Step 5 discusses this in further detail.

5. **Estimate Savings:** The last step of employing a RCT design is to estimate the savings from the intervention. Evaluators should calculate this as the difference in energy use between the subjects who were initially assigned to the treatment versus control group. To be able to calculate an unbiased savings estimate, evaluators must compare the energy use from the entire group of subjects who were originally randomly assigned to the treatment group to the entire group of subjects who were originally randomly assigned to the control group. For example, the savings estimate would be biased if evaluators used only data from utility customers in the treatment group who actually chose to participate in the study.

The difference in energy use between the treatment and control groups, which is usually referred to as an intent-to-treat (ITT) effect, is an unbiased estimate of savings because of the random assignment of subjects to the treatment and control groups. The effect is an ITT because in contrast to many randomized clinical medical trials, it is not possible to ensure that treatment group subjects in most BB programs comply with the treatment. For example, some households may opt out of an energy reports program, or they may fail to notice or simply ignore the energy reports. As a consequence, the effect is ITT; and the evaluator would base the results on the initial assignment of subjects to the treatment group, whether or not subjects actually complied with the treatment.

3.4 Common Designs

The following section describes some of the RCT designs commonly used in BB programs.

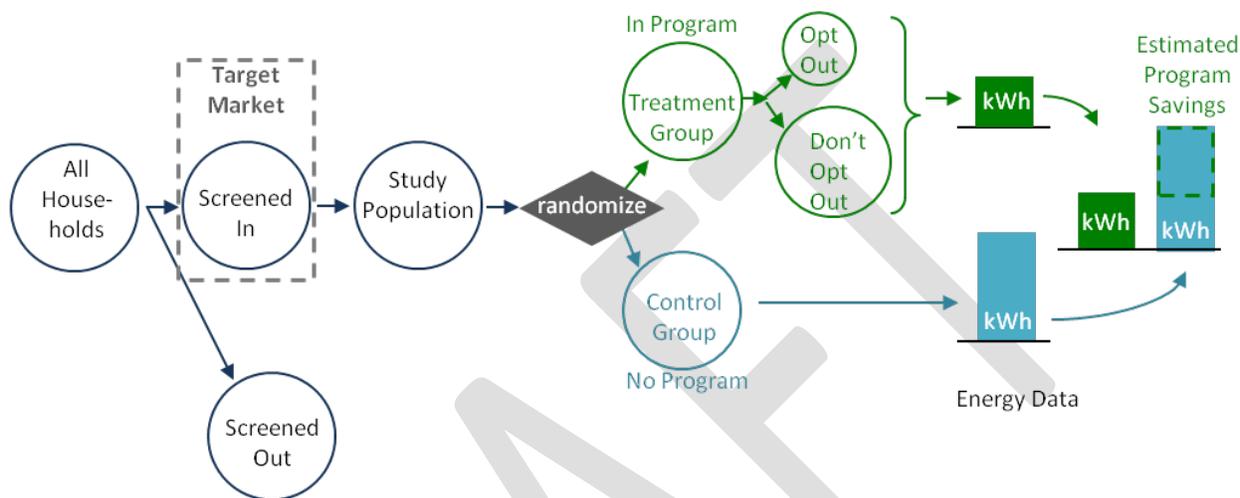
RCT With Opt-Out Program Design

One common type of RCT includes the option for treated subjects to opt out of receiving the program treatment. This design reflects the most realistic description of how most BB programs

employ experimental designs. For example, in energy reports programs, treated customers may ask the utility to stop sending them reports.

Figure 1 depicts the process flow of a RCT in which treated customers can opt out of the program. In this illustration, the utility performed an initial screening of utility customers to refine the study population.

Figure 1. Illustration of RCT With Opt-Out Program Design¹⁴



It will be necessary for the utility to screen the utility population if it wishes to only offer the program intervention to certain customer segments, such as when only single-family homes are eligible to participate. Programs designers can base eligibility on dwelling type (e.g., single family, multifamily), geographic location, completeness of recent billing history, heating fuel type, utility rate class, or other energy-use characteristics.

Customers that pass the screening constitute the study population or sample frame. The estimate of savings will apply to this population. Alternatively, the utility may want to study only a sample of the screened population, in which case a third party should sample randomly from the study population. The analysis sample must be large enough to meet the minimum size requirement for the treatment and control groups. The program savings goals and desired statistical power will determine the size of the treatment group. .

The next steps in a RCT with opt-out program design are to randomly assign subjects in the study population to the program treatment and control groups, administer the program treatments, and collect energy-use data.

The distinguishing feature of this experimental design is that customers can opt out of the program. As Figure 1 shows, evaluators should include opt-out subjects in the energy-savings analysis to ensure internal validity of the savings estimates. Evaluators can then calculate savings

¹⁴ This graphic and the following ones are variations of those that appeared in SEE Action (2012). A coauthor of the SEE Action report and the creator of that reports' figures is one of the authors of this protocol.

as the difference in average energy use between treatment group customers, including opt-out subjects, and control group customers. Removing opt-out subjects from the analysis would bias the estimate of savings because evaluators can assume that the same percentage of subjects in the control group would have also chosen to opt out had they been a part of the treatment group. The resulting savings estimate is therefore an average of the savings of treated customers who remain in the program and those of customers who opted out.

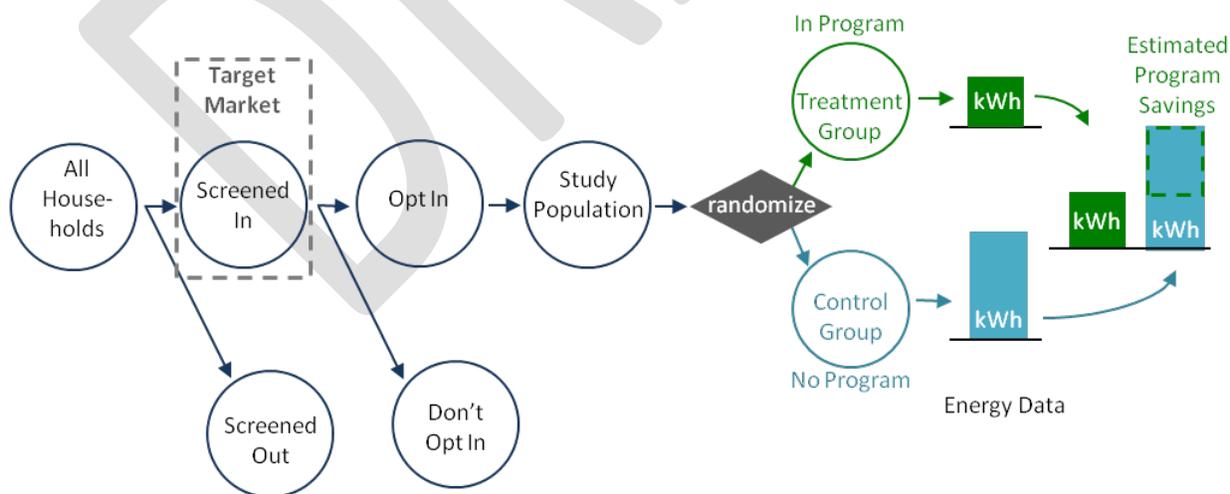
Depending on the type of BB program, the percentage of customers who opt out may be small, and may not affect the savings estimates significantly (e.g., few customers generally opt out of energy reports programs).

RCT With Opt-In Program Design

RCT with opt-out subjects assumes that it is possible to administer the BB program treatments to subjects without their agreement. This is the case for programs in which, for example, a utility mails energy reports to customer homes or leaves door hangers with energy-savings tips on customer homes. However, the utilities cannot administer some interventions without customer consent. Examples of such BB interventions include offering web-based home-audit or energy-consumption tools; programmable, communicating thermostats with wireless capability; an online class about energy rates and efficiency; or in-home displays. All of these examples require that customers opt in to the program.

There is a modified approach to the program design for an opt-in RCT, shown in Figure 2. The results in this case are an unbiased estimate of the ITT effect for customers who opt-in to the program. While the estimate of savings will have internal validity, it will not have external validity, because it will not apply to subjects who do not opt-in. Figure 2. Figure 2 illustrates the approach.

Figure 2. Illustration of RCT With Opt-In Program Design



Implementing opt-in RCTs is very similar to implementing opt-out RCTs. The first two steps, random assignment and sample size, are the same. The third step, data collection, is different. Upon marketing the program, some eligible customers will agree to participate. Then, a third party randomly assigns these customers to either a treatment group that receives the intervention or a control group that does not. The utility delays or denies participation for customers assigned to the control group. Thus, only customers assigned to the treatment group and opt in to the program will participate in the experiment.

Randomizing only opt-in customers ensures that the treatment and control groups are equivalent in their energy-use characteristics. In contrast, other quasi-experimental approaches, such as matching participants to nonparticipants, cannot guarantee this equivalence and thus the internal validity of the savings estimates.

After the random assignment, the opt-in RCT proceeds the same as a RCT with opt-out subjects: the utility administers the intervention to the treatment group. The evaluator collects energy-use data from both treatment and control groups, then estimates energy savings as the difference in energy use between the groups. The evaluator does not collect energy-use data for customers who do not opt in to the program.

An important difference between the opt-in RCTs and RCTs with opt-out subjects is how to interpret the savings estimates. In the RCT with opt-out subjects, the evaluator will base the savings estimate on a comparison of the energy use between treatment and control groups, which pertains to the entire study population. In contrast, in the opt-in RCT, the savings estimate pertains to the subset of customers who opted into the program, and the difference in energy use represents the treatment effect on customers who opted in to the program. While opt-in RCT savings estimates have internal validity, they do not apply to customers who did not opt in to the program.

Randomized Encouragement Design

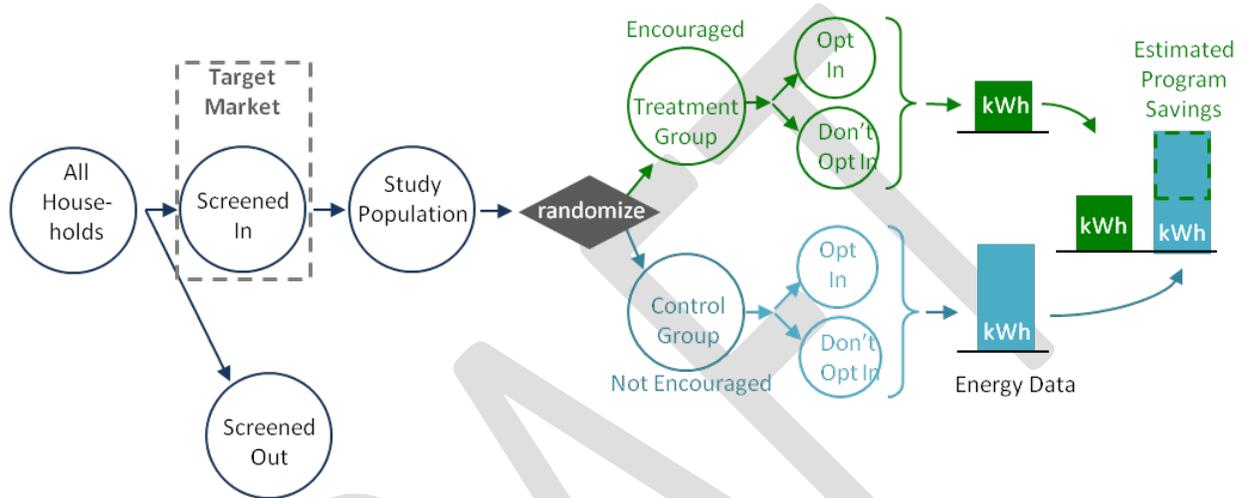
For some opt-in BB programs, it may be undesirable to delay or deny participation to some customers. In this case, neither the opt-out nor opt-in RCT design would be appropriate, and this protocol recommends a RED. Instead of randomly assigning subjects to receive or not receive an intervention, a third party randomly assigns them to a treatment group encouraged to accept the intervention (i.e., to participate in a program or adopt a measure), or to a control group not encouraged to participate. Customers who receive the encouragement can refuse the intervention, just as control group customers who learn about the intervention can choose to participate (depending on the program design).

As explained below, the RED yields an unbiased estimate of the effect of encouragement on energy use and, depending on the program design, can also provide an unbiased estimate of either the effect of the intervention on customers who accept it because of the encouragement or the effect of the intervention on all customer who accept it, regardless of the reason for their acceptance.

Figure 3 illustrates the process flow for a program using RED. As with the RCT with opt-out and opt-in RCT, the first two steps are to identify the sample frame and select an analysis sample.

Next, like the RCT with opt-out, a third party assigns subjects to a treatment group, which receives encouragement, or control group. For example, a utility might employ a direct-mail campaign encouraging treatment group customers to use an online audit tool. The utility would administer the intervention to treatment group customers who opt-in. While customers in the control group do not receive encouragement, some of them may learn of the program and decide to sign up. The program design shown in Figure 3 allows for control group customers to receive the behavioral intervention.

Figure 3. Illustration of RED Program Design



In Figure 3, the difference in energy use between homes in the treatment and control groups is an estimate of savings from the encouragement, not from the intervention. However, evaluators can also use the difference in energy use to estimate savings for customers who accept the intervention because of the encouragement, as opposed to those having another reason to adopt the intervention. To see this, consider that the study population comprises three types of subjects: 1) always takers, or those who would accept the intervention whether encouraged or not; 2) never takers, or those who would never accept the intervention even if encouraged; and 3) compliers, or those who would only accept the intervention if encouraged. Compliers are customers who participate only after receiving the encouragement.

Because eligible subjects are randomly assigned to whether they receive encouragement, assume that the treatment and control groups have equal frequencies of always takers, never takers, and compliers in expectation. The only difference between the treatment and control groups is that compliers in the treatment group accept the treatment while those in the control group do not. In both groups, always takers accept the treatment and never takers always refuse the treatment. Therefore, the difference in energy use between the groups reflects the treatment effect of encouragement on compliers.

To estimate the effect of the intervention on compliers (known as the local average treatment effect [LATE]), it is necessary to scale the treatment effect of the encouragement by the following factor:

1/(% of encouraged customers that accepted - % of not encouraged customers that did not accept)

The LATE does not capture the program effect on always takers because always takers in the control group are permitted to participate.

For BB programs with REDs that do not permit control group customers to participate, evaluators can estimate the treatment effect on the treated (TOT). The TOT is the effect of the program intervention on all customers who accept the intervention. In this case, the difference in energy use between the treatment and control groups reflects the impact of the encouragement on the always takers and compliers in the treatment group. Scaling the difference by the inverse of the percentage of customers who accepted the intervention yields an estimate of the TOT impact.

Successful application of an RED requires that compliers constitute a sufficiently large percentage of the encouraged population.¹⁵ If the RED generates too few compliers, it will not be possible to precisely estimate the effects of the encouragement and receiving the intervention. Therefore, before employing an RED, evaluators should ensure that the sample size is sufficiently large and that the encouragement will result in the required number of compliers. If the risk of an RED generating too few compliers is significant, evaluators may want to consider alternative approaches, including quasi-experimental methods.

Persistence Design

Many utilities want to know what happens to BB program savings after the intervention ends. For example, studies of home energy reports programs show that program savings are durable while homes continue to receive reports.¹⁶ The utility needs to measure whether savings persist after the Utility stops sending reports and for how long, as well as the rate of the savings decrease, if any.

This protocol recommends that evaluators employ RCTs to estimate the persistence of savings, that is, the effect of the intervention on energy use after the intervention stops. The application of an RCT to a savings persistence study proceeds similarly to the application of RCTs previously discussed.

It is assumed that the utility implemented the BB program as a RCT with opt-out design, that is, customers from the study population were randomly assigned to a treatment group that received an intervention or to a control group that did not. Customers are able to opt-out of the program (see Figure 1).

The persistence study starts with identifying the study population, in this case, the population of treated customers that received the intervention. The utility may choose to screen this population and study persistence by energy use or by socio-demographic characteristics. It is important that

¹⁵ For an example of the successful application of an RED, see SMUD (2013).

¹⁶ Allcott and Rodgers (2012) researched the durability and persistence of savings from a home energy reports program. They define “durability” as the extent to which savings maintain or increase while homes receive interventions, and “persistence” as the extent to which savings last after the interventions end.

the persistence study population include customers that opted out, because evaluators will need to make energy-use comparisons between the persistence study population and the original control group, which includes customers who would have opted out.

The next step is to randomly assign customers in the persistence study population to a treatment group or control group. Customers in the persistence treatment group will stop receiving the intervention, while customers in the persistence control group will continue receiving intervention. The utility then administers the study and collects energy-use data after sufficient time has passed to observe the persistence effects.

Then to estimate savings persistence, the evaluator makes two comparisons:

1. The energy use of customers in the persistence control group with the energy use of customers in the original control group. This represents the savings for customers who continued to receive the intervention.
2. The energy use of customers in the persistence treatment group with the energy use of customers in the original control group. This represents the savings for customers who no longer receive the intervention; and

The difference in the savings is an estimate of the unrealized savings because the intervention stops.

4 Savings Estimation

Evaluators should estimate BB program savings as the difference in energy use between treatment and control group subjects in the analysis sample. Energy savings for a household in the BB program is the difference between the energy the household used and the energy the household would have used if it had not participated. However, it is not possible to observe the energy use of a household in two different states. Instead, to estimate savings, we compare the energy use of households in the treatment group to the energy use of a group of households that are statistically the same but did not receive the treatment (the homes randomly assigned to the control group). Since the treatment and control group assignments are random, evaluators can expect the control group to use the same amount of energy that the treatment group would have used without the treatment and that their estimates of the energy savings is unbiased between the difference between the two groups.

It is possible to estimate savings using either energy use data from only during the treatment or energy use from before and during the treatment. If energy use from only the treatment period is used, evaluators will estimate the savings as a simple difference (D). If the analysis also controls for energy use before the treatment, evaluators will estimate the savings as a difference-in-differences (D-in-D). Evaluators will choose D or D-in-D estimation depending on the availability of energy use data for the period before the treatment.

Both approaches will result in unbiased estimates of savings, but D-in-D estimation generally results in more precise savings estimates (i.e., in expectation, the two methods yield the same savings estimate, but the D-in-D estimate has a smaller standard error) because it allows for removing time-invariant energy-use characteristics that contribute significantly to the variance of energy use between subjects.

In addition, D-in-D estimation removes time-invariant differences in the mean energy use between treatment and control group subjects who are introduced through the process of randomly assigning subjects to the treatment or control group. While the assignments are random evaluators may not expect such differences, the differences may nevertheless exist in the analysis sample.

4.1 Sample Design

Utilities should integrate the design of the analysis sample with program planning, as there are numerous considerations that must be addressed before the program begins. These considerations include the size of the analysis sample and the method of recruiting customers to the program.

Sample Size

First, the analysis sample should be large enough to detect the minimum hypothesized program effect with desired probability.¹⁷ To determine the minimum number of subjects required, researchers should employ a statistical power analysis. The inputs for this calculation are: the

¹⁷ The utility may also base the number of subjects in the treatment group on the amount of savings it desires to achieve.

hypothesized program effect, the coefficient of variation of energy use, the specific analysis approach to be used (e.g., simple differences of means or a repeated measure analysis), tolerances for Type I and Type II statistical errors (as discussed in the Experimental Research Designs section), and the correlation of a subject's energy-use observations. Most statistical software, including SAS, STATA, and R, include packages for performing statistical power analyses.¹⁸

If the BB program will operate for more than several months or allow subjects to opt out, it may be necessary to account for attrition or the loss of some subjects from the analysis sample due to account closures and opt out customers.

Random Assignment to Treatment and Control Groups by Independent Third Party

After determining the appropriate sizes of the treatment and control group samples, an independent and experienced third-party evaluator should perform the random assignment of subjects to the treatment or the control group.

Equivalency Check

It is important for evaluators performing the random assignment to verify that the characteristics of subjects in the treatment group are balanced with those in the control group, including energy use. If subjects in the groups are not equivalent overall, there is the potential for bias in the energy-savings estimates.

To verify the equivalence of energy use, this protocol recommends that evaluators test for differences between treatment and control group subjects in both the mean pre-treatment period energy use and in the distribution of pre-treatment energy use. In addition, evaluators should test for differences in other available covariates, such as home floor area or heating fuel type.

If significant differences are found, the third party should perform the random assignment again.

If the evaluator is not the third party who performed the random assignment, the evaluator should also perform an equivalency check. Using statistical methods, it may still be possible to control for pre-existing differences in energy use that are found after the program is underway.¹⁹

4.2 Data Requirements and Collection

¹⁸ If statistical software is not available and one wishes to calculate the sample sizes by hand, Sergici and Faruqui (2011) provides formulas for calculating sample sizes using statistical power calculations. The formulas also appear in Hilbe (1993) and Seed (1997).

¹⁹ If energy-use data are available for the periods before and during the treatment, it is possible to control for time-invariant differences between sampled treatment and control group subjects using subject fixed effects.

Energy Use Data

Estimating BB program impacts using an experimental research design requires collecting energy-use data from subjects in the analysis sample. This protocol recommends that evaluators collect multiple energy-use measurements for each sampled unit for the periods before and during the treatment.²⁰

These data are known as a panel. Panels can consist of multiple hourly, daily, or monthly energy-use observations for each sampled unit. In this protocol, a panel refers to a dataset that includes energy measurements for each sampled unit either for the pre-treatment and treatment periods or for just the treatment period. The time period for panel data collection will depend on the program timeline, the frequency of the energy-use data, and the amount of data collected.

Panel data have the several following advantages for use in measuring BB program savings:

- **Ease of collection.** It is usually easy and inexpensive to collect multiple energy-use measurements for each sampled unit from utility billing systems. Collecting panel data such as 12 months each of pre-treatment and treatment period bills is usually not onerous or expensive.
- **Can estimate savings during specific times.** If the panel collects enough energy-use observations per sampled unit, it may be possible to estimate savings at specific times during the treatment period. For example, hourly energy-use data may enable the estimation of precise savings during utility system peak hours. Monthly energy-use data may enable the development of precise savings estimates for each month of the year.
- **Savings estimates are more precise.** Evaluators can more precisely estimate energy savings with a panel, as it is possible to control for the time-invariant differences in energy use between treatment and control group subjects that increases the precision of energy-use savings estimates in other types of studies.
- **Allows for smaller analysis samples.** All else being equal, the number of units required to detect a minimum level of savings is less in a panel study than in a cross-section analysis. Thus, collecting panel data may enable studies with relatively small analysis samples.

There are, however, some disadvantages of using panel data relative to a single measurement per household. One is that evaluators must correctly cluster the standard errors within each household or unit (as described in the following section). In addition, panel data requires statistical software to analyze, whereas it may be possible to estimate savings using single measurements in a basic spreadsheet software program.

This protocol also recommends that evaluators collect energy-use data for the duration of the treatment, ensuring that evaluators can observe the treatment effect for the duration of the experiment. Ideally, an energy-efficiency BB program lasts for a year or more because the energy end uses affected by BB programs vary seasonally. For example, these programs may

²⁰ A single measurement of energy-use for each sampled unit during the treatment period will also result in an unbiased estimate of program savings. The statistical significance of the savings estimate will depend on the variation of the true but unknown savings and the number of sampled units.

influence weather-sensitive energy uses, such as space heating or cooling, so collecting less than one year of data to reflect every season may yield incomplete results.

It may not be possible to collect data for an entire year because some BB programs do not last that long. For these programs, it is only possible to obtain an unbiased estimate of savings for the time period of analysis. Evaluators should exercise caution s in extrapolating those estimates to seasons or months outside of the analysis period, especially if the BB program affected weather-sensitive or seasonally variant end uses.

Makeup of Analysis Sample

It is important to note that evaluators must collect energy-use measurements for every household or unit that is initially assigned to a control or treatment group, whether or not the household or unit later decided not to participate in the treatment (e.g., the unit opted out). It is not sufficient to collect energy-use data for households initially placed in a treatment group, but then decided not to participate; doing so will result in a biased savings estimate.

Other Data Requirements

In addition to energy-use measurements, it is necessary to collect program information about each program participant. This data must include whether the subject was assigned to the treatment or control group, when the treatments were administered, and if and when the subject opted out.

Temperature and other weather data may also be useful but are not necessary. Collect weather data for each household from the nearest weather station.

Data Collection Method

Collect the energy-use measurements used in the savings estimation from the utility, not from the program implementers, at the end of the program evaluation period. Depending on the program type, utility billing system, and evaluation requirements, the data frequency may be at 15-minute, one-hour, daily, or monthly intervals.

4.3 Analysis Methods

This protocol recommends using panel regression analysis to estimate savings from BB field experiments where subjects were randomly assigned to either a treatment or control group. Evaluators typically prefer regression analysis to simply calculating differences in unconditional mean energy use, because it generally results in more precise savings estimates. A significant benefit of field experiments is that regression-based savings estimates are usually not very sensitive to the type of model specification.

This section addresses issues in panel regression estimation of BB program savings, including model specification and estimation, standard errors estimation, robustness checks, and savings estimation. It illustrates some specifications as well as the application of energy-savings estimation.

Panel Regression Analysis

In panel regressions, the dependent variable is usually the energy use of a subject (a home, apartment, dormitory) per unit of time such a month, day, or hour. The right side of the equation includes an independent variable to indicate whether the subject was assigned to the treatment or control group. This variable can enter the model singularly or be interacted with another independent variable, depending on the analysis goals and the availability of energy-use data from before treatment. The coefficient on the term with the treatment indicator is the energy savings per subject per unit of time. Difference-in-differences models of energy savings must also include an indicator for whether the period occurred before or during the treatment period.

Many panel regressions also include fixed effects. Subject fixed effects capture unobservable energy use specific to a subject that does not vary over time. For example, home fixed effects may capture variation in energy use that is due to differences such as the sizes of homes or makeup of a home's appliance stock. Time-period fixed effects capture unobservable energy use specific to a time period that does not vary between subjects. Including time or subject fixed effects in a regression of energy use of subject randomly assigned to the treatment or control group will increase the precision but not the unbiasedness of the savings estimates.

There are several ways to incorporate fixed effects into panel regression. As an example for subject fixed effects, one approach can include a separate dummy variable or intercept for each subject in the model. The estimated coefficient on a subject's dummy variable represents the subject's time-invariant energy use. This approach, known as Least Squares Dummy Variables (LSDV), may, however, not be practical for evaluations with a large number of subjects, because the model requires the inclusion of thousands of dummy variables that may overwhelm available computing resources.

Another approach to modeling fixed effects is to apply the fixed-effect estimator, which requires transforming the dependent variable and all of the independent variables by subtracting subject-specific means and then running Ordinary Least Squares (OLS) on the transformed data.²¹ This approach is equivalent to LSDV.

A third way of modeling fixed effects is to estimate a first difference or annual difference of the model. Differencing removes the subject fixed effect and is equivalent to the dummy variable approach if the fixed-effects model is correctly specified.²²

²¹ See Chapter 11 of Greene (2011) for more details.

²² In standard econometric formulations, assume that fixed effects account for unobservable factors that are correlated with one or more independent variable in the model. This correlation assumption is what distinguishes fixed-effects panel model estimation from other types of panel models. Fixed effects eliminate bias that would result from omitting unobserved time-invariant characteristics from the model. In general, to avoid omitted variable bias, it is necessary to include fixed effects. In an RCT, however, fixed effects are unnecessary to the claim that the estimate of the treatment effect is unbiased because fixed effects are uncorrelated with the treatment by design. While fixed effects regression is unnecessary, it will increase precision by reducing model variance.

Panel Regression Model Specifications

The text below outlines common regression approaches for estimating treatment effects from residential BB programs. Unless otherwise stated, assume that the BB program was implemented as a field experiment with a randomized control trial or randomized encouragement design.

Simple Differences Regression Model of Energy Use

Consider a BB program in which the evaluator has energy-use data for during the treatment period only, and wishes to estimate the average energy savings per period from the treatment. Let $t=1, 2, \dots, T$, where t denotes the time periods during the treatment for which data are available²³, and let $i=1, 2, \dots, N$, where i denotes the treatment and control group subjects. For simplicity, assume that all treated subjects started the treatment at the same time.

A basic specification to estimate the average energy savings per period from the treatment is:

Equation 1

$$y_{it} = \beta_0 + \beta_1 * Tr_i + \varepsilon_{it}$$

Where,

y_{it} = The metered energy use of subject i in period t .

β_0 = The average energy use per unit of time for subjects in the control group.

β_1 = The average treatment effect of the program. The energy savings per subject per period equals $-\beta_1$.

Tr_i = An indicator for whether subject i received the treatment. The variable equals 1 for subjects in the treatment group and equals 0 for subjects in the control group.

Some evaluators may be tempted to choose to use random-effects estimation, which assumes time- or subject-invariant factors are uncorrelated with other variables in the model. However, fixed-effects estimation has important advantages over random-effects estimation:

First, it is robust to the omission of any time-invariant regressors. If the evaluator has doubts about whether the assumptions of the random-effects model are satisfied, the fixed-effects estimator is better.

Second, it yields consistent savings estimates when the assumptions of the random-effects model holds. The converse is not true, making the fixed-effects approach more robust.

Because weaker assumptions are required for the fixed effects model to yield unbiased estimates, this protocol generally recommends the fixed effects estimation approach. In remainder of this protocol presents panel regression models that satisfy the fixed-effects assumptions.

²³ For a treatment that is continuous, an example might be $t=1$ on the first day that the treatment starts, $t=2$ on the second day, etc.; for a treatment that only occurs during certain days (e.g., a day when the utility's system peaks), an example might be $t=1$ during the first critical event day, $t=2$ during the second, etc.

ε_{it} = The model error term, representing random influences on the energy use of customer i in period t .

In this simple model, the error term ε_{it} is uncorrelated with Tr_i because subjects were randomly assigned to the treatment or control group. OLS estimation of this model (with standard errors clustered at the subject) will result in an unbiased estimate of β_1 .

This specification does not include subject fixed effects. Since the available energy-use data is for during the treatment period only, it is not possible to identify the program treatment effect and account for subject fixed effects. However, as previously noted, because of the random assignment of subjects to the treatment group, any time-invariant characteristics affecting energy use will be uncorrelated with the treatment, so omitting that type of fixed effects will not bias the savings estimates.

Using Equation 1, however, it would be possible to obtain more precise estimates of savings by replacing the coefficient β_0 with time-period fixed effects. By incorporating time-period fixed effects, the model captures more of the variation in energy use over time, resulting in greater precision. The interpretation of β_1 , the treatment effect per home per time period, is unchanged.

Simple Differences Regression Estimate of Savings during Each Time Period

To estimate the average energy savings from the treatment during each period, the evaluator can interact the treatment indicator with indicators variables for the time periods using the following variation of Equation 1²⁴:

Equation 2

$$y_{it} = \sum_{j=1}^T \beta_t Tr_i * d_{jt} + \sum_{j=1}^T \theta_t d_{jt} + \varepsilon_{it}$$

Where,

β_t = The average savings per subject specific to period t (e.g., the average savings per subject during month 4 or during hour 6).

d_{jt} = An indicator variable for period j , $j=1, 2, \dots, T$. d_{jt} equals 1 if $j=t$ (i.e., the period is the t^{th}) and equals 0 if $j \neq t$ (i.e., the period is not the t^{th}).

θ_t = The average effect on consumption per subject specific to period t .

Estimate equation 2 by including a separate dummy variable and an interaction between that dummy variable and Tr_i for each time period t , where $t=1, 2, \dots, T$. When the time period is in months, refer to the time-period variables as month-by-year fixed effects. The coefficient on the interaction variable for period t , β_t , is the average savings per subject specific to period t . Again,

²⁴ If the number of time periods is very large, the number of time period indicator variables in the regression may overwhelm the capabilities of the available statistical software. Another option for estimation is to transform the dependent variable and all of the independent variables by subtracting time period-specific means and then running Ordinary Least Squares (OLS) on the transformed data.

because ε_{it} is uncorrelated with the treatment after accounting for the average energy use in period t , OLS estimation of Equation 2 (with standard errors clustered at the subject level) results in an unbiased estimate of the average treatment effect during each period.

Difference-in-Differences Regression Model of Energy Use

This section outlines a D-in-D approach to estimating savings from BB field experiments. This protocol recommends D-in-D estimation to the simple differences approach, but it requires information about the energy use of treatment and control group subjects during the pre-treatment period. This energy-use data enables the evaluator to:

- Include subject fixed effects to account for differences between subjects in time-invariant energy use
- Obtain more precise savings estimates
- Test identifying assumptions of the model

Assume there are N subjects and $T + 1$ periods, $T > 0$, in the pre-treatment period denoted by $t = -T, -T+1, \dots, -1, 0$, and T periods in the treatment period, denoted by $t = 1, 2, \dots, T$. A basic D-in-D panel regression with subject fixed effects could be specified as:

Equation 3

$$y_{it} = \alpha_i + \beta_1 P_t + \beta_2 P_t * Tr_i + \varepsilon_{it}$$

Where,

- α_i = Unobservable, time-invariant energy use for subject i . These effects are controlled for with subject fixed effects.
- β_1 = The average energy savings per subject during the treatment period that was not caused by the treatment.
- P_t = An indicator variable for whether time period t occurs during the treatment. It equals 1 if treatment group subjects received the treatment during period t , and equals 0 otherwise.
- β_2 = The average energy savings per subject due to the treatment per unit of time.

The model includes fixed effects to account for differences in energy use between subjects that do not vary over time. Including subject fixed effects would likely explain a significant amount of the variation in energy use between subjects and result in more precise savings estimates. The interaction of P_t and Tr_i equals one for subjects in the treatment group during time periods when the treatment is in effect, and 0 for all other periods.

This regression equation has a D-in-D structure; for control group subject i , the expected energy use is α_i during the pre-treatment period and $\alpha_i + \beta_1$ during the treatment period. The difference in expected energy use, also known as naturally-occurring savings, is β_1 . If that same subject i had been in the treatment group, the expected energy use would have been α_i during the pre-

treatment period and $\alpha_i + \beta_1 + \beta_2$ during the treatment period. The expected savings would have been $\beta_1 + \beta_2$, which is the sum of naturally occurring savings from the BB program. The D-in-D yields β_2 , a D-in-D estimate of program savings. OLS estimation results in an unbiased estimate of β_2 .

D-in-D Estimate of Savings for Each Time Period

By re-specifying Equation 3 with time-period fixed effects, it is possible to estimate savings during each period and test the model assumption that treatment and control group subjects are equivalent except for receiving the treatment. Consider the following D-in-D regression specification:

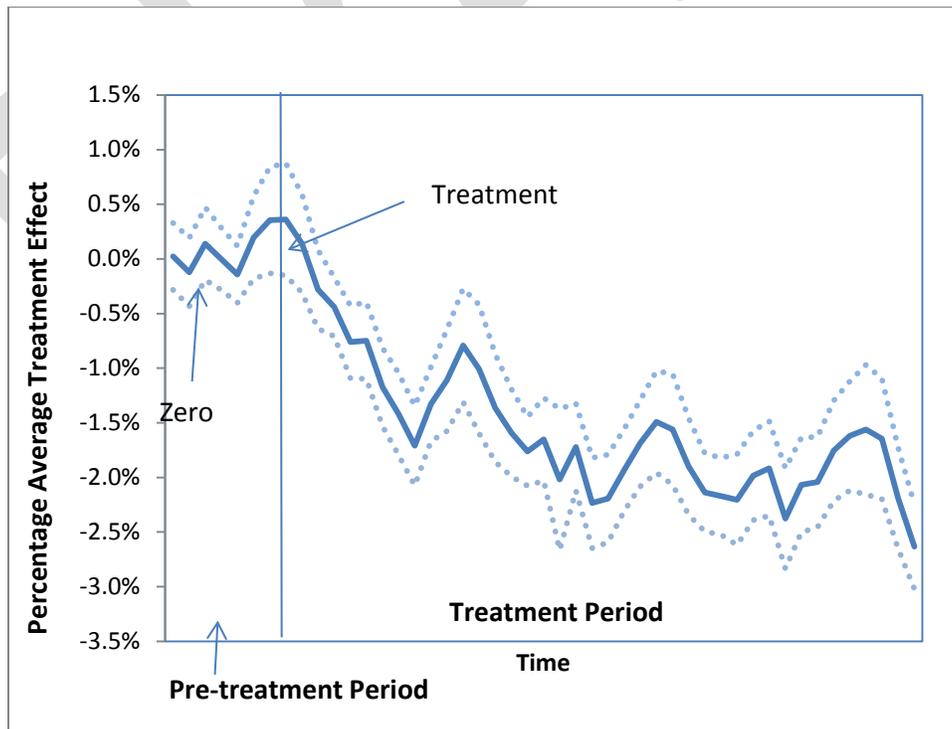
Equation 4

$$y_{it} = \alpha_i + \sum_{j=-T}^T \theta_j d_{jt} + \sum_{j=-T}^{-1} \beta_j Tr_i^* d_{jt} + \sum_{j=1}^T \beta_j Tr_i^* d_{jt} + \varepsilon_{it}$$

Estimate this model by including a separate dummy variable and an interaction between the dummy variable and Tr_i for each time period t , where $t = -T, -T+1, \dots, -1, 0, 1, 2, \dots, T$. The coefficient on the interaction variable for period t , β_t^T , is the D-in-D savings for period t .

Unlike the simple differences regression model, this model yields an estimate of BB program savings during all periods except for one, e.g., $t=0$, for a total of $2T-1$ savings estimates. Figure 4 shows an example of savings estimates obtained from such a model. The dotted lines show the 95% confidence interval for the savings estimates.

Figure 4. Example of Difference-in-Differences Regression Savings Estimates



Use estimates of savings before the treatment to test the assumption of equivalency between treatment and control group subjects. Before utilities administer the treatment, BB programs should not have statistically significant savings without the treatment. BB program pretreatment saving estimates that are statistically different from zero suggest a design flaw in the experiment and that the evaluator incorrectly specified the regression model. For example, an error in the randomization process could have resulted in assignments of subjects to the treatment and control groups that were correlated with their energy use.

Randomized Encouragement Design

Some field experiments involve a RED in which subjects are only encouraged to accept a BB measure, in contrast to RCTs in which a program administers a BB intervention. This section outlines the types of regression models that are appropriate for REDs, how to interpret the coefficients, and how to estimate savings from RED programs.

Evaluators can apply the model specifications previously described for RCTs to REDs. The model coefficients and savings are interpreted differently, however, and it takes an additional step to estimate average savings for subjects who accept an intervention. Treatment in an RED is defined as receiving encouragement to adopt the BB intervention, rather than actually receiving the intervention as with RCTs.

Consider a field experiment with a RED that has energy-use data for treatment and control group subjects available for the periods before and during the treatment. Using Equations 1 through 4, it is possible to estimate the treatment effect, or the average energy-use effect on those receiving encouragement. The estimate only captures savings from compliers, as never takers would never accept the intervention, and always takers would accept the intervention with or without encouragement.

To recover the LATE, the savings from subjects who accept the treatment because of the encouragement, scale the estimate of β_2 must by the inverse of the difference between the percentage of subjects in the treatment group who accept the intervention and the percentage of subjects in the control group who accept the intervention (which is zero if control group subjects are prohibited from accepting the intervention). Estimate this as:

Equation 5

$$\beta_2 / (\pi_T - \pi_C)$$

Where,

π_T = The percentage of treatment group subjects who accept the intervention.

π_C = The percentage of control group subjects who accept the intervention.

Models for Estimating Savings Persistence

A utility offering a residential BB program may want to know what happens to savings during the second or third year of the program or after the program is discontinued. There are two kinds of savings effects to measure: (1) the effect of continuing the intervention on consumption called *savings durability* and (2) the effect on consumption after discontinuing the intervention called *savings persistence*.²⁵

Suppose a utility implemented a BB program as a RCT and wants to measure the persistence of savings after the BB intervention stops. The utility started the treatment in period $t=1$ and administered it for t^* periods. Beginning in period $t = t^*+1$, the utility stopped administering the intervention for a random sample of treated subjects. Evaluators can estimate the average savings c for subjects who continue to receive the treatment (continuing treatment group) and for those who stopped receiving the treatment after period t^* (discontinued treatment group)..

Assuming that pretreatment energy-use data is available, estimate savings persistence and savings durability using the following regression equation:

Equation 6

$$kWh_{it} = \alpha_i + \tau_t + \beta_1 P_{1,t} * Tc_i + \beta_2 P_{1,t} * Td_i + \beta_3 P_{2,t} * Tc_i + \beta_4 P_{2,t} * Td_i + \varepsilon_{it}$$

Where,

τ_t = The time-period fixed effect (an unobservable that affects the consumption of all subjects during time period t). The time period effect can be estimated by including a separate dummy variable for each time period t , where $t = -T, -T+1, \dots, -1, 0, 1, 2, \dots, T$.

β_1 = The average energy savings per continuing subject caused by the treatment during periods $t=1$ to $t=t^*$.

$P_{1,t}$ = An indicator variable for whether subjects in the continued and discontinued treatment groups received the treatment during period t . It equals 1 if period t occurs between periods $t=1$ and $t=t^*$ and equals 0 otherwise.

Tc_i = An indicator for whether subject i is in the continuing treatment group. The variable equals 1 for subjects in the continuing treatment group and equals 0 for subjects not in the continuing treatment group.

β_2 = The average energy savings per discontinuing subject caused by the treatment during periods $t=1$ to $t=t^*$.

Td_i = An indicator for whether subject i is in the discontinuing treatment group. The variable equals 1 for subjects in the discontinuing treatment group and equals 0 for subjects not in the discontinuing treatment group.

²⁵ Allcott and Rodgers (2012) employ this terminology.

β_3 = The average energy savings from the treatment for subjects in the continuing treatment group when $t > t^*$.

$P_{2,t}$ = An indicator variable for whether continuing treatment group subjects received the treatment and discontinued treatment group subjects did not receive the treatment during period t . It equals 1 if period t occurs after $t = t^*$ and equals 0 otherwise.

β_4 = The average energy savings for subjects in the discontinued treatment group when $t > t^*$.

OLS estimation of Equation 6 yields unbiased estimates of the savings durability (β_3) and continuance (β_4) because original treatment group subjects were assigned randomly to the continuing and discontinued treatment groups. Subjects in these groups will be equivalent to each other and to subjects in the control group in every respect except for whether they receive the treatment indefinitely, temporarily, or not at all. Evaluators can expect that $\beta_4 \geq \beta_3$, that is, the average savings of the continuing treatment group will be larger than that of the discontinued treatment group.

Evaluators can test the equivalence of the continuing and discontinued treatment groups before utilities discontinued the reports by comparing their respective savings between period $t=1$ and t^* . If the groups are equivalent, expect their savings during this period to be equal.

Standard Errors

As panel data have multiple energy-use observations for each subject, each subject's energy-use data are very likely to exhibit correlations. Many factors affecting energy use persist over time, and the strength of within-subject correlations usually increases with the frequency of the data. When calculating standard errors for panel regression model coefficients, it is important to account for within-subject energy-use correlations. Failing to do so will lead to savings estimates with overstated precision.

This protocol strongly recommends that evaluators estimate robust standard errors clustered on subjects (the randomized unit in field trials) to account for within-subject correlation. Most statistical software packages, including STATA, SAS, and R, have regression packages that output regression-clustered standard errors.

Clustered standard errors account for having less information about energy use in a panel with N subjects and T observations than in a dataset with $N \cdot T$ independent observations. Because clustered standard errors account for within-subject energy-use correlations, they are typically larger than OLS standard errors. When there is within-subject correlation, OLS standard errors are biased downwards and overstate the statistical significance of the estimated regression coefficients.²⁶

²⁶ Bertrand, Duflo, and Mullainathan (2004) show when D-in-D studies ignore serially correlated errors, the probability of finding significant effects when there are none (Type I error) increases significantly.

Opt Out Subjects and Account Closures

Many BB programs allow subjects to opt out and stop receiving the treatment. This section addresses how evaluators should treat opt out subjects in the analysis, as well as account closures.

As a general rule, include all subjects initially assigned to the treatment or control group in the savings analysis. For example, evaluators should keep opt-out subjects in the BB program analysis. Opt-out subjects may have different energy-use characteristics than subjects who remain in the program, and dropping them from the analysis would result in having treatment and control groups that are no longer equivalent. To preserve equivalence and ensure the internal validity of the savings, opt-out subjects should be kept in the analysis sample.

Sometimes a treatment or control group subject closes their billing account after the program starts. It is usually safe to assume that account closures are unrelated to the BB program or savings; most account closures are due to households moving residences. Subjects in the treatment group will experience account closures for the same reasons and at the same rates as subjects in the control group; for this reason, evaluators can safely drop treatment and control group subjects who close their accounts from the analysis sample.

However, if savings are correlated with the probability of an account closure, it is best to keep subjects with account closures in the analysis sample. For example, if young households, which are the most mobile and likely to close their accounts, are also most responsive to BB programs, then dropping these households from the analysis would bias the savings estimates downward,²⁷ and evaluators should keep these households in the analysis.

If evaluators drop customers who close their accounts during the treatment from the regression estimation, they should still count the savings from these subjects (from during the treatment period before customers closed their accounts). To illustrate, when estimating savings for a one-year BB program, evaluators can estimate the savings from subjects who closed their accounts and from those who did not as the weighted sum of the conditional average program treatment effects in each month:

Equation 7

$$\text{Savings} = \sum_{m=1}^{12} -\beta_m * \text{Days}_m * N_m$$

Where,

m = Indexes the months of the year

$-\beta_m$ = The conditional average daily savings in month m (obtained from a regression equation that estimates the program treatment effect on energy use in each month, such as in Equation 2 or Equation 4)

²⁷ See State and Local Efficiency Action Network (2012), p. 30.

Days_m = The number of days in month m

N_m = The number of subjects with an active account receiving the treatment in month m or in a previous month

This approach assumes that savings in a given month from subjects who end up closing their accounts is equal to savings of subjects who do not end up closing their accounts.

4.4 Energy-Efficiency Program Uplift and Double Counting of Savings

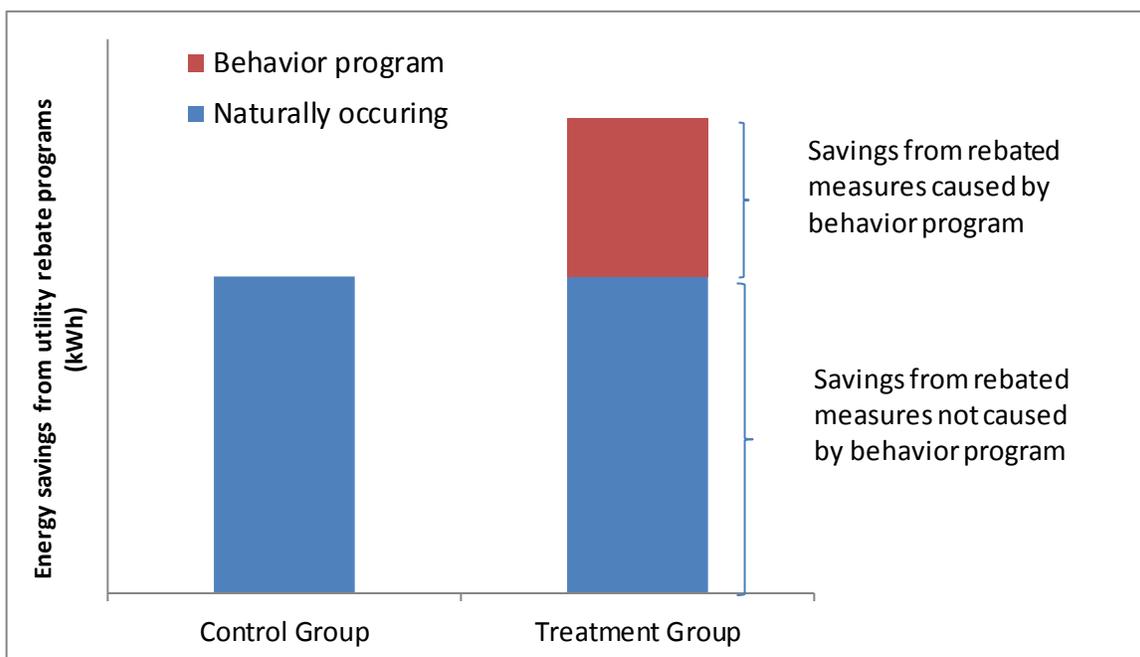
BB programs may increase participation in other utility energy-efficiency programs; this additional participation is known as program uplift. For example, many energy-reports programs encourage BB program participants to purchase efficient measures, such as furnaces, air conditioners, wall insulation, windows, and/or CFLs, in exchange for cash rebates. When a household participates in an efficiency program because of this encouragement, it is possible to count their savings twice: once in the regression-based estimate of BB program savings and again in the estimate of net savings for the efficiency program.

To avoid double-counting savings, it is necessary to estimate and subtract the savings from program uplift from the efficiency program portfolio savings. It is important to understand that the double-counted savings are real and are caused by the BB program as the savings would not have occurred in the program's absence.²⁸ It is important that utilities distinguish between program participation caused by the BB program and those caused by another reason.

It is conceptually straightforward to estimate the double-counted savings from BB programs with experimental research designs. For example, if the treatment and control groups have an equal number of subjects and the utility markets energy-efficiency Measure A to both groups identically through a separate rebate program. Subjects in the treatment group also receive behavioral messaging encouraging them to adopt efficiency measures, including Measure A. Because customers were randomly assigned to the treatment and control groups, and both groups are equivalent except whether they received behavior treatment, evaluators should attribute any differences between the groups in their installation of Measure A to the BB program. Figure 5 illustrates this logic for calculating double-counted savings.

²⁸ The amount of participation in other utility efficiency programs caused by the BB program is dependent on the BB program incentive amount. While the behavior-based program is necessary to cause the uplift, it may or may not be sufficient on its own. Because the incentive amount is typically not randomized, it is unclear whether the incentive program is necessary to cause the uplift; however, it is known for sure that it alone is not sufficient. Program uplift may be greater with larger rebates.

Figure 5. Calculation of Double-Counted Savings



The net savings from Measure A that the BB program and the other utility program count is the difference in the number of treatment and control group subjects adopting Measure A multiplied by the per-unit deemed net savings.

Evaluators can use this concept to estimate savings from program uplift for efficiency measures that the utility tracks at the customer level. Most of the measures that utilities rebate—such as high-efficiency furnaces, windows, insulation, and air conditioners—fit this description. To estimate savings from program uplift, the evaluator should match the BB program treatment and control group subjects to the utility energy-efficiency program tracking data. Next, the evaluator should calculate the difference between savings in the treatment and control groups per subject. That difference is multiplied by the number of treated subjects, accounting for account closures, when the treatment started, and when subjects installed measures.

It is more difficult to estimate savings from program uplift for measures that the utility does not track at the customer level. The most important such measures are high-efficiency lights such as CFLs and LEDs that are rebated through utility upstream programs. Most utilities provide incentives directly to retailers for purchasing these measures, and the retailers then pass on these price savings in the form of retail discounts. In order to estimate BB savings in upstream efficiency programs, it is necessary to collect data on the purchases of rebated measures by treatment and control group subjects. Evaluators can use household surveys for this purpose. However, because the individual difference in the number of upstream measure purchases between treatment and control group subjects may be small, it is necessary to survey a large number of subjects to detect the BB program effect. Instead of surveys, utilities might use algorithms that detect specific end uses through hourly interval data to determine whether lighting subjects installed measures.

5 Reporting

BB program evaluators should carefully document the research design, data collection and processing steps, analysis methods, and plan for calculating savings estimates. Specifically, evaluators should describe:

- The program implementation and the hypothesized effects of the behavioral intervention;
- The experimental design, including the procedures for randomly assigning subjects to the treatment or control group;
- The sample design and sampling process;
- Processes for data collection and preparation for analysis including all data cleaning steps;
- Analysis methods including the application of statistical or econometric models and key assumptions used to identify savings, including tests of those key identification assumptions; and
- Results of savings estimate including point estimates of savings and standard errors and full results of regressions used to estimate savings.

A good rule-of-thumb is that evaluators should report enough detail such that a different evaluator could replicate the results with the study data. Every detail does not have to be provided in the body of the report; provide much of the data collection and savings estimation details in a technical appendix.

6 References

Allcott, H. (2011). Social Norms and Energy Conservation. *Journal of Public Economics*, 95(2), 1082-1095.

Allcott, H., T. Rodgers. (2012). The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation. National Bureau of Economic Research Paper 18492.

Bertrand, M., E. Duflo, and S. Mullainathan. (2004). How Much Should We Trust Difference-in-Differences Estimates? *Quarterly Journal of Economics* 119 (1), 249-275.

Consortium for Energy Efficiency database. (2013). <http://library.cee1.org/content/2013-behavior-program-summary-public-version>

Davis, M. (2011). Behavior and Energy Savings: Evidence from a Series of Experimental Interventions. Environmental Defense Fund Report.

Electric Power Research Institute. (2010). Guidelines for Designing Effective Energy Information Feedback Pilots: Research Protocols. EPRI: Palo Alto, CA. 1020855.

Electric Power Research Institute and Lawrence Berkeley National Laboratories, forthcoming. Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines. LBNL, Berkeley CA and EPRI, Palo Alto, CA.

Greene, W. (2011). *Econometric Analysis*. New Jersey: Prentice Hall.

Harding, M. and A. Hsiaw. (2012). Goal Setting and Energy Efficiency. Stanford University working paper.

Hilbe, J.M. (1993). Sample Size Determination for Means and Proportions. *Stata Technical Bulletin* 11, 17-20.

Ignelzi, P., J. Peters, L. Dethman, K. Randazzo, and L. Lutzenhiser. (2013). Paving the Way for a Richer Mix of Residential Behavior Programs. Prepared for Enernoc Utility Solutions. CALMAC Study SCE0334.01.

List, John A., Sally Sadoff, and Mathis Wagner. (2010). So You Want to Run an Experiment, now What? Some Simple Rules of Thumb for Optimal Experimental Design. National Bureau of Economic Research Paper 15701.

Rosenberg, Mitchell, G. Kennedy Agnew, and Kathleen Gaffney. Causality, Sustainability, and Scalability – What We Still Do and Do Not Know about the Impacts of Comparative Feedback Programs. Paper prepared for 2013 International Energy Program Evaluation Conference, Chicago.

Sacramento Municipal Utility District. (2013). SmartPricing Options Interim Evaluation. Prepared for the U.S. Department of Energy and Lawrence Berkeley National Laboratory by SMUD and Freeman, Sullivan, & Co.

Seed, P.T. (1997). Sample Size Calculations for Clinical Trials with Repeated Measures Data. Stata Technical Bulletin 40, 16-18.

Sergici, Sanem and Ahmad Faruqui. (2011). Measurement and Verification Principles for Behavior-Based Efficiency Programs. San Francisco: The Brattle Group, Inc.

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>.

DRAFT