Chapter 17: Residential Behavior Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

September 2011 – August 2020

This version supersedes the version originally published in January 2015 and subsequent 2017 publication. The content in this version has been updated.

James Stewart[1] and Annika Todd[2]

1 Cadmus, Waltham, Massachusetts
2 Lawrence Berkeley National Laboratory, Berkeley, California

NREL Technical Monitor: Charles Kurnik

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Subcontract Report
NREL/SR-7A40-77435
October 2020

Chapter 17: Residential Behavior Evaluation Protocol

The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures

September 2011 – August 2020

James Stewart[1] and Annika Todd[2]

1 Cadmus, Waltham, Massachusetts
2 Lawrence Berkeley National Laboratory, Berkeley, California

NREL Technical Monitor: Charles Kurnik

Suggested Citation
Stewart, James and Annika Todd. 2020. *Chapter 17: Residential Behavior Evaluation Protocol, The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures: September 2011 – August 2020*. Golden, CO: National Renewable Energy Laboratory. NREL/SR-7A40-77435. https://www.nrel.gov/docs/fy21osti/77435.pdf.

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Contract No. DE-AC36-08GO28308

Subcontract Report
NREL/SR-7A40-77435
October 2020

National Renewable Energy Laboratory
15013 Denver West Parkway
Golden, CO 80401
303-275-3000 • www.nrel.gov

NOTICE

This work was authored in part by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Office of Electricity Delivery and Energy Reliability (OE). The views expressed herein do not necessarily represent the views of the DOE or the U.S. Government.

Cover Photos by Dennis Schroeder: (clockwise, left to right) NREL 51934, NREL 45897, NREL 42160, NREL 45891, NREL 48097, NREL 46526.

Disclaimer

These methods, processes, or best practices ("Practices") are provided by the National Renewable Energy Laboratory ("NREL"), which is operated by the Alliance for Sustainable Energy LLC ("Alliance") for the U.S. Department of Energy (the "DOE").

It is recognized that disclosure of these Practices is provided under the following conditions and warnings: (1) these Practices have been prepared for reference purposes only; (2) these Practices consist of or are based on estimates or assumptions made on a best-efforts basis, based upon present expectations; and (3) these Practices were prepared with existing information and are subject to change without notice.

The user understands that DOE/NREL/ALLIANCE are not obligated to provide the user with any support, consulting, training or assistance of any kind with regard to the use of the Practices or to provide the user with any updates, revisions or new versions thereof. DOE, NREL, and ALLIANCE do not guarantee or endorse any results generated by use of the Practices, and user is entirely responsible for the results and any reliance on the results or the Practices in general.

User agrees to indemnify DOE/NREL/Alliance and its subsidiaries, affiliates, officers, agents, and employees against any claim or demand, including reasonable attorneys' fees, related to user's use of the practices. The practices are provided by DOE/NREL/Alliance "As is," and any express or implied warranties, including but not limited to the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall DOE/NREL Alliance be liable for any special, indirect or consequential damages or any damages whatsoever, including but not limited to claims associated with the loss of profits, that may result from an action in contract, negligence or other tortious claim that arises out of or in connection with the access, use or performance of the practices.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). UMP provides model protocols for determining energy and demand savings that result from specific energy-efficiency measures implemented through state and utility programs. In most cases, the measure protocols are based on a particular option identified by the International Performance Verification and Measurement Protocol; however, this work provides a more detailed approach to implementing that option. Each chapter is written by technical experts in collaboration with their peers, reviewed by industry experts, and subject to public review and comment. The protocols are updated on an as-needed basis.

The UMP protocols can be used by utilities, program administrators, public utility commissions, evaluators, and other stakeholders for both program planning and evaluation.

To learn more about the project, visit https://energy.gov/eere/about-us/ump-home, or download the UMP introduction document at http://www.nrel.gov/docs/fy17osti/68557.pdf.

Acknowledgments

The authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on and previous draft versions of this updated 2020 protocol:[1]

- Ingo Bensch of Evergreen Economics
- Debbie Brannan, Bill Provencher, Kevin Cooney, Carly Olig, and Frank Stern of Guidehouse (formerly, Navigant)
- Mimi Goldberg of DNV GL
- Tim Guiterman and John Backus Mayes of Energy Savvy
- Cheryl Jenkins of Vermont Energy Investment Corporation
- M. Sami Khawaja of Cadmus
- Charlie Buck, Julia Friedman, Matt Frades, Maggie McCarey, Marisa Uchin, and Alessandro Orfei of Oracle Utilities (Opower)
- Julie Michals of Northeast Energy Efficiency Partnerships (now with E4TheFuture)
- Drew Blumenthal of Opinion Dynamics Corporation
- Joe Loper and Mike Rufo of Itron
- Dan Ouellet of BC Hydro
- Jonathan Hoechst of Tetra Tech.

[1] Reviewers' affiliations were those when they provided comments on the draft protocols and may not reflect their current affiliations.

Acronyms

BB	behavior-based
DiD	difference in differences
IPMVP	International Performance Measurement and Verification Protocol
ITT	intent to treat
IV	instrumental variable
LATE	local average treatment effect
LED	light-emitting diode
NREL	National Renewable Energy Laboratory
OLS	ordinary least squares
PG&E	Pacific Gas & Electric
RCT	randomized control trial
RED	randomized encouragement design
SEE Action	State and Local Energy Efficiency Action
TOT	treatment effect on the treated
TRM	technical resource manual
UMP	Uniform Methods Project

Protocol Updates

The original version of this protocol was published in January 2015. The authors updated the protocol in 2017 by making the following changes:

- Incorporated findings from recent research comparing the accuracy of savings estimates from randomized experiments and quasi-experiments
- Presented new developments in the estimation of energy savings from behavior-based programs, including the postperiod-only model with preperiod controls (Allcott 2014)
- Updated the discussion of randomized encouragement designs to emphasize the importance of having large sample sizes or a sufficient proportion of compliers as well as the application of instrumental variables two-stage least squares for obtaining estimates of the local average treatment effect
- Incorporated new research regarding the calculation of statistical power and sizing of analysis samples
- Provided more guidance about estimating impacts of behavior-based programs on participation in other energy efficiency programs
- Edited the text in various places to improve organization or to clarify concepts and recommendations.

The authors updated the protocol again in 2020 by making these changes regarding estimation of energy efficiency program uplift and behavior program savings persistence and measure life:

- Recommended that evaluators use difference in differences rather than simple differences to estimate the lift in efficiency program participation and savings from behavior-based measures
- Overviewed recent research concerning behavior-based savings persistence and measure life
- Presented concepts related to behavior-based savings persistence and measure life and a savings accounting framework to demonstrate how program administrators can perform home energy report (HER) savings accounting with a multiyear measure life
- Provided specific guidance to evaluators about econometrically estimating savings persistence and measure life
- Discussed practical difficulties of estimating savings persistence and measure life.

Table of Contents

List of Figures

List of Tables

1 Measure Description

Residential behavior-based (BB) programs use strategies grounded in the behavioral and social sciences to influence household energy consumption. These may include providing households with feedback about their real-time or historical energy consumption; reframing of consumption information in different ways; supplying energy efficiency education and tips; rewarding households for reducing their energy use; comparing households to their peers; and establishing games, tournaments, and competitions.[2] BB programs often target multiple energy end uses and encourage energy savings, demand savings, or both. Savings from BB programs are usually a small percentage of energy use, typically less than 3%.[3]

Utilities introduced the first large-scale residential BB programs in 2008. Since then, scores of utilities have offered these programs to their customers.[4] Although program designs differ, many share these features:

- They are implemented as randomized experiments wherein eligible homes are 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 analyses of their historical consumption, energy savings tips, and energy efficiency comparisons to neighboring homes, either in personalized home reports or through a web portal, or offer incentives for savings energy.
- They are typically implemented by outside vendors.[5]

Utilities will continue to implement residential BB programs as large-scale, randomized control trials (RCTs); however, 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 (for example, Facebook or other social network competitions to reduce consumption).[6] These programs will create new evaluation challenges and may require different evaluation methods than those presented in this protocol. Quasi-experimental methods require stronger assumptions to yield valid savings estimates and may not measure savings with the same degree of validity and accuracy as randomized experiments.

[2] See Ignelzi et al. (2013) for a classification and descriptions of different BB intervention strategies and Mazur-Stommen and Farley (2013) for a survey and classification of current BB programs. Also, a Minnesota Department of Commerce, Division of Energy Resources white paper (2015) defines, classifies, and benchmarks behavioral intervention strategies.
[3] See Allcott (2011), Davis (2011), and Rosenberg et al. (2013) for savings estimates from residential BB programs.
[4] See the 2018 Consortium for Energy Efficiency (CEE) database for a list of utility behavior programs; it is available for download: https://library.cee1.org/content/2018-behavior-program-summary-public-version
[5] Vendors that offer residential BB programs include Aclara, C3 Energy, ICF, Oracle Utilities (Opower), and Uplight.
[6] The 2018 CEE database includes descriptions of many residential BB programs with alternative designs, such as community-focused programs, college dormitory programs, K-12 school programs, and programs relying on social media.

2 Application Conditions of Protocol

This protocol recommends the use of RCTs or randomized encouragement designs (REDs) for estimating savings from BB programs. A significant body of research indicates that randomized experiments result in unbiased and robust estimates of program energy and demand savings. Moreover, recently evaluators have conducted studies comparing the accuracy of savings estimates from randomized experiments and quasi-experiments or observational studies. These comparisons suggest that randomized experiments produce the most accurate savings estimates.[7]

This protocol applies to BB programs that satisfy the following conditions:[8]

- Residential utility customers are the target.
- Energy or demand savings are the objective.
- An appropriately sized analysis sample can be constructed.
- Treated customers can be identified and accurate energy use measurements for sampled customers are available.
- It must be possible to isolate the treatment effect when measuring savings.

This protocol applies only to residential BB programs. Although the number of nonresidential BB programs is growing, utilities offer more residential BB programs to a much larger number of residential customers.[9] As evaluators accumulate more experience with nonresidential programs, the National Renewable Energy Laboratory (NREL) or others could expand this protocol to cover these programs.

This protocol addresses best practices for estimating energy and demand savings. There are no significant conceptual differences between measuring energy savings and measuring demand savings when interval data are available; 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 and engagement, educating customers about their energy use, or increasing awareness of energy

[7] Allcott (2011) compares RCT difference in differences (DiD) savings estimates with quasi-experimental simple differences and DiD savings estimates for several home energy reports programs. He found large differences between the RCT and quasi-experimental estimates. Also, Baylis et al. (2016) analyzed data from a California utility time-of-use and critical peak pricing pilot program and found that RCT produced more accurate savings estimates than quasi-experimental methods such as DiD and propensity score matching that relied on partly random but uncontrolled variation in participation.

[8] As discussed in the "Considering Resource Constraints" section of the UMP *Chapter 1: Introduction*, small utilities (as defined under U.S. Small Business Administration regulations) may face additional constraints in undertaking this protocol. Therefore, alternative methodologies should be considered for such utilities.

[9] Evaluators may be able to apply the methods recommended in this protocol to the evaluation of some nonresidential BB programs. For example, Pacific Gas and Electric (PG&E) offers a Business Energy Reports Program, which it implemented as an RCT (Seelig 2013). Also, Xcel Energy implemented a business energy reports program as an RCT (Stewart 2013b). Other nonresidential BB programs may not lend themselves to evaluation by randomized experiment. For example, many strategic energy management programs enroll large industrial customers with unique production and energy consumption characteristics for which a randomized experiment would not be feasible (NREL 2017).

efficiency.[10] But these program outcomes could be studied in a complementary fashion alongside the energy savings.

This protocol also requires that the analysis sample be large enough to detect the expected savings with high probability. Because most BB programs result in small percentage savings, a large sample size (often in the thousands or tens of thousands of customers) is required to detect savings. This protocol does not address evaluations of BB programs with a small number of participants.

Finally, this protocol requires that the energy use of participants or households affected by the program (for the treatment and control groups) can be clearly identified and measured. Typically, the analysis unit is the household; in this case, treatment group households must be identifiable and individual household energy consumption must be metered. However, depending on the BB program, the analysis units may not be households. For example, for a BB program that generates an energy competition between housing floors or residential buildings at a university, the analysis unit may be floors or buildings; in this case, the energy consumption of these units must be metered.

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[11]
- The specific behavior-modification theory or strategy
- The channel(s) through which program information is communicated.

Although this protocol strongly recommends RCTs or REDs, it also recognizes that implementing these methods may not always be feasible. Government regulations or program designs may prevent the utilization of randomized experiments for evaluating BB programs. In these cases, evaluators must employ quasi-experimental methods, which require stronger assumptions than do randomized experiments to yield valid savings estimates.[12] If these assumptions are violated, quasi-experimental methods may produce biased results. The extent of the biases in the estimates is not knowable *ex ante*, so results will be less reliable. Because there is currently not enough evidence of quasi-experimental methods that perform well, this protocol refrains from recommending non-RCT evaluation methods. A good reference for applying quasi-experimental methods to BB program evaluation is State and Local Energy Efficiency Action (SEE Action) (2012) or Cappers et al. (2013). As more evidence accumulates about the efficacy of quasi-experiments, NREL may update this protocol as appropriate.

[10] Process evaluation objectives may be important, and omission of them from this protocol should not be interpreted as a statement that these objectives should not be considered by program administrators.
[11] In opt-in programs, customers enroll or select to participate. In opt-out programs, the utility enrolls the customers, and the customers remain 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.
[12] For example, Harding and Hsiaw (2012) use variation in timing of adoption of an online goal-setting tool to estimate savings from the tool.

2.1 Examples of Protocol Applicability

Examples of residential BB programs for which the evaluation protocol applies follow:

- **Example 1**. A utility sends energy reports encouraging conservation to thousands of randomly selected residential customers.
- **Example 2**. A utility sends email or text alerts to residential customers with tips about reducing energy consumption when their energy consumption is on track to exceed normal levels for the billing period.
- **Example 3**. A utility invites thousands of residential customers to use its web portal to track their energy consumption in real time, set goals for energy saving, find ideas about how to reduce their energy consumption, and receive points or rewards for saving energy.
- **Example 4**. A utility sends voice, text, and email messages to thousands of residential utility customers encouraging—and providing tips for— reducing energy consumption during an impending peak demand event.

Examples of programs for which the protocol does not apply follow:

- **Example 5**. A utility uses a mass-media advertising campaign that relies on radio and other broadcast media to encourage residential customers to conserve energy.
- **Example 6**. A utility initiates a social media campaign (for example, using Facebook or Twitter) to encourage energy conservation.
- **Example 7**. A utility runs a pilot program to test the savings from in-home energy-use displays and enrolls too few customers to detect the expected savings.
- **Example 8**. A utility runs a BB program in a large, master-metered college dormitory to change student attitudes about energy use. The utility randomly assigns some rooms to the treatment group and other rooms to the control group.

The protocol does not apply to Example 5 or Example 6 because the evaluator cannot identify who received the messages. The protocol does not apply to Example 7 because too few customers are in the pilot to accurately detect energy savings. The protocol does not apply to Example 8 because energy-use data are not available for the specific rooms assigned to the treatment and control groups.

3 Savings Concepts

The protocol recommends RCTs and REDs to develop unbiased and robust estimates of energy or demand savings from BB programs that satisfy the applicability conditions described in Section 2. Unless otherwise noted, all references in this protocol to savings are to net energy or demand savings.

Section 3.1 defines some key concepts and Section 3.2 describes specific evaluation methods.

3.1 Definitions

The following key concepts are used throughout this protocol.

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

Experimental design.[13] Randomized experiments rely on observing the energy use of subjects who were randomly assigned to program treatments or interventions in a controlled process.

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

Internal validity. Savings estimates are internally valid if the savings estimator is expected to yield an estimate of the causal effect of the program on consumption.

Opt-in program. A program in which customers must enroll themselves. Utilities use opt-in BB programs if the customers must agree to participate, and the utility cannot administer treatment without consent.

Opt-out program. A program in which a utility can automatically enroll customers. Utilities use opt-out BB programs if the utility does not need prior agreement from the customer to participate. The utility can administer treatment without the customer's consent, and customers remain enrolled until they ask the utility to stop the treatment.

Quasi-experimental design. Quasi-experimental designs 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. An RCT uses random variation in which subjects are exposed to the program treatment to obtain an estimate of the program treatment effect. By randomly assigning subjects to treatment, an RCT controls for all factors that could confound measurement of the treatment effect. An RCT is expected to yield an unbiased estimate of program savings. Evaluators randomly assign subjects from a study population to a treatment group or a control group. Subjects in the treatment group receive one program treatment (there could be multiple treatments and treatment groups), whereas subjects in the control group receive no treatment. The RCT ensures that receiving the treatment is uncorrelated with the subjects' pretreatment

[13] When this protocol uses the term randomized experiments, it refers to RCTs or REDs, not other experimental evaluation approaches such as natural experiments or quasi-experiments.

energy use, and that evaluators can attribute any difference in energy use between the groups to the treatment.

Randomized encouragement design. In a RED, 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. The RED yields unbiased estimates of the effect on energy consumption from the encouragement and the effect on energy consumption from participating in the program for subjects who participated because of the encouragement. This latter estimate is known as the local average treatment effect (LATE).

Treatment. A 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 accept an intervention.

Treatment effect. This is the effect of the BB program intervention(s) on energy consumption for a specific population and time period. The treatment effect may persist after the period in which the intervention is administered. This means that for long-running programs, some savings may be attributable to treatments administered in previous periods. Section 4.6 of this protocol addresses BB program savings persistence and measure life.

Treatment group. The experimental group of subjects who received the treatment.

3.2 Randomized Experimental Research Designs

This section outlines the application of randomized experiments for evaluating BB programs. The most important benefit of an RCT or RED is that, if carried out correctly, the experiment results in an unbiased estimate of the program's causal impact.[14] Unbiased savings estimates have internal validity. A result is internally valid if the evaluator can expect 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 about who receives a program intervention. RCTs and REDs yield unbiased savings estimates because they ensure that receiving the program intervention is uncorrelated with the subjects' energy consumption.

Randomized experiments 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 features.[15] Program administrators should exercise caution in applying BB program savings estimates for one population to another or to the same population at a later time, because differences in population characteristics, weather, or naturally occurring efficiency can cause savings to change.

A benefit of randomized field experiments is their versatility: evaluators can apply them to a wide range of BB programs regardless of whether they are opt-in or opt-out programs. Evaluators can apply randomized experiments to any program where the objective is to achieve

[14] List (2011) describes many of the benefits of employing randomized field experiments.
[15] Allcott (2015) analyzes the external validity of savings estimates from evaluations of 111 RCTs of home energy reports programs in the United States and shows that the first utilities implementing the programs achieved higher savings than utilities that implemented such programs subsequently.

energy or demand savings; evaluators can construct an appropriately sized analysis sample; and accurate measurements of the energy consumption of sampled units can be obtained.

Randomized experiments, particularly those with large sample sizes, 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 an RCT or RED to evaluate program savings should depend on several factors, including whether it is an opt-in or opt-out program, the expected number of program participants, and the utility's tolerance for subjecting customers to the requirements of an experiment. For example, using an 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.

Implementing an RCT or RED design requires upfront planning. Program evaluation must be an integral part of the program planning process, as described in the randomized experiment research design descriptions in Section 3.3.

3.3 Basic Features

This section outlines several types of RCT research designs, which are simple but very powerful research tools. The core feature of RCT is the random assignment of study subjects (for example, 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.

Section 3.3.1 outlines some common features of RCTs and discusses specific cases.

3.3.1 Common Features of Randomized Control Trial Designs

The key requirements of an RCT are incorporated into the following steps:

1. **Identify the study population**. The program administrator screens the utility population if the program intervention is only offered to certain customer segments, such as single-family homes. Programs designers can base eligibility on dwelling type (for example, single family, multifamily), geographic location, completeness of recent billing history, heating fuel type, utility rate class, or other energy use characteristics.
2. **Identify the treatment effect the experiment will measure and the measurement approach**. Is the BB treatment designed to reduce peak demand, energy consumption, or both? For what periods will savings be measured? A year? Each month of the year or the sample? Hour of the day?
3. **Determine sample sizes**. The numbers of subjects to assign to the treatment and control groups should depend on the type of randomized experiment (for example, REDs and opt-out RCTs generally require more customers), the hypothesized savings, the variance of consumption, and tolerance for error. The number of subjects assigned to the treatment versus control groups should be large enough to detect the hypothesized program effect with the required probability (the statistical power of the experiment), though it is not necessary for the treatment and control groups to be equally sized. Furthermore, some jurisdictions or program administrators may require savings estimates to achieve certain levels of confidence and precision such as 90% confidence with +/-10% precision. An

experiment may have sufficient statistical power, but not yield estimates that meet the required confidence and precision.[16]

Evaluators can use savings estimation simulations to calculate statistical power and confidence and precision for a sample of a given size. Repeated simulations for different sample sizes can be used to obtain minimum sample sizes for the treatment and control groups that meet the desired level of statistical power and confidence and precision. Program administrators and regulators should specify requirements for statistical significance and/or confidence and precision before a program is designed so evaluators can size the experiment appropriately.[17] It is not uncommon for BB programs with expected savings of less than 3% to require thousands of subjects in the treatment and control groups.[18]

4. **Randomly assign subjects to treatments and control**. Study subjects should be randomly assigned to treatment and control groups. To maximize the credibility and acceptance of BB program evaluations by regulators and program administrators, this protocol recommends that a qualified independent third party perform the random assignment.[19] Also, to preserve the integrity of the experiment, customers must not choose their assignments. The procedure for randomly assigning subjects to treatment and control groups should be transparent and well documented.

5. **Verify equivalence**. 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 a minimum, evaluators should verify that before the intervention there are no statistically significant differences between treatment and control homes in average pretreatment energy consumption and in the distribution of pretreatment energy use. Evaluators should conduct analogous tests using customer demographic and housing characteristics data if such data are available.

6. **Administer the treatment**. The intervention must be administered to the treatment group and withheld from 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 research subject and intervention type, the utility may administer treatment once or repeatedly and for different durations.

---

[16] The number of subjects in the treatment group may also depend on the savings goal for the program.

[17] Evaluators can also use statistical software packages to calculate statistical power as a function of the hypothesized program effect, the coefficient of variation of energy use, the specific analysis approach that will be used (for example, simple differences of means, a repeated measure analysis where there are multiple observations of energy consumption at different time periods for the same subject [i.e., a panel analysis]), and tolerances for Type I and Type II statistical errors. 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 et al. (2010). Most statistical software (including SAS, STATA, and R) now include packages for performing statistical power analyses.

[18] The Electric Power Research Institute (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.

[19] This protocol encourages program administrators to have a third party conduct the random assignment but for some programs it has proven difficult to coordinate with evaluators at the program design stage. In this case, the evaluator should perform an ex-post verification of the random assignment using billing consumption data and demographic data, if available.

However, the treatment period should be long enough for evaluators to observe any effects of the intervention.

7. **Collect data**. Data must be collected from all randomized study subjects, not only from those who chose to participate or only from those who participated for the whole study or experiment.

   Preferably, evaluators should collect multiple pre and post-treatment energy consumption measurements. Such data enable the evaluator to control for time-invariant differences in average energy use between the treatment and control groups to obtain more precise savings estimates. Step 8 discusses this in further detail.

8. **Estimate savings**.[20] Evaluators should calculate savings as the difference in energy consumption or difference in differences (DiD) of energy consumption between the subjects who were initially assigned to the treatment and those assigned to the control group. To obtain an unbiased savings estimate, evaluators must compare the energy consumption 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 chose to participate in the study.

   The difference in energy consumption between the treatment and control groups, usually called an intent-to-treat (ITT) effect, is an unbiased estimate of savings because subjects were randomly assigned to the treatment and control groups. The effect is an ITT because, in contrast to many randomized clinical medical trials, ensuring that treatment group subjects in most BB programs comply with the treatment is impossible. For example, some households may opt out of an energy reports program, or they may fail to notice or open the energy reports. Thus, the effect is ITT, and the evaluator should base the results on the initial assignment of subjects to the treatment group, whether or not subjects actually complied with the treatment.

   The savings estimation should be well documented, transparent, and performed by an independent third party.

3.4 Common Designs

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

3.4.1 Randomized Control Trial 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 work. For example, in energy reports programs, some treated customers may ask the utility to stop sending them reports.

[20] This protocol focuses on estimating average treatment effects; however, treatment effects of behavior programs may be heterogeneous. Costa and Kahn (2010) discuss how treatment effects can depend on political ideology and Allcott (2011) discusses how treatment effects can depend on pretreatment energy use.

Figure 1 depicts the process flow of an RCT in which treated customers can opt out of the program. In this illustration, the utility initially screened its customers to refine the study population.[21]

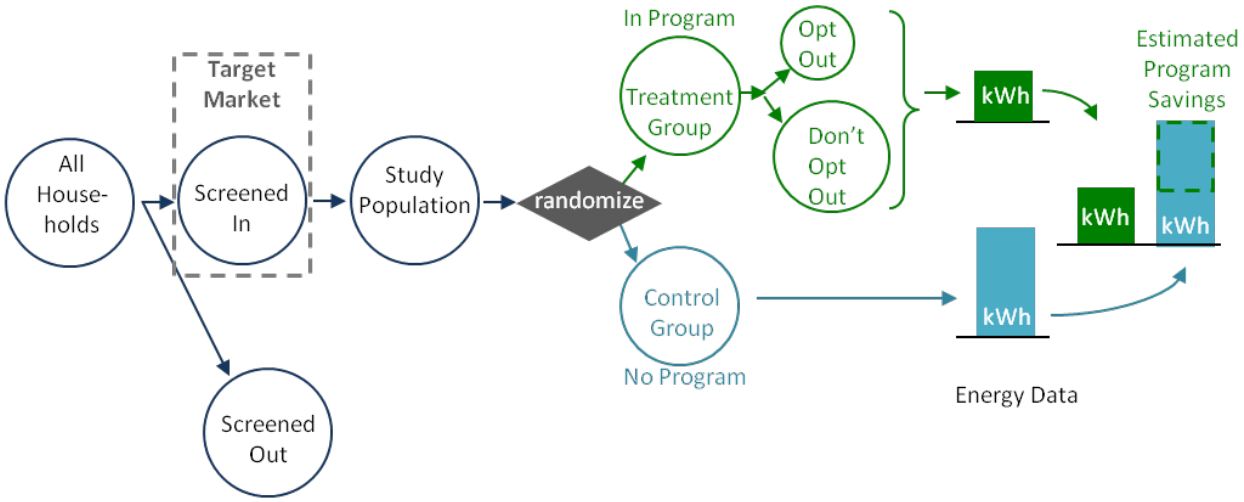

Figure 1. Illustration of RCT with opt-out program design

Customers who pass the screening comprise the study population or sample frame. The ITT savings estimate will apply to this population. Alternatively, the utility may want to study only a sample of the screened population, in which case customers from the population should be sampled randomly. The analysis sample must be large enough to meet the minimum required sizes for the treatment and control groups.

The next steps in an opt-out RCT are to (1) randomly assign subjects in the study population to the program treatment and control groups, (2) administer the program treatments, and (3) collect energy use data.

The distinguishing feature of this randomized 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 obtain unbiased savings estimates. Evaluators can then calculate savings as the difference in average energy consumption between treatment group customers, including opt-out subjects and control group customers. Removing opt-out subjects from the analysis would bias the savings estimate because certain subjects in the control group would have opted out if they had been treated but it is impossible to know who that might be in the control group. The resulting savings estimate is therefore an average of the savings of treated customers who remain in the program and of customers who opted out.

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

[21] This graphic and the following ones are variations of those that appeared in SEE Action (2012). A co-author of the SEE Action report and the creator of that reports' figures is one of the authors of this protocol.

3.4.2 Randomized Control Trial With Opt-In Program Design

Some BB programs require utility customers to enroll before they can be treated. Examples include web-based home audit or energy consumption tools; online courses about energy rates and home efficiency; or in-home displays. These interventions contrast with interventions such as home energy reports that can be administered to subjects without having their prior agreement.

An opt-in RCT (Figure 2) can accommodate the necessity for customers to opt into some BB programs. This design results in an unbiased estimate of the ITT effect for customers who opt into the program. The estimate of savings will have internal validity; however, it will not necessarily have external validity because it will not apply to subjects who do not opt in.

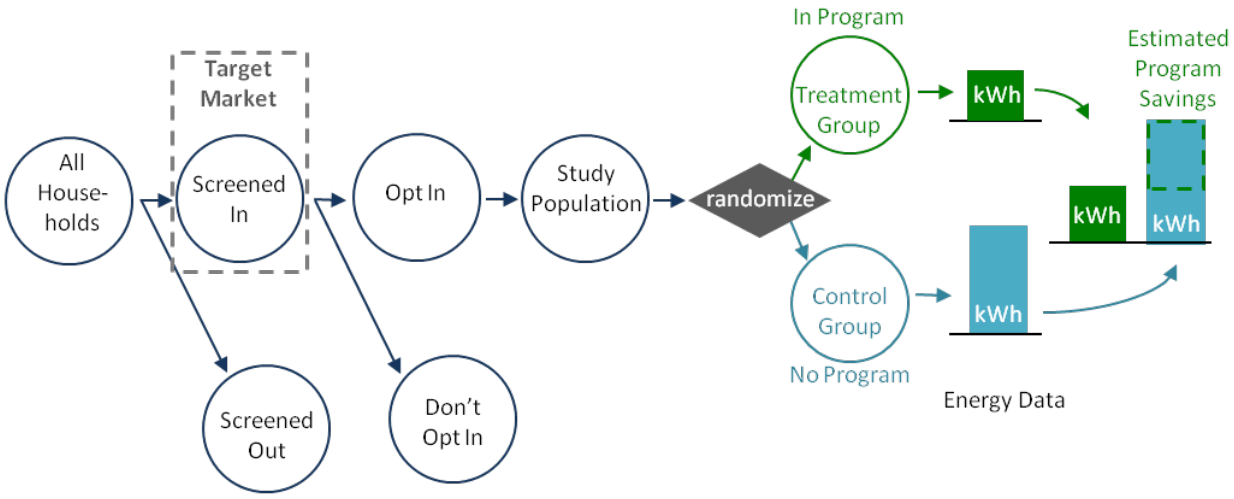

Figure 2. Illustration of RCT with opt-in program design (kilowatt-hour [kWh])

Implementing opt-in RCTs is very similar to implementing opt-out RCTs. The first step, screening utility customers for eligibility to determine the study population, is the same. The next step is to market the program to eligible customers. Some eligible customers may then agree to participate. Then, an independent 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 in the program to customers assigned to the control group. Thus, only customers who opted in and were assigned to the treatment group will receive the treatment.

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 either this equivalence or the internal validity of the savings estimates.

After the random assignment, the opt-in RCT proceeds the same as an RCT with opt-out subjects: the utility administers the intervention to the treatment group. The evaluator collects energy consumption data from the treatment and control groups, then estimates energy savings as

the difference in energy consumption between the groups. The evaluator does not collect energy consumption data for customers who do not opt into the program.

An important difference between the opt-in RCT and opt-out RCT is how to interpret the savings estimates. In the opt-out RCT, the evaluator bases the savings estimate on a comparison of the energy consumption between the randomized 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 in the study population who opted into the program, and the difference in energy consumption represents the treatment effect for customers who opted in to the program. Opt-in RCT savings estimates have internal validity; however, they do not apply to customers who did not opt into the program.

3.4.3 Randomized Encouragement Design

Some BB interventions require participants to opt into treatment but delaying or denying participation to some customers may be undesirable. In this case, neither the opt-out nor the opt-in RCT design would be appropriate, and this protocol recommends an RED. Instead of randomly assigning subjects to receive or not receive the intervention, a third party randomly assigns them to a treatment group that is *encouraged* to accept the intervention (that is, to participate in a program or adopt a measure), or to a control group that does not receive encouragement. Examples of common kinds of encouragement include direct paper mail or email informing customers about the opportunity to participate in a BB program. Customers who receive the encouragement can refuse to participate, and, depending on the program design, control group customers who learn about the program may be able to participate.

The RED yields an unbiased estimate of the effect of encouragement on energy consumption and, depending on the program design, can also provide an unbiased estimate of either the effect of the intervention on customers who accept the intervention because of the encouragement or the effect of the intervention on all customers who accept it. Necessary conditions for a RED to produce an unbiased estimate of savings from the BB intervention is that the encouragement only influence the decision to accept the BB intervention and not energy consumption. For example, the RED must be such that customers who receive a direct mailing encouraging them to log into a website with personalized energy efficiency recommendations only save energy if they decide to log into the site; the mailing itself must not cause the customer to save energy if the customer never logs on. If the encouragement causes customers to save energy, it may be impossible to isolate the savings from the intervention. Programs designed as REDs should design and distribute encouragement materials that will not affect consumption. If evaluators expect that the encouragement will cause energy savings, they could send the similar messaging that excludes the program enrollment option to the control group or to a second randomized treatment group. Evaluators could use the second randomized treatment group to test whether the encouragement produces savings.

Figure 3 illustrates the process flow for a RED program evaluation. As with the RCT with opt-out and opt-in RCT, the first two steps are to identify the sample frame and select a study population. Next, like the RCT with opt out, a third party randomly assigns subjects to a treatment group, which receives encouragement, or to a control group, which does not. For example, a utility might employ a direct mail campaign that encourages treatment group customers to use an online audit tool. The utility would administer the intervention to treatment group customers who opt in. Although customers in the control group did not receive

encouragement, some may learn about the program and decide to sign up. The program design shown in Figure 3 allows for control group customers to receive the behavioral intervention.

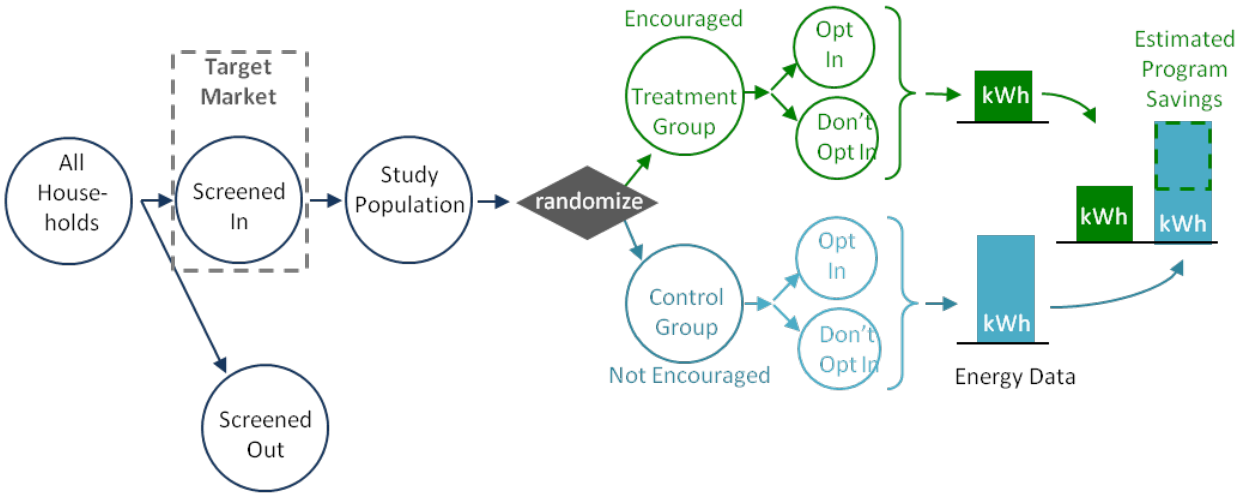

Figure 3. Illustration of RED program design

In Figure 3, the difference in energy consumption 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 consumption to estimate savings for customers who accept the intervention because of the encouragement. 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 accept the intervention only if encouraged. Compliers participate only after receiving the encouragement.

Because eligible subjects are randomly assigned to groups depending on whether they receive encouragement, the treatment and control groups are expected to have equal frequencies of always takers, never takers, and compliers. After treatment, the only difference between the treatment and control groups is that compliers in the treatment group accept the treatment and compliers 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 consumption between the groups reflects the treatment effect of encouragement on compliers (known as LATE).

Furthermore, for the study to have enough statistical power to detect the expected effect, there must be very large encouraged and nonencouraged groups relative to a RCT or quasi-experimental design and/or a high proportion of compliers in the treatment group; a power calculation should be done to ensure that there are enough customers in the encouraged and nonencouraged groups to produce significant savings estimates for the expected take-up rate.[22]

To estimate the effect of the intervention on compliers, evaluators can either employ instrumental variables (IV), using the random assignment of customers to receive encouragement as an instrument for the customer's decision to accept the intervention (that is, to participate).

[22] For an example of a power calculation for REDs, see Fowlie (2010).

The IV approach is presented in Section 4.3. Another option is that evaluators can scale the treatment effect of the encouragement by the difference between treatment and control groups in the percentage of customers who receive the intervention (note that in this equation, if the nonencouraged customers are not allowed to take up the treatment, the second term in the denominator will be zero):[23]

$$1/(\%\text{ of encouraged customers who accepted} - \%\text{ of nonencouraged customers who accepted}). \quad (1)$$

If customers in the control group are permitted to participate if they find out about the treatment even though they did not receive encouragement, the LATE does not capture the program effect on always takers. (Note, however, in most programs, the control group is not permitted to take up the treatment). If customers in the control group are permitted to participate, the LATE may differ from the average treatment effect unless the savings from the intervention is the same for compliers and always takers. However, the LATE will be equal to the average treatment effect if the control group customers (nonencouraged customers) are not permitted to take up the treatment.

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.[24]

Successful application of a RED requires that compliers comprise a percentage of the encouraged population that is sufficiently large given the number of encouraged customers.[25] If the RED generates too few compliers, the effects of the encouragement and receiving the intervention will not be precisely estimated. Therefore, before employing a 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 a RED generating too few compliers is significant, evaluators may want to consider alternative approaches, including quasi-experimental methods.

[23] This approach of estimating savings from the intervention because of encouragement assumes zero savings for customers who received encouragement but did not accept the intervention. If encouraged customers who did not accept the intervention reduced their energy use in response to the encouragement, the savings estimate for compliers will be biased upward.

[24] If the effect of program participation is the same for compliers as for others, those who would have participated without encouragement (always takers) and those who do not participate (never takers), then the RED will yield an unbiased estimate of the population average treatment effect.

[25] For an example of the successful application of a RED, see Sacramento Municipal Utility District (2013).

3.5 Evaluation Benefits and Implementation Requirements of Randomized Experiments

This protocol strongly recommends the use of randomized field experiments (RCTs or REDs) for evaluating residential BB programs. Table 1 summarizes the benefits and requirements of evaluating BB programs using RCTs and REDs, as described in Sections 3.1–3.4.

Table 1. Benefits and Implementation Requirements of Randomized Experiments

Evaluation Benefits	Implementation Requirements
• Yield unbiased, valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • If experiment is sized correctly, provides high confidence that study will result in measurement of treatment effects of interest • Yield savings estimates that are robust to changes in model specification • Are versatile, and can be applied to opt-out and opt-in BB programs • Are widely accepted as the "gold standard" of good program evaluations • Result in transparent analysis and evaluation • Can be designed to test specific research questions such as persistence of savings after treatment ends	• An appropriately sized analysis sample • Accurate energy consumption measurements for sampled units • Advance planning and early evaluator involvement in program design • Restricted program participation or marketing to randomly selected customers

The principal benefit of randomized experiments is that they yield unbiased and robust estimates of program savings or other treatment effects. They are also versatile, widely accepted, and straightforward to analyze. The principal requirements for implementing randomized experiments include the availability of accurate energy consumption measurements and a sufficiently large study population.[26]

Also, this protocol specifically recommends REDs or RCTs for estimating BB program savings as both designs yield unbiased savings estimates. The choice of RED or RCT will depend primarily on program design and implementation considerations, in particular, whether the program has an opt-in or opt-out design. RCTs work well with opt-out programs such as residential energy reports programs. Customers who do not want to receive reports can opt out without adversely affecting the evaluation. RCTs also work well with opt-in programs, for which customer participation can be delayed (for example, customers are put on a "waiting list") or denied. For situations in which delaying or denying a certain subset of customers is impossible or costly, REDs may be more appropriate. REDs can accommodate all interested customers, but have the disadvantages of requiring larger analysis samples, two analysis steps to yield a direct

[26] A frequent objection to the use of randomized experiments is that some utility customers may not have the opportunity to participate in a program. However, programs are often limited to a certain subset of customers; for example, a program may start out as limited to customers in a certain county or other geographic location. REDs allow any customers who would like to participate the opportunity to do so, even if they are in the control group. In our view, limiting the availability of the program to certain customers in RCTs is done with the worthy objective of advancing the utility's knowledge of program savings effects and making future allocation of scarce efficiency resources more optimal.

estimate of the behavioral intervention's effect on energy use, and a high proportion of compliers among encouraged customers.

Table 2 lists some issues to consider when choosing a RCT or RED.

Table 2. Considerations in Selecting a Randomized Experimental Design

Experimental Design	Evaluation Benefits	Implementation and Evaluation Requirements
RCT	• Yields unbiased, robust, and valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • Simple to understand • Works well with opt-out programs • Works well with opt-in programs if customers can be delayed or denied	• May require delaying or denying participation of some customers if program requires customers to opt in
RED	• Yields unbiased, robust, and valid estimates of causal program impacts, resulting in a high degree of confidence in the savings • Can accommodate all customers interested in participating • Works well with opt-in and opt-out programs	• More complex design and harder to understand • Requires a more complex analysis • Requires larger analysis sample • Requires a proportion of compliers that is sufficient given the number of encouraged customers to estimate savings • Encouragement to participate should not cause customers to save energy

4 Savings Estimation

Energy savings for a household in a BB program is the difference between the energy the household consumed and the energy the household would have consumed if it had not participated. However, the energy consumption of a household cannot be observed under two different states. Instead, to estimate savings, evaluators should compare the energy consumption of households in the treatment group to that of a group of households that are statistically the same but did not receive the treatment. In a randomized experiment, assignment to the treatment is random; thus, evaluators can expect control group subjects to consume, on average, the same amount of energy that the treatment group would have consumed without the treatment. The difference in their energy consumption will therefore be an unbiased estimate of energy savings.

Savings can be estimated using energy consumption data from the treatment period only or from before and during the treatment. If energy consumption data from only the treatment period are used, evaluators estimate the savings as a simple difference. If data on energy consumption before treatment is administered are available, evaluators can estimate the savings as a DiD or a simple difference that controls for pretreatment energy consumption. The approach that estimates savings conditional on pretreatment consumption is sometimes referred to as a "post-only model with pre-period controls."[27] The availability of energy consumption data for the period before the treatment will determine the approach, but incorporating pretreatment consumption data in the analysis is strongly advised when such data are available.

Both approaches result in unbiased estimates of savings (that is, in expectation, the two methods are expected to yield an estimate equal to the true savings). However, estimators using pretreatment data generally result in more precise savings estimates (that is, the estimators using pretreatment data will have a smaller standard error) as they account for time-invariant energy use that contributes significantly to the variance of energy consumption between utility customers.[28]

Evaluators should collect at least one full year of historical energy use data (the 12 months immediately before the program start date) to ensure baseline data fully reflect seasonal energy use effects.

Regulators usually determine the frequency of program evaluation. Although requirements vary between jurisdictions, most BB programs are evaluated once per year. Annual evaluation will likely be necessary for the first several years of many BB programs such as HER programs because savings tend to increase for several years before leveling off. However, some program

[27] The model with pretreatment consumption control variables is a more efficient estimator (that is, it is expected to have smaller variance) than the DiD estimator when the model errors are independent and identically distributed or when serial correlation of consumption is low (Burlig, Preonas, and Woerman 2017). This model is more efficient because it uses one degree of freedom rather than multiple degrees of freedom—one for each study subject—to account for between-subject differences in consumption. However, when serial correlation of customer consumption is high, there is little or no gain in efficiency over the fixed effects in the DiD approach.

[28] Postonly or DiD estimation with customer fixed effects also accounts for differences in mean energy use between treatment and control group subjects that are introduced when subjects are randomly assigned to the treatment or control group. Evaluators may not expect such differences with random assignment; however, these differences may nevertheless arise.

administrators may desire measurement or evaluation more frequently than annually to closely track program performance and optimize the program delivery.

4.1 International Performance Measurement and Verification Protocol Option

This protocol's recommended evaluation approach aligns best with International Performance Measurement and Verification Protocol (IPMVP) Option C, which recommends statistical analysis of data from utility meters for whole buildings or facilities to estimate savings. Option C is intended for projects with expected savings that are large relative to consumption. This protocol recommends regression analysis of residential customer consumption and statistical power analysis to determine the analysis sample size necessary to detect the expected savings.

4.2 Sample Design

Utilities should integrate the design of the analysis sample with program planning, because numerous considerations, including the size of the analysis sample, the method of recruiting customers to the program, and the type of randomized experiment, must be addressed before the program begins.

4.2.1 Sample Size

The analysis sample should be large enough to detect the minimum hypothesized program effect with desired probability.[29] If the sample is too small, evaluators risk being unable to detect the program's effect and possibly wrongly accepting an hypothesis of no effect or there may be substantial uncertainty about the program's effect at the end of the study, and it may be necessary to repeat the study with a larger sample. On the other hand, if the sample size is too large, researchers may risk wasting scarce program resources.[30] Oversizing the sample is primarily a concern for pilot programs, for which determining the savings is often a primary objective.

To determine the minimum number of subjects required and the number of subjects to be assigned to the treatment and control groups, researchers should employ a statistical power analysis. Statistical power is the likelihood of detecting a program impact of minimum size (the minimum detectable effect). Typically, researchers design studies to achieve statistical power of 80% or 90%. A study with 80% statistical power has an 80% probability of detecting the hypothesized treatment effect.

Statistical power analysis can be conducted in two ways. First, if data on consumption or another outcome of interest before treatment are available for the study population, researchers can use simulation to estimate the probability of detecting an effect of a certain size (for example, 1%) for possible treatment and control groups sizes, N_T and N_C.

[29] A program can comprise a collection of randomized cohorts or waves in which the treatment effect of interest is at the program level and not at the level of individual cohorts. In this case, power calculations and tests of statistical significance can be applied to the collection of cohorts. Examples of this design include behavioral programs that consist of several waves launched over time or rolling enrollment waves.

[30] The utility may also base the number of subjects in the treatment group on the total savings it desires to achieve.

Simulation follows these steps:

1. Researchers should divide the pretreatment sample period into two parts, corresponding to a simulation pretreatment and post-treatment period. For example, an evaluator with monthly billing consumption data for 24 pretreatment months could divide the pretreatment period into months 1 to 12 and months 13 to 24 and designate the first 12 months as the simulation pretreatment period.
2. From the eligible program population, researchers should randomly assign N_T subjects to the treatment group and N_C subjects to the control group.
3. Researchers should decide upon the minimum detectable treatment effect (for example, 2 kWh/period/subject), and a distribution of treatment effects (for example, normal distribution with mean 2 and standard deviation 1). For each treatment customer, the researcher should simulate the program treatment effect, taken randomly from the distribution of treatment effects, during the simulation treatment period. (One could also assume the treatment effect is the same for all customers and merely apply the same effect to all households; however, the power calculation is likely to underestimate the number of households needed because it assumes zero variance for the treatment effect).
4. Researchers should randomly sample with replacement N_T customers from the treatment group and N_C subjects from the control group.
5. Researchers should estimate the program treatment effect for the sample only using data from the simulation pretreatment and simulation post-treatment periods and record the estimate and whether the estimate was statistically significant for a given Type 1 error.
6. Researchers should repeat steps 4 and 5 many times (for example, >250), and calculate the percentage of iterations when the estimated treatment effect was statistically different than zero. This is the statistical power of the study, the probability of detecting savings of *x* with treatment group size N_T and control group size N_C.

It is important that the estimation method used in the statistical power simulation adheres as closely as possible to the method evaluators plan to use for the actual savings estimation. Otherwise, the statistical power analysis may be misleading about the likelihood of detecting the savings.

The second approach to calculating statistical power uses analytic formulas. Researchers employing panel data methods and using statistical power formulas are advised to use the formulas in Burlig et al. (2017). Though more demanding to implement than those in Frison and Pocock (1992), the statistical power formulas in Burlig et al. (2017) are more accurate because they account for both intracluster correlations and arbitrary serial correlations of customer consumption over time. The required inputs for the power calculation are:

- The minimum detectable treatment effect
- The coefficient of variation of energy use, taken from a sample of customers
- The specific analysis approach to be used (for example, simple differences of means or a repeated measure analysis)

- The numbers of pretreatment and post-treatment observations per subject
- The tolerances for Type I and Type II statistical errors (as discussed in Section 3.3)
- The intracluster correlation of an individual subject's energy use or error term covariances for pretreatment and post-treatment periods and between periods.

Many statistical software applications, including SAS, STATA, and R, include packages for performing statistical power analyses.

Researchers conducting statistical power analyses should keep in mind the following:

- For a given program population, statistical power will be maximized if 50% of subjects are assigned to the treatment group and 50% are assigned to the control group. However, especially for large programs, researchers may obtain acceptable levels of statistical power with unbalanced treatment and control groups. The principal benefit of a smaller control group is that more customers are available to participate in the program.
- If the BB program will operate for more than several months and repeated measurements are planned, researchers should adjust the required sample sizes to account for attrition (the loss of some subjects from the analysis sample because of account closures or withdrawal from the study).

Finally, many studies will not estimate statistically significant savings. This null result could mean that the program did not save energy or that the evaluation did not detect the savings. During the program and evaluation design phase, if clear guidelines are not already available, program administrators, regulators, and evaluators should reach agreement about how statistically insignificant savings estimates should be treated and reported and whether all or some savings based on such estimates can be claimed.

4.2.2 Random Assignment to Treatment and Control Groups by an Independent Third Party

After determining the appropriate sizes of the treatment and control group samples, researchers should randomly assign subjects to the treatment and control groups. For the study to have maximum credibility and acceptance, this protocol recommends that an independent and experienced third party, such as an independent evaluator, perform the randomization. If there is a significant risk that the random assignment will result in unbalanced treatment and control groups with statistically different consumption, this protocol recommends that evaluators first stratify the study population by pretreatment energy consumption levels and then randomly assign subjects in each stratum to treatment and control groups. Stratifying the sample will increase the likelihood that treatment and control group subjects have similar pretreatment means and variances.[31]

This protocol also recommends that the unit of analysis (for example, a household) should be the basis for random assignment to treatment or control group. For example, in an analysis of individual customer consumption, it is better to randomly assign individual customers instead of all customers in the same neighborhood (for example, in a zip code or census block) to receive the treatment. However, for some BB programs, it may not be feasible to randomize the unit of

[31] Shadish et al. (2002) discuss the benefits of stratified random assignment. Bruhn and McKenzie (2009) compared stratified random assignment and re-randomization methods and found that stratification is superior.

analysis. For example, in some multifamily housing BB programs, the unit of analysis may be individual customers but all customers in the same multifamily building may receive the treatment. In this case, it will be necessary to randomly assign multifamily buildings to the treatment or control group. In this case, researchers will need to account for correlations in consumption between customers in the same housing units.

Although this protocol recommends that an independent and experienced third party perform the random assignment, circumstances sometimes make this impossible. In such cases, a third-party evaluator should verify that the assignment of treatment and control group subjects was done correctly and did not introduce bias into the selection process.

4.2.3 Equivalency Check

The third party performing the random assignment must verify that the characteristics of subjects in the treatment group, including pretreatment energy consumption, are balanced with those in the control group. If subjects in the groups are not equivalent, the energy savings estimates may be biased. Evaluators should perform two equivalency checks: (1) for all customers who were randomly assigned to the treatment and control groups; and (2) for all randomized customers who remain in the analysis sample after data cleaning and preparation are completed. Ideally, the consumption data used for the equivalency checks should cover 12 months preceding the start treatment and equivalency should be checked for each month of the year.

To verify the equivalence of energy consumption, this protocol recommends that the third-party test for differences between treatment and control group subjects in both the mean pretreatment period energy consumption and in the distribution of pretreatment energy consumption. Evaluators should attempt to verify equivalence of energy consumption using the same frequency of data to be used in the savings analysis. For example, evaluators should use hour interval consumption data to verify equivalence if the study objective is to estimate peak hour energy savings. Evaluators should also test for differences in other available covariates, such as energy efficiency program participation, home floor area, heating fuel type, and customer demographics. These tests can be used to further demonstrate that the treatment and control groups are well-balanced, as would be expected if assignment to treatment or control group was random. Evaluators can use t-tests or the following regression equation of energy consumption to verify the randomization.

Suppose the evaluator has monthly billing consumption data for all treatment and control group customers for the 12 months, m, m=1, 2, … 12, before treatment began.

$$y_{im} = \Sigma^{12}_{m=1}\beta_{1m} * Tr_i + \mu_m + \varepsilon_{im} \quad (2)$$

where:

y_{im} = The metered energy consumption of subject i in month m

β_{1m} = The average difference in daily energy consumption between the treatment and control groups in month m of the pretreatment period

Tr_i = An indicator for whether subject i was randomly assigned to receive the treatment; the variable equals 1 for subjects in the treatment group and equals 0 for subjects in the control group

μ_m = A month-year fixed effect; the model controls for the month-year fixed effects with a separate intercept for each month, which represents the average daily consumption of the control group in month m

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

In this simple model, the coefficient β_{1m} provides an estimate of the difference in average daily consumption between the treatment and control group in month m of the pretreatment period. Because of the random assignment to treatment, it is expected that the differences will be close to zero and statistically insignificant. Ordinary least squares (OLS) estimation of this model will result in an unbiased estimate of β_{1m}. The standard errors should be clustered on the customer or subject.[32]

Evaluators can check for differences in time-invariant (e.g., demographic or home) characteristics between treatment and control group customers by replacing the dependent variable with the time-invariant characteristics and replacing the month-year fixed effects with a constant β_0 and $\Sigma^{12}_{m=1}\beta_{1m}*Tr_i$ with β_1*Tr_i. The coefficient β_1 will measure the average difference between the treatment and control groups.

If significant differences are found, and it is possible to perform the random assignment again before treatment starts, the third party should consider doing so. Ideally, random assignment should not result in differences; however, differences occasionally appear, and it is better to redo the random assignment than to proceed with unbalanced treatment and control groups, which may lead to biased savings estimates.[33] As noted in Section 4.2.2, stratifying the study population by pretreatment energy use will increase the probability that the groups are balanced.

If the evaluator is not the third party who performed the random assignment, they should perform an equivalency check before estimating the savings. The evaluator may be able to use statistical methods to control for differences in pretreatment energy consumption found after the program is underway.[34]

4.3 Data Requirements and Collection

4.3.1 Energy Use Data

Estimating BB program impacts using a field experiment requires collecting energy consumption data from subjects in the analysis sample. This protocol recommends that evaluators collect

[32] Although the methods recommended in this protocol minimize the potential for violations of the assumptions of the classical linear regression model, evaluators should be aware of—and take steps to minimize—potential violations. The clustering of standard errors accounts for the correlation of individual customer consumption across time periods. In general, it is incorrect to treat observations of a customer's consumption readings as being independent of one another.

[33] Evaluators should keep in mind that at a statistical significance level of 10%, it is expected that statistically significant differences from random assignment will be found 10% of the time as a result of random chance.

[34] 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.

multiple energy consumption measurements for each sampled unit for the periods before and during the treatment.[35]

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

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

- **Relative ease of collection**. Collecting multiple energy consumption measurements for each sampled unit from utility billing systems is usually easy and inexpensive.
- **Can estimate savings during specific times**. If the panel collects enough energy consumption observations per sampled unit, estimating savings at specific times during the treatment period may be possible. For example, hourly energy consumption data may enable the estimation of precise savings during utility system peak hours. Monthly energy consumption 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 because they may be able to control for the time-invariant differences in energy consumption between subjects that contribute to higher variance.
- **Allows for smaller analysis samples**. All else being equal, fewer units are required to detect a minimum level of savings in a panel study than in a cross-section analysis. Thus, collecting panel data may enable studies with smaller analysis samples and data collection costs.

Using panel data has some disadvantages relative to a single measurement per household in a cross-sectional analysis. First, evaluators must correctly cluster the standard errors within each household or unit (as described in the following section). Second, panel data generally require statistical software to analyze, whereas estimating savings using single measurements in a basic spreadsheet software program may be possible.

This protocol also recommends that evaluators collect energy consumption data for the duration of the treatment to ensure they can observe the treatment effect for the entire study period. Ideally, an energy efficiency BB program will last for a year or more because the energy end uses affected by BB programs may vary seasonally. For example, these programs may influence weather-sensitive energy end uses, such as space heating or cooling, so collecting less than one year of data may yield incomplete results.

4.3.2 Makeup of Analysis Sample

Evaluators must collect energy consumption measurements for every household or unit that is initially assigned to a control or treatment group, whether or not the household or unit later opts

[35] A single measurement of energy use for each sampled unit during the treatment period also results in an unbiased estimate of program savings. The statistical significance of the savings estimate depends on the variation of the true but unknown savings and the number of sampled units.

out. Not collecting energy consumption data for opt-out households will result in imbalanced treatment and control groups and could bias the savings estimates.

4.3.3 Other Data Requirements

Program information about each participant must also be collected. Evaluators will need to collect data on customer assignments to the treatment or control group and when the treatments began. Evaluators must have this information to accurately construct regression analysis model variables and to estimate savings. Also, depending on the research design and evaluation objectives, evaluators may also want to collect data on how many and when individual treatments were administered, if and when customers opted out, or details about the specific information included in the treatment. For example, evaluators will need information about the number of reports delivered to customers to estimate the impact of varying the number of delivered reports. Information about how many and which customers opted out may be helpful for evaluating opt-out behavior programs when the opt-out rate is high. The treatment effect for customers who received treatment (LATE) may be different than the ITT effect.

Temperature and other weather data may allow for more precise savings estimates but are often not necessary for estimating savings. Typically, researchers can use dummy variables for individual time periods to account for the effect of weather on household energy consumption. In a regression with time period fixed effects, weather data will improve the precision of the savings estimates only if there is significant variance between customers in weather. If weather data will be collected, evaluators should obtain them from the weather station nearest to each household.

4.3.4 Data Collection Method

Energy use measurements used in the savings estimation should be collected directly from the utility, not from the program implementer, at the end of the program evaluation period. Depending on the program type, utility billing system, and evaluation objectives, the data frequency can be at 15-minute, 1-hour, daily, or monthly intervals.

4.4 Analysis Methods

This protocol recommends using panel regression analysis to estimate savings from BB field experiments where subjects were randomly assigned to either treatment or control groups. Panel regression analysis is preferred to calculating savings differences of unconditional mean energy use, because regression results in more precise savings estimates. A significant benefit of randomized field experiments is that regression-based savings estimates are usually quite insensitive to the type of model specification.

Section 4.3.1 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.

4.4.1 Panel Regression Analysis

In panel regressions, the dependent variable is usually the energy use of a subject (a utility customer home, apartment, or 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. DiD 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 consumption specific to a subject that does not vary over time. For example, home fixed effects may capture variation in energy consumption that is caused by differences such as home sizes or makeup of a home's appliance stock. Time-period fixed effects capture unobservable energy consumption specific to a time period that does not vary between subjects. Including time or subject fixed effects in a regression of energy consumption of subjects randomly assigned to the treatment or control group will increase the precision but not the expected unbiasedness of the savings estimates.[36]

Fixed effects can be incorporated into panel regression in several ways, as follows:

- 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 average energy use. This approach, known as least squares dummy variables, may, however, not be practical for evaluations with a large number of subjects, because the model requires thousands of dummy variables that may overwhelm available computing resources.
- Transform the dependent variable and all independent variables (except for the fixed effects) by subtracting the subject-specific mean of each variable from the variable and then running OLS on the transformed data.[37] This approach is equivalent to least squares dummy variables.
- 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.

[36] Standard econometric formulations assume that fixed effects account for unobservable factors that are correlated with one or more independent variables in the model. This correlation assumption 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, fixed effects must be included to avoid omitted variable bias. 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. Although fixed effects regression is unnecessary, it will increase precision by reducing model variance.
Some evaluators may be tempted 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: (1) 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; and (2) 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. The remainder of this protocol presents panel regression models that satisfy the fixed-effects assumptions.
[37] Greene (2011) Chapter 11 provides more details.

4.4.2 Panel Regression Model Specifications

This section outlines common regression approaches for estimating treatment effects from residential BB programs. Unless otherwise stated, assume that the BB program was implemented as an RCT or RED field experiment.

4.4.3 Simple Differences Regression Model of Energy Use

Consider a BB program in which the evaluator has energy consumption data for the treatment period only and wishes to estimate the average energy savings per period from the treatment. Let t, t = 1, 2, …, T, denote the time periods during treatment for which data are available,[38] and let i, i = 1, 2, …, N, denote the treatment and control group subjects in the analysis sample. For simplicity, assume that all treated subjects started the treatment at the same time.

A basic specification to estimate the average energy savings per treated customer per period is:

$$y_{it} = \beta_0 + \beta_1 * Tr_i + \varepsilon_{it} \quad (3)$$

where:

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

β_0 = The average energy consumption 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

ε_{it} = The model error term, representing random influences on the energy consumption 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. The OLS estimation of this model will result in an unbiased estimate of β_1. The standard errors should be clustered on the subject (customer).[39]

This specification does not include subject fixed effects. Because the available energy consumption data only apply to the treatment period, it is not possible to identify the program treatment effect and to incorporate subject fixed effects into the model. However, as previously noted, because of the random assignment of subjects to the treatment group, any time-invariant

[38] 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, and so on; for a treatment that occurs during certain days only (for example, 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, and so on.
[39] Although the methods recommended in this protocol minimize the potential for violations of the assumptions of the classical linear regression model, evaluators should be aware of—and take steps to minimize—potential violations.

characteristics affecting energy use will be uncorrelated with the treatment, so omitting that type of fixed effects will not bias the savings estimates.

However, in Eq. 2, more precise estimates of savings could be obtained by replacing the coefficient β_0 with time-period fixed effects. The model would capture more of the variation in energy consumption over time, resulting in greater precision in the savings estimate. The interpretation of β_1, the average treatment effect per home per time period, is unchanged.

4.4.4 Simple Differences Regression Estimate of Heterogeneous Savings Impacts

Suppose that the evaluator still has energy consumption data that apply to the treatment period only, but wishes to obtain an estimate of savings from the treatment as a function of some exogenous variable, such as preprogram energy consumption, temperature, home floor space, or pretreatment efficiency program participation (to determine, for example, whether high energy users save more or less energy than low energy users). If data for treatment and control group subjects on the exogenous variable of interest are available, the evaluator may be able to estimate the treatment effect as a function of this variable.

Let m_{ij} be an indicator that subject i belongs to a group j, j = 1, 2, …, J, where membership in group j is exogenous to receiving the treatment. Then the average treatment effect per subject for subjects in group j can be estimated using the following regression equation:

$$y_{it} = \beta_0 + \Sigma^{J}_{j=1}\, \beta_{1j} * Tr_i * m_{ij} + \Sigma^{J-1}_{j=1}\, \gamma_j m_{ij} + \varepsilon_{it} \qquad (4)$$

where:

m_{ij} = An indicator for membership of subject i in group j; it equals 1 if customer i belongs to group j and equals 0, otherwise

β_{1j} = The average treatment effect for subjects in group j; energy savings per subject per period j equals $-\beta_{1j}$

γ_j = The average energy consumption per period for subjects in group j, j = 1, 2, …J-1.

All of the other variables are defined as in Eq. 2.

This specification includes a separate intercept for each group indicated by γ_j and the treatment indicator Tr_i interacted with each of the m_{ij} indicators. The coefficients on the interaction variables β_{1j} show average savings for group j relative to baseline average energy use for group j. It is important that the equation include the uninteracted indicator variables for the groups if average energy consumption varies between groups; otherwise, the treatment effect for group *j* will be incorrectly estimated relative to the average consumption of all control subjects rather than control subjects in group j.

4.4.5 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 indicator variables for the time periods as in the following equation[40]:

$$y_{it} = \Sigma^{T}_{j=1}\ \beta_j\ Tr_i^*\ d_{jt} + \Sigma^{T}_{j=1}\ \theta_j d_{jt} + \varepsilon_{it} \quad (5)$$

where:

β_t = The average savings per subject for period j (for example, the average savings per subject during month 4 or during hour 6)

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

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

Equation 4 can be estimated by including a separate dummy variable and an interaction between the dummy variable and Tr_i for each time period t, where t = 1, 2, ..., T. When the time period is in months, the time-period variables are referred to as month-by-year fixed effects. The coefficient on the interaction variable for period t, β_t, is the average savings per subject for period j. Again, because ε_{it} is uncorrelated with the treatment after accounting for the average energy consumption in period t, the OLS estimation of Eq. 4 (with standard errors clustered on subjects) results in an unbiased estimate of the average treatment effect for each period.

Evaluators with smart meter data can use this specification to estimate BB program demand savings during specific hours of the analysis period. The coefficient β_j would indicate the demand savings from the treatment during hour *j*. Examples of research that estimates savings during hours of peak usage include Stewart (2013a), Todd (2014), and Brandon et al. (2019).

4.4.6 Difference-in-Differences Regression Model of Energy Use

This section outlines a DiD approach to estimating savings from BB field experiments. This protocol recommends DiD estimation to the simple differences approach but DiD requires information about the energy use of treatment and control group subjects during the pretreatment and treatment periods. These energy use data enable 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.

[40] 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 the OLS on the transformed data.

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

$$y_{it} = \alpha_i + \beta_1 P_t + \beta_2 P_t * Tr_i + \varepsilon_{it} \quad (6)$$

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 resulting from the treatment per subject per unit of time.

The model includes fixed effects to account for differences in average energy consumption between subjects. Including subject fixed effects would likely explain a significant amount of the variation in energy consumption 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 periods when the treatment is in effect, and 0 for other periods and all control subjects.

Equation 5 is a DiD specification. For control group subject i, the expected energy use is α_i during the pretreatment period and $\alpha_i + \beta_1$ during the treatment period. The difference in expected energy use between pretreatment and treatment periods, 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 pretreatment 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 and savings from the BB program. Taking the difference yields β_2, a DiD estimate of program savings. The OLS estimation of Eq. 5 results in an unbiased estimate of β_2.

A more general form of Eq. 5 would allow the treatment period to vary for each subject and substitute time-period fixed effects (such as a separate indicator variable for each day or month of the analysis period) for the stand-alone postperiod variable. The specification with time-period fixed effects in Eq. 6 can be handy when subjects begin the treatment at different times, such as with rolling program enrollments or if it is difficult to define when treatment would have begun for a control group subject.

$$y_{it} = \alpha_i + \tau_t + \beta_2 P_{it} * Tr_i + \varepsilon_{it} \quad (7)$$

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 of T-1 time periods t, where t = -T, -T+1, …, -1, 0, 1, 2, ..., T; one time period dummy variable must be dropped to avoid collinearity

P_{it} = An indicator variable for whether time period t occurs during the treatment for subject i; it equals 1 if treatment group subject i received the treatment during period t, and equals 0 otherwise.

As in Eq. 4, the coefficient β_2 represents the average savings per treated customer per time period. The interpretations of the other variables and coefficients in the model remain unchanged.

4.4.7 Difference-in-Differences Estimate of Savings for Each Time Period

By re-specifying Eq. 5 with time-period fixed effects, savings can be estimated during each period and the identifying assumption tested to determine that assignment to the treatment was random. Consider the following DiD regression specification:

$$y_{it} = \alpha_i + \Sigma^{T}_{j=-T}\, \theta_j d_{jt} + \Sigma^{-1}_{j=-T}\, \beta_j\, Tr_i^* \, d_{jt} + \Sigma^{T}_{j=1}\, \beta_j\, Tr_i^* \, d_{jt} + \varepsilon_{it} \quad (8)$$

Savings in each period are estimated 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, …, -1, 0, 1, 2, ..., T. The coefficient on the interaction variable for period t, β_t^T, is the DiD savings for period t.

Unlike the simple differences regression model, this model yields an estimate of BB program savings during all periods except one, which must be excluded to avoid collinearity, for a total of 2T-1 period 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 using standard errors clustered on utility customers.

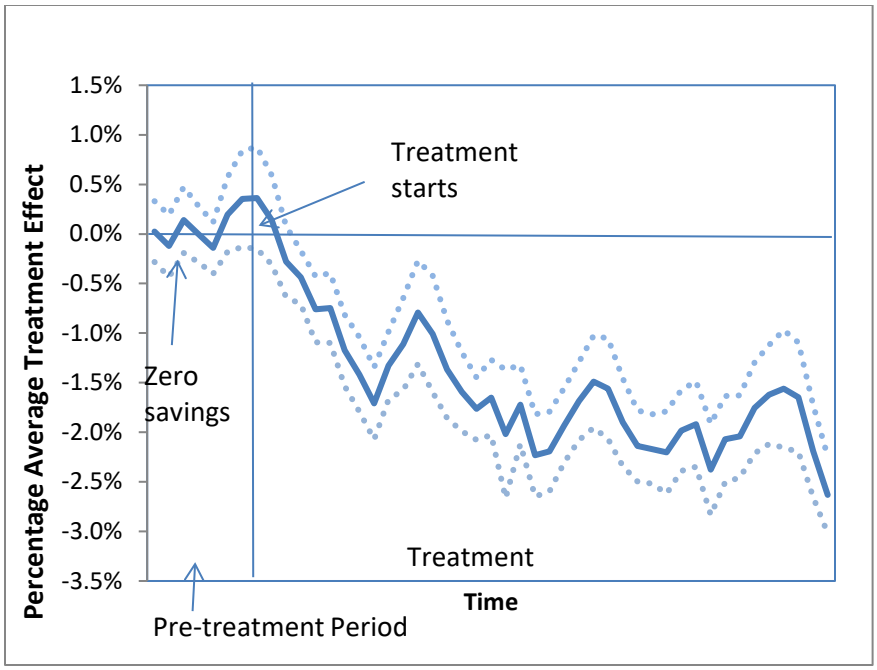

Figure 4. Example of DiD regression savings estimates

Estimates of pretreatment savings can be used to test the assumption of random assignment to the treatment. Before utilities administer the treatment, it should not be possible to reject the hypothesis of statistically significant differences in energy consumption between treatment and control group subjects, that is, the confidence intervals should contain the x axis. BB program pretreatment saving estimates that were statistically different from zero might suggest a flaw in the experiment design or implementation or the evaluator's understanding of the experiment.

As with Eq. 3, this specification can be used to estimate demand savings during specific hours. Energy consumption data for hours before the treatment are required, however.

4.4.8 Simple Differences Regression Model with Pretreatment Energy Consumption

In addition to estimating energy savings as a DiD, evaluators can estimate savings as a simple difference conditional on average pretreatment energy consumption. This estimator, often referred to as a post-only model with preperiod controls or lagged dependent variable, includes pretreatment energy consumption as an independent variable in the regression to account for differences between subjects in their post-treatment consumption, serving a purpose similar to that of customer fixed effects in the DiD model. However, many researchers favor the post-only estimator because it usually has smaller variance than the standard fixed effects DiD estimator when energy consumption is uncorrelated or weakly correlated over time.[41] However, evaluators

[41] Some researchers refer to this model as a "post-only" model; however, this name is misleading because the model uses pretreatment consumption as an explanatory variable. In a personal correspondence with the authors, Hunt Allcott, who introduced this method in evaluation of Home Energy Reports, points out that if seasonal effects are being estimated, this model "has slightly smaller standard errors and can be better at addressing naturally occurring randomization imbalances that may result in the baseline pretreatment energy usage differing between the control and treatment group."

can estimate both specifications and compare results. In large samples, the models should produce very similar estimates.

Consider the following post-only with preperiod controls regression specification:

$$y_{it} = \tau_t + \beta_1 * Tr_{it} + \rho \overline{y_i^{pre}} + \varepsilon_{it} \quad (9)$$

where:

τ_t = The time-period fixed effect (an unobservable that affects 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, …, -1, 0, 1, 2, ..., T

β_1 = Coefficient for the average treatment effect of the program; the energy savings per subject per period equals $-\beta_1$

Tr_{it} = An indicator variable for whether subject i received the treatment in period t; the variable equals 1 for subjects who receive the treatment in period t and equals 0 otherwise

ρ = Coefficient indicating the effect of average pretreatment consumption on consumption during the treatment period

$\overline{y_i^{pre}}$ = Average consumption during the corresponding pretreatment period for subject i; for example, if the dependent variable was a customer's average daily consumption in July during the treatment period, $\overline{y_i^{pre}}$ would equal the customer's average daily consumption for July in the pretreatment period

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

With random assignment of subjects to treatment and control groups, the OLS estimation of Eq. 8 is expected to produce an unbiased estimate of the average savings per subject per period.

Evaluators can estimate slightly different versions of this model:

- **Savings for each treatment period**. Evaluators can include a treatment indicator variable for each period instead of a treatment indicator variable for the entire treatment period. This specification will produce an estimate of average savings per subject for each treatment period.
- **Additional pretreatment consumption control variables**. Instead of one pretreatment consumption variable, evaluators can include multiple pretreatment consumption variables, such as pretreatment consumption for different seasons or months of a year, days of the week, or hours of the day.
- **Additional control variables**. Evaluators can add other variables such as weather to the model. The addition of such variables might help to improve the precision of the savings estimates.

4.4.9 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 estimating savings from 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, an additional step is required to estimate average savings for utility customers who accept the behavioral intervention. Treatment in a 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 consumption data for treatment and control group subjects available for the pretreatment and treatment periods. Equations 1 through 4 can be used to estimate the treatment effect, or the average energy consumption effect on those receiving encouragement. If control group customers can participate, the estimate only captures savings from *compliers*, because, as discussed previously, *never takers* never accept the intervention, and *always takers* accept the intervention with or without encouragement.

To recover an estimate of the LATE—the savings from subjects who accept the treatment because of the encouragement—evaluators can scale the estimate of β_2 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:

$$\text{LATE} = \beta_2 / (\pi_T - \pi_C) \quad (10)$$

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.

A related approach for obtaining an estimate of savings for the BB intervention in a RED study is instrumental variables, two-stage least squares (IV-2SLS). This approach uses the random assignment of subjects to the treatment as an instrumental variable for the decision by encouraged customers to participate in the program. The instrumental variable provides the exogenous variation necessary to identify the effect of endogenous participation on energy consumption. Participation is endogenous because the encouraged customers' decisions to participate is not random and depends on unobserved characteristics that may be correlated with energy consumption. For encouragement to be a valid instrument, it must be that encouragement affects only energy consumption through its impact on BB program participation.

In the first stage, the evaluator regresses a binary program participation decision variable on an indicator for whether the customer was randomly assigned to receive encouragement and other exogenous independent variables from the second-stage energy consumption equation. The evaluator then uses the regression to predict the likelihood of participation for each subject and time period. In the second stage, the evaluator estimates the energy consumption equation, substituting the first-stage predicted likelihood of participation for the variable indicating actual

program participation. The estimated coefficient on the predicted likelihood of participation is the LATE for the BB intervention.

For a detailed method of using an IV approach, see Cappers et al. (2013) and for a real-world example of the IV-2SLS approach applied to a home weatherization program implemented as a RED, see Fowlie et al. (2018).

4.4.10 Standard Errors

Panel data have multiple energy consumption observations for each subject; thus, the energy consumption data are very likely to exhibit within-subject correlations. Many factors affecting energy consumption persist over time, and the strength of within-subject correlations usually increases with the frequency of the data. When standard errors for panel regression model coefficients are calculated, the within-subject correlations must be accounted for. Failing to do so will lead to savings estimates with standard errors that are biased.

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 programs, including STATA, SAS, and R, have regression packages that output clustered standard errors.

Clustered standard errors account for the fact that in a panel with N subjects and T observations per subject there is less information about energy consumption than in a data set with N*T independent observations. Because clustered standard errors account for these within-subject energy-use correlations, they are typically larger than OLS standard errors. When there is within-subject correlation, OLS standard errors are biased downward and overstate the statistical significance of the estimated regression coefficients.[42]

4.4.11 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 customers in the analysis, as well as utility customers whose billing accounts close during the analysis period.

As a general rule, evaluators should include all subjects initially assigned to the treatment and control groups in the savings analysis.[43] Specifically, evaluators should keep opt-out subjects in the analysis sample. Opt-out subjects may have different energy consumption characteristics than subjects who remain in the program and dropping them from the analysis would result in nonequivalent treatment and control groups. To ensure the internal validity of the savings estimates, opt-out subjects should be kept in the analysis sample.

Sometimes treatment or control group subjects close their billing accounts after the program starts. Account closures are usually unrelated to the BB program or savings; most are a result of

[42] Bertrand et al. (2004) show when DiD studies ignore serially correlated errors, the probability of finding significant effects when there are none (Type I error) increases significantly.

[43] This protocol urges evaluators not to arbitrarily drop outlier energy consumption observations from the analysis unless energy consumption was measured incorrectly, the customer was not a residential customer, or the sample size is small enough that the outlier strongly influences the estimated savings. If an outlier is dropped from the analysis, the reasons for dropping the outlier and the effects of dropping it from the analysis on the savings estimates should be clearly documented. Evaluators should test the sensitivity of the results to dropping observations.

households changing residences. Subjects in the treatment group should experience account closures for the same reasons and at the same rates as subjects in the control group; evaluators can thus safely drop treatment and control group subjects whose accounts close from the analysis sample.

When dropping customers who close their accounts during the treatment from the regression estimation, evaluators should still count the savings from these subjects for periods during treatment when their accounts were active. To illustrate, when estimating savings for a 1-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:

$$\text{Savings} = \sum_{m=1}^{12} -\beta_m * \text{Days}_m * N_m \quad (11)$$

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 consumption in each month)

Days_m = The number of days in month m

N_m = The number of treatment group subjects with active accounts in month m.

4.5 Energy Efficiency Program Uplift and Double Counting of Savings

Many BB programs cause participants to increase their participation in other utility energy efficiency programs, a phenomenon often referred to as *efficiency program uplift*. For example, most home energy report programs encourage recipients to participate in other utility energy efficiency programs that provide cash rebates in exchange for adopting efficiency measures, such as efficient furnaces, air conditioners, wall insulation, windows, and light-emitting diodes (LEDs). The savings from this efficiency program participation caused by HERs are often referred to as joint savings or uplift savings.

Quantifying the effects of BB programs on efficiency program participation is important for two reasons:

- Uplift can be an important effect of BB programs and a potential additional source of energy savings.
- Savings from efficiency program uplift may be double counted. When a utility customer participates in an efficiency program because of a BB program intervention, the utility may count the program savings twice: once in estimating BB program savings and again in estimating the rebate program savings. To avoid double counting, evaluators must estimate savings from program uplift and subtract these savings from the behavior program savings or the uplifted program savings or from both programs.[44]

[44] This protocol does not take a position on which program gets credit for the uplift. When a BB intervention causes participation in an energy efficiency program, we know that the program participation would not have occurred

4.5.1 Estimating Uplift Energy Savings

For BB programs implemented as randomized experiments, estimating savings from uplift is conceptually straightforward. To illustrate, suppose that a utility markets an energy efficiency measure to treatment and control group subjects identically through a separate rebate program. Customers in the behavioral treatment group also receive messaging encouraging them to adopt the measure. Because customers were randomly assigned to the treatment and control groups, the groups are expected to be equivalent except for the treated customers having received the BB program encouragement. Therefore, in comparing BB program treatment and control group customers, evaluators can attribute any difference in the uptake of the measure between the groups to the behavioral treatment. To improve the accuracy of the uplift estimate, evaluators can estimate the impact as a DiD, by comparing the change in uptake of the measure between the pre and post-treatment periods for treatment and control group customers. The DiD estimate will account for any preexisting differences between treatment and control groups in the tendency to adopt the measure. If data are not available on the installation of the measure in the pretreatment period (for example, if it was not rebated at that time), evaluators should estimate uplift savings based only on post-treatment differences.

Figure 5 illustrates the logic for calculating behavior program savings from the efficiency program as a DiD. The figure shows energy savings from utility rebate program participation for treatment and control group customers during the pretreatment and treatment periods. Although customers had been randomly assigned to receive treatment, treatment group customers had a slightly higher tendency to participate and greater savings (=5) during the pretreatment period than the control group (=4). In this case, estimating program uplift by taking the simple difference in post-treatment savings between the treatment and control groups (8-4) would ignore the higher savings for the treatment group that would have occurred in the absence of the BB treatment and yield a slightly biased uplift savings estimate of 4. The true uplift savings equal 3, and an accurate estimate can be obtained as a DiD: (8-5) – (4-4).

without the intervention. However, the amount of uplift caused by the BB intervention may depend on the dollar incentives provided by the efficiency program. For example, the BB program may produce greater lift in participation for a program incentive of $200 than $100. To determine the relationship between uplift and the incentive amount, it would be necessary to randomize the incentive amount and to study participation as a function of incentives and who receives the BB intervention. It is possible to subtract the uplift savings from either the behavior program or the uplifted program. However, it is common practice for evaluators to attribute all joint or uplift savings to other energy efficiency programs by subtracting them from the BB program savings. This is a simple and convenient approach for avoiding double counting of savings.

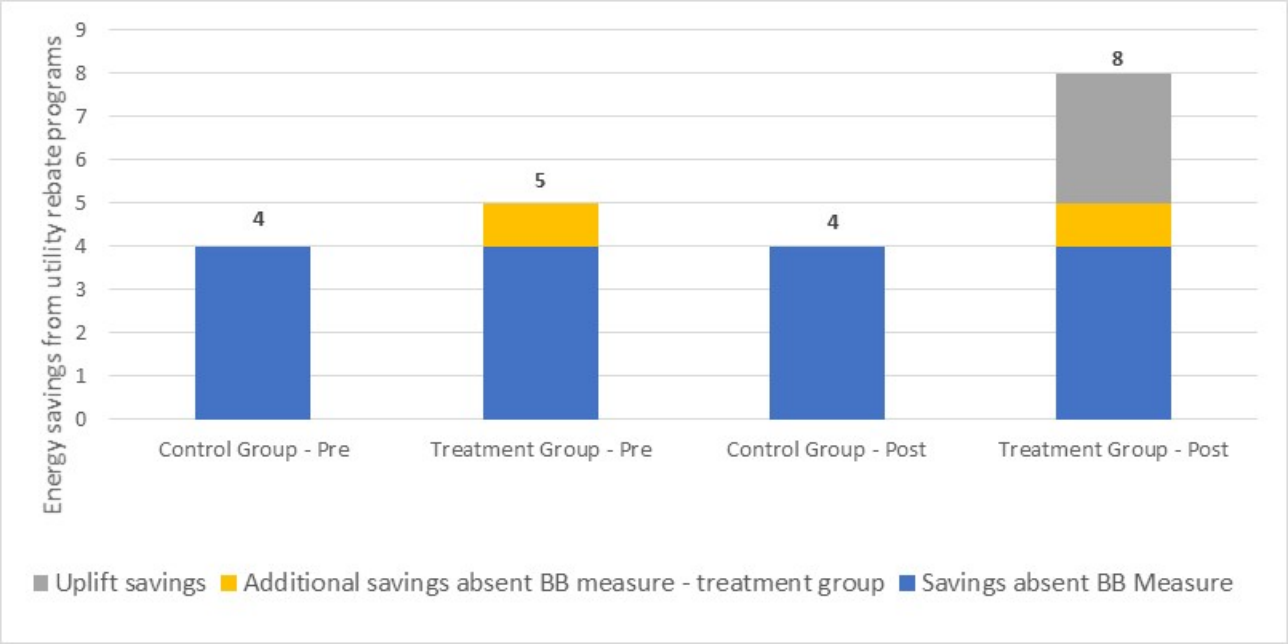

Figure 5. Calculation of double-counted savings

To estimate BB program savings from efficiency program uplift, evaluators should take the following steps:

1. Collect energy efficiency program tracking data for treatment and control group customers for the year before treatment and all years of treatment.[45] Match the BB program treatment and control group subjects to the utility energy efficiency program tracking data.
2. Calculate the average uplift savings per treatment group customer as the DiD between treatment and control groups in average efficiency program savings per customer, where the savings are obtained from the utility tracking database of installed measures.[46] The averages should be calculated over all treatment group customers and all control group customers, not just those who participated in efficiency programs. Evaluators can calculate the average uplift savings per treatment group customer as a difference in unconditional means between treatment group and control group customers or in a regression. As described in the next few paragraphs, it may be necessary to adjust the deemed savings values in the utility tracking data for measures installed for less than 1 year.
3. Multiply the uplift savings per treatment group customer by the number of customers who were in the treatment group to obtain the total uplift savings.

[45] These data should include a customer account number and premise number for linking the records to individual customers and homes, a measure description and category, the installation date, the quantity installed, and a unit annual savings value.

[46] A simple difference can be used if evaluators verify that pretreatment energy efficiency program participation and savings are equal for treatment and control group customers or if pretreatment energy efficiency program data are not available. Pretreatment data will be unavailable for new programs.

Evaluators can estimate BB program uplift savings for efficiency measures that the utility tracks at the customer level. Most measures for which utilities offer rebates—such as high-efficiency furnaces, windows, insulation, and air conditioners—fit this description. Also, evaluators can perform the uplift analysis for individual efficiency measures or programs or in aggregate across all programs and measures. Performing the analysis for individual measures or programs may provide useful insights about interactions between the BB program and other efficiency programs that an aggregate analysis cannot provide.

Evaluators should be mindful of specific reporting conventions for efficiency program measures in utility tracking databases. For example, many jurisdictions require utilities to report weather-normalized and annualized measure savings, which do not reflect when measures were installed during the year or the actual weather conditions that affect savings. In contrast, regression-based estimates of energy savings, such as from Eq. 4, will reflect installation dates of measures and actual weather. Evaluators should therefore adjust the annual deemed savings in the program reporting database to account for when measures were installed during the year and weather.[47]

In addition, for BB programs treating customers for longer than 1 year, evaluators should account for the savings from uplift in previous years if uplift savings are subtracted from the behavior program. Measures with a multiyear life installed in previous program years will continue to save energy for the remaining life of the measure. Depending on the utility's conventions for reporting savings, it may be necessary to account for savings from program lift in previous program years from the BB program savings estimate.[48]

4.5.2 Estimating Uplift for Upstream Programs

Upstream measures are those that the utility does not track at the customer level. The most important of such measures are high-efficiency lights such as 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 discounts to utility customers at the point of sale. Estimating behavior program savings for upstream measures is conceptually similar to that for downstream measures but requires a different data collection approach. Data on the purchases of rebated measures by treatment and control group subjects can be collected through customer surveys, store intercept surveys, or home site visits.[49]

Evaluators wanting to estimate the lift in LED adoption from upstream programs should be aware that it may be necessary to collect data for large numbers of customers to detect small BB program treatment effects. If evaluators perform surveys, they should size their survey samples with the objective of being able to detect small but economically significant effects. However, if

[47] For an example of a HER program evaluation that makes these adjustments, see Cadmus (2018) and DNV-GL (2018).

[48] Most utilities only claim the first-year savings for program measures, but each measure has an average useful life value associated with it, with the assumption that each measure continues to save energy for its useful life. To be consistent with how the utility tracks savings for rebated measures, if the evaluator subtracts all uplift savings from the BB program, meaning that the energy efficiency rebate program claims all uplift savings, the evaluator should subtract the uplift savings from the HER savings in subsequent years for the life of rebated measures; otherwise, the joint savings will be double counted. See Cadmus (2018) and DNV-GL (2018) for evaluations that account for savings from rebated measures in previous years. If the uplift savings are subtracted from the first-year savings of the uplifted energy efficiency program, the uplift savings should be subtracted once in the year the measures were installed.

[49] See PG&E (2013) for an example of a study employing home visits.

the treatment effect is small, the uplift savings from LEDs will also be small, and it may not be worth conducting surveys to measure it. Also, evaluators should adjust the lighting purchases impact estimates for in-service rates and the percentage of high-efficiency lamps sold in the utility service area that received rebates.[50] Evaluators should also be aware that some energy savings from purchasing LEDs may be offset by reductions in the hours of use of those bulbs by treated customers. LEDs may save less because treated customers light their homes less than before.

4.6 Savings Persistence and Measure Life

Most behavior-based program administrators and utility regulators assume a 1-year measure life for HERs and other residential BB measures. Administrators and regulators have been conservative in their assumptions about measure life for several reasons. First, doubts exist about the persistence of behavioral savings after treatment ends for utility customers who had been changing thermostat settings, turning lights off in unoccupied rooms, or modifying other energy consumption behaviors. Second, until recently, there has been a lack of evidence demonstrating that BB savings persist. Finally, HERs and other BB measures are fundamentally different than home improvements, such as LEDs or air source heat pumps. This difference is because BB measures attempt to influence behaviors, which often requires repeated treatments to be effective. Further, their effects can decay. For all these considerations, the default assumption for most BB program administrators has been that behavioral savings do not persist and that measure life is 1 year.

However, in the last 7 years, researchers have conducted highly credible RCT studies demonstrating that HER customers continue to save energy after treatment ends and that savings may persist for several years.[51] In addition, researchers have developed frameworks for estimating BB savings persistence and implementing a multiyear measure life (Khawaja and Stewart 2014; Jenkins et al. 2017). These frameworks account for repeated, multiyear program treatments and the gradual decay of behavior-based measure savings.

Because of this research, some BB program administrators and regulators have begun to reconsider the assumption of a 1-year measure life and allowed for savings persistence. For example, the Illinois Technical Resource Manual (TRM) was recently updated to require

[50] Upstream lighting savings captured in the BB program savings calculation equals the product of the BB treatment effect on upstream lighting purchases (in bulbs, estimated from the comparison of treatment group and control group purchases), the in-service rate, and the unit savings. The portion of these upstream lighting savings claimed by the upstream lighting program equals the product of the upstream lighting savings, the ratio of upstream sales to total market sales, and the upstream program net-to-gross ratio.

[51] For studies showing that energy savings may persist after customers stop receiving home energy reports, see Integral Analytics (2011), Allcott and Rodgers (2014), Navigant (2016, 2017), DNV GL (2016, 2018), Skumatz (2016), and NMR Group (2017). Allcott and Rodgers (2014) estimate a savings decay rate of about 19% per year. All persistence savings estimates are measured with uncertainty, and evaluators should keep this uncertainty in mind when conducting savings accounting with a multiyear measure life. Also, there is less agreement about the causes of savings persistence, that is, whether customers have formed long-lasting savings habits or made long-lasting home energy efficiency improvements. Based on analysis of 38 HER program RCT evaluations, Brandon et al. (2017) provide evidence that up to half of HER savings persistence may be attributable to physical capital improvements to homes. However, HER program administrators use a variety of messaging to encourage different savings activities, and the share of HER savings from physical home improvements may vary significantly across programs.

adjustments to HERs savings for persistence. Other states previously or recently adopted a multiyear measure life for HERs or other BB measures or are proposing to adopt one.[52]

This part of the protocol provides evaluators with guidance about HER savings accounting and designing experiments to estimate BB savings persistence and measure life and about estimating BB savings when savings from previous treatments persist. The protocol does not recommend specific savings decay or measure life values.

4.6.1 BB Savings Persistence and Measure Life Concepts

BB savings persistence and measure life concepts presented in this section are meant to be illustrative of how program administrators can perform BB savings accounting with a multiyear measure life. BB savings accounting methods are still evolving, and there is not yet consensus about the details. Sections 4.6.1.1 & 4.6.1.2 describe approaches that two states have implemented for performing HER savings accounting.

Figure 6 illustrates the measure life, savings persistence, and savings decay concepts for a multiyear BB program. The figure shows the average annual savings per customer for the first five program years. Suppose in the first year that the BB treatment generates 100 kWh of savings. Assume that savings from this treatment and all subsequent treatments partially persist, decaying at a 20% annual rate. In the second year, the BB treatment generates 150 kWh, but not all these savings are attributable to the second year 2 treatment. Eighty kWh of savings are from the year 1 treatment, and the remaining 70 kWh of savings are new savings attributable to the year 2 treatment.[53] In year 3 (and years 4 and 5), the same logic applies. Only a portion of the annual savings are attributable to that year's treatment. In year 3, 64 kWh of savings are from year 1 treatment, 56 kWh of savings are from the year 2 treatment, and 68 kWh are new savings, attributable to treatment in year 3.

[52] As of July 2019, Illinois, Connecticut, New Hampshire, and Minnesota have adopted a multiyear measure life for home energy reports. Pennsylvania is considering an HER multiyear measure life.

[53] The new savings can be further divided into avoided decay and incremental savings. Avoided decay is the difference between the previous year's savings and the savings that would have occurred if treatment had not resumed. Incremental savings are the difference in annual savings between the current and previous year.

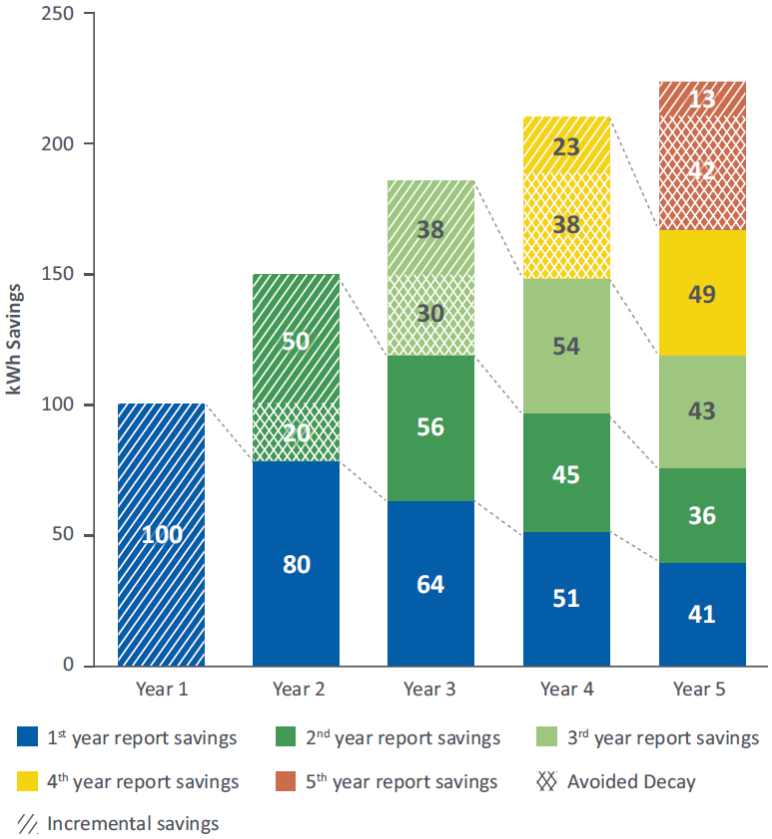

Figure 6. Illustration of savings persistence

Figure 6 shows: (1) for each year of BB treatment new savings are generated and that a fraction of the new savings persist in future years, and (2) for BB programs in which the same customers receive treatments in multiple years, some annual savings may be attributable to treatments provided in previous years. This implies that after year 1 only a portion of annual savings will be attributable to treatment in that year.

In this example, with a constant annual savings decay rate, the lifetime savings from the year *t* treatment is the sum of year *t* new savings ($s_{n,t}$) and savings in future years from persistence of year *t* savings:

$$Lifetime\ savings = s_{n,t} + s_{n,t}(1-\delta) + s_{n,t}(1-\delta)^2 + \cdots = \frac{s_{n,t}}{\delta} \quad (12)$$

where δ, $0 \leq \delta < 1$, is the savings decay rate. For instance, if δ=0.2, lifetime savings would equal $5s_{n,t}$.

Also, with measurements of the annual savings in year t, new savings from previous years of treatment, and the savings decay rate, it is possible to deduce new savings in the program's t^{th} year.

$$New\ savings\ in\ year\ t = s_t - \sum_{k=1}^{t}(1-\delta)^k s_{n,t-k} \quad (13)$$

where s_t is an estimate of the annual savings.

For example, the new savings in program year 3 equals:

$$s_{n,3} = s_3 - (1-\delta)^2 s_{n,1} \pm (1-\delta)\, s_{n,2}$$

In the example, to estimate the year 3 new savings, the evaluator would estimate the year 3 annual savings (188 kWh) by regression analysis and then subtract the decay-adjusted year 1 and year 2 new savings, which equal 64 kWh and 56 kWh, respectively. The year 3 new savings equal 68 kWh.

The previously mentioned formulas for lifetime and new savings neglect that most behavioral energy-efficiency programs experience attrition in the number of program participants because of customers moving residences and closing their accounts. A more accurate estimate of these savings would account for this attrition. If the annual rate of customer attrition equals α, $0 \leq \alpha < 1$:

$$Lifetime\ savings = s_{n,t} + s_{n,t}(1-\delta)(1-\alpha) + s_{n,t}(1-\delta)^2(1-\alpha)^2 + \cdots$$

$$= \frac{s_{n,t}}{\delta + \alpha - \delta\alpha}$$

and

$$s_{n,t} = s_t - \sum_{k=1}^{t}(1-\delta)^k(1-\alpha)^k s_{n,t-k}$$

For example, with customer attrition, new savings in program year 3 would equal:

$$s_{n,3} = s_3 - (1-\delta)^2(1-\alpha)^2 s_{n,1} - (1-\delta)(1-\alpha)\, s_{n,2}$$

Evaluators may also need an estimate of BB measure life. The measure life of a behavior-based treatment (e.g., reports sent in the second year of a program) can be defined as the lifetime savings expressed in terms of first-year savings equivalents[54]:

$$Measure\ life_{n,t} = \frac{lifetime\ savings_{n,t}}{s_{n,t}}$$

$$= \frac{1}{\delta + \alpha - \delta\alpha}$$

For example, for δ = 0.2 and α = 0.1, measure life for a behavior-based treatment in year *t* would equal 3.57 years. The lifetime savings from the year *t* treatment equal approximately 3.5 times the new savings in year *t*.

[54] Energy efficiency program administrators typically define measure life using the concept of effective useful life: "the median length of time (in years) that an energy efficiency measure is functional." (Hoffman et al. 2015) Because it is not possible to directly observe functionality of BB measures in contrast to a efficiency product, it is necessary to estimate BB measure life in terms of first-year savings.

This illustration of the savings persistence and measure life concepts has assumed that savings decay indefinitely at a constant annual rate and the customer attrition rate is constant, but these assumptions, while simplifying the savings accounting, may not hold and need not be used. For example, the savings persistence rate may change over time, savings may persist for a finite number of periods, or customer attrition rates may vary. Evaluators can relax the assumptions and adapt the previously mentioned framework to their own situations. However, even with alternative assumptions, the concepts of new savings, lifetime savings, and measure life described earlier are still valid, and with modifications, the formulas for these concepts can be applied.

The following section describes how Illinois and Pennsylvania have conducted HER savings accounting with a multiyear measure life.[55]

4.6.1.1 Illinois

The Illinois TRM incorporates the HER savings accounting framework with several modifications.[56] The TRM assumes that HER electricity and gas savings only persist for 5 years and that the electric savings decay at 20% in the first year after treatment and then at a higher rate for the second, third, fourth, and fifth years after treatment. After the fifth year, the savings completely decay. Gas savings decay at a faster rate.

Table 3 presents the Illinois TRM persistence factors for new savings as a function of years since the savings were first realized. The savings persistence factors equal one minus the cumulative savings decay rate.

Table 3. Illinois TRM HER Savings Persistence Factors

Fuel	Persistence factor for year *t* new savings[a] in year t	Persistence factor for year *t* new savings in year t+1	Persistence factor for year *t* new savings in year t+2	Persistence factor for year *t* new savings in year t+3	Persistence factor for year *t* new savings in year t+4
Electricity	100%	80%	54%	31%	15%
Natural Gas	100%	45%	20%	9%	4%

[a]New savings are the sum of avoided decay savings and incremental savings in year t.
Source: Illinois TRM (2019)

For example, with an annual customer attrition rate of α, new savings in program year 3 in Illinois would equal:

$$s_{n,3} = s_3 - 54\% * (1 - \alpha)^2 s_{n,1} - 80\% * (1 - \alpha)\, s_{n,2}$$

In this calculation, before accounting for attrition in the number of treated customers, 80% of year 2 new savings are assumed to persist to year 3 and 54% of year 1 new savings (all savings are new) are assumed to persist to year 3.

[55] Also, see NMR (2017) for application of this protocol's framework to a HER program in Connecticut and ADM (2018) for application to a Utah HER program.
[56] See Jenkins et al. (2017) for a description of the Illinois TRM framework development.

The Illinois TRM determined the HER savings persistence factors based on empirically estimated HER savings persistence factors for electricity and gas utilities inside and outside of Illinois. The TRM persistence factors will be updated as findings from new studies about HER savings persistence become available.

4.6.1.2 Pennsylvania

The Pennsylvania TRM also assumes a multiyear HER measure life and incorporates, with modifications, the previously described savings accounting framework. The TRM assumes that HER savings decay *continuously* at a linear rate of 31.3% per year for program populations treated for 2 or more years. The savings decay factor was based on analysis of HER savings decay for Pennsylvania electric utility HER programs that paused delivery of energy reports.[57] The savings persistence rate is assumed to be 0% for the first year of treatment.

Table 4. Pennsylvania TRM HER Electricity Savings Persistence Factors[b]

Program Year	Persistence factor for year *t* new savings[2] in year t	Persistence factor for year *t* new savings in year t+1	Persistence factor for year *t* new savings in year t+2	Persistence factor for year *t* new savings in year t+3	Persistence factor for year *t* new savings in year t+4
First year	100%	0%	0%	0%	0%
Second and later years	100%	84.4%	53.1%	21.8%	0%

[b]The savings persistence factors were calculated using the default annual savings decay assumption of 31.3% and the persistence formulas in the Pennsylvania TRM.
[c]New savings are the sum of avoided decay savings and incremental savings in year *t*.
Source: Pennsylvania Act 129 Phase IV TRM (2019)

The continuous linear decay rate means implies that following the second treatment year, 15.65% of second-year savings, all which are assumed to be "new," decay in the next year or, equivalently, that 84.4% (=1-0.313*(1-0.5)) persist.[58] Similarly, 53.1% (=1-0.313*(2-0.5)) of second-year savings persist after 2 years.

As an example, with an annual customer attrition rate of α, new savings in program year 4 would equal:

$$s_{n,4} = s_4 - 53.1\% * (1 - \alpha)^2 s_{n,2} - 84.4\% * (1 - \alpha) s_{n,3}$$

Because it is assumed none of the annual savings from the first program year persist, the first-year savings do not enter the calculation of savings for program year 4.

4.6.2 Estimating BB Savings Persistence

This section describes how evaluators can design studies to obtain estimates of savings persistence and savings decay for BB measures.

[57] See Pennsylvania Statewide Evaluation Team (2018).
[58] If 31.3% of HER savings decay after 1 year, the average rate of savings decay over the year is 31.3%*0.5.

4.6.2.1 Study Design

This protocol recommends that evaluators employ RCTs to estimate the persistence of BB savings after participants stop receiving treatment. The implementation of an RCT to estimate savings persistence should proceed similarly to the implementation of RCTs previously discussed in this protocol.

Figure 7 illustrates an RCT savings persistence experiment. The program administrator is assumed, as in Figure 1, to have implemented the BB program as an RCT with an opt-out design: customers from the study population were randomly assigned to receive the treatment or to a control group, and treated customers can opt out of the program. To economize on space, Figure 7 does not show the utility's option at the beginning of the program to screen customers or that after treatment begins customers can opt out of the program.

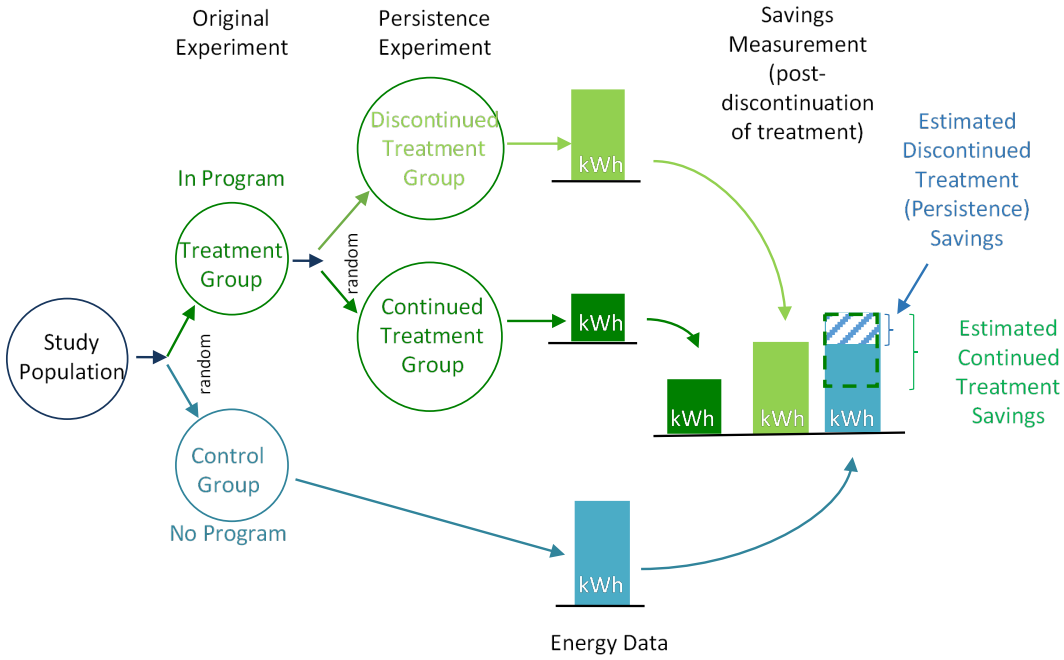

Figure 7. Illustration of savings persistence study design

The persistence study starts after treatment group customers have received treatment for some duration (e.g., 1, 2, or 3 years). Though not illustrated, the utility may choose to screen the treatment group (and the control group) and study persistence for a specific subpopulation (e.g., by an energy use, socio-demographic, or housing characteristic). Also, the persistence study population must include treatment group customers who 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 if they had been treated.

The next step is to randomly assign customers in the persistence study population to one of two groups. Customers in the "discontinued treatment" group will stop receiving the treatment;

customers in the "continued treatment" group will continue receiving it. Evaluators should size, that is, assign enough customers to, the continued and discontinued customer treatment groups to detect the expected savings. The utility then administers the experiment and after enough time has passed collects energy consumption data for the report discontinuation period for control customers, discontinued treatment group customers, and continued treatment group customers to estimate savings persistence.

To estimate savings for discontinued customers ("persistence savings"), the evaluator should compare the energy consumption of customers in the discontinued treatment group with the energy consumption of customers in the original control group during the discontinuation period. Under *Savings Measurement* in Figure 7, this difference is shown as the dashed, light-green box and represents the post-treatment savings for customers who no longer received the treatment.

The savings persistence rate can be estimated in two ways. This protocol recommends that evaluators compare the savings of the continued and discontinued treatment groups after treatment is discontinued. The continued treatment group savings represent the savings that the discontinued treatment group would have achieved if treatment had continued. Therefore, the ratio of the savings shows the percentage of the continued treatment group savings that persist after customers stop receiving treatment.

Evaluators can also estimate savings persistence by comparing the savings of the discontinued treatment group after treatment was discontinued with the group's savings before treatment was suspended. For evaluators wanting to measure savings persistence after a program administrator stops treating all customers in the behavior program, this approach is the only option. A limitation of this approach is, however, that savings may depend on weather, program implementation changes, or other time-varying factors, which, if not accounted for when comparing savings over time, can bias estimates of the savings persistence.

Both ways of calculating savings persistence only measure savings persistence rates for customers whose treatment was discontinued after a certain length of treatment (e.g., 2 years). Evaluators would need to conduct a series of discontinuation experiment to measure savings persistence for customers receiving treatment for fewer or greater number of years.

4.6.2.2 Estimating Savings Persistence

Suppose a utility started the treatment in period t = 1 and administered it for t*>0 periods. Beginning in period t = t*+1, the utility stopped administering the intervention for a random sample of treated customers. Evaluators can estimate the average savings per customer for a customer who continues to receive the treatment (continuing treatment group) and for those who stopped receiving the treatment after period t* (discontinued treatment group).

Assuming pretreatment energy consumption data are available, the following fixed effects DiD regression equation can be used to estimate savings during treatment and savings after treatment stops. This specification is estimated with consumption data for treatment and control group customers:[59]

[59] Evaluators can also implement a variant of the lagged dependent variable model (Eq. 7) to estimate savings persistence.

$$kWh_{it} = \alpha_i + \tau_t + \beta_1 P_{1,t}*Tc_i + \beta_2 P_{1,t}*Td_i + \beta_3 P_{2,t}*Tc_i + \beta_4 P_{2,t}*Td_i + \varepsilon_{it} \quad (14)$$

where:

kWh_{it} = electricity consumption by customer i in period t

α_i = A customer fixed effect (an unobservable that affects energy use for customer i); these effects can be estimated by including a separate intercept for each customer

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, …, -1, 0, 1, 2, ..., T

β_1 = The average energy savings per continued customer caused by the treatment during periods t = 1 to t = t*

$P_{1,t}$ = An indicator variable for periods when customers in the continued *and* discontinued treatment groups received the treatment; it equals 1 if period t occurs between periods t = 1 and t = t* and equals 0 otherwise

Tc_i = An indicator for whether customer i is in the continued treatment group; the variable equals 1 for customers in the continued treatment group and equals 0 for customers not in the continued treatment group

β_2 = The average energy savings per discontinued customer caused by the treatment during periods t = 1 to t = t*

Td_i = An indicator for whether customer i is in the discontinued treatment group; the variable equals 1 for customers in the discontinued treatment group and equals 0 for customers not in the discontinued treatment group

β_3 = The average energy savings from the treatment for customers in the continued treatment group when t>t*

$P_{2,t}$ = An indicator variable for periods when continued treatment group customers received the treatment and discontinued treatment group customers did not receive the treatment; it equals 1 if period t occurs after t = t* and equals 0 otherwise

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

If the persistence study is implemented as an RCT, OLS estimation of Eq. 9 is expected to yield unbiased estimates of savings for customers in the continued treatment group (β_3) and discontinued treatment group (β_4) after the discontinued group stops receiving treatment.[60] To

[60] Evaluators can test the identifying assumption that assignment of treatment group customers to the discontinued treatment group was random by comparing the consumption of continuing and discontinuing treatment group subjects prior to the first treatment. If assignment was done at random, there should not be statistically significant differences in consumption between the two groups during this period.

estimate savings persistence after treatment stops, evaluators can take the difference between savings during treatment (β_2) and post-treatment savings (β_4) for subjects in the discontinued treatment group or the difference between post-treatment savings for the discontinued treatment group (β_4) and the same period savings for the continued treatment group (β_3).

4.6.3 Practical Evaluation Considerations

Evaluators conducting experiments to measure BB savings persistence should be mindful of several issues. First, stopping delivery of HERs or other BB treatments to estimate savings persistence may involve loss of some energy savings from discontinued customers, especially if the measure life is 1 year or program administrators are prevented from claiming persistence savings from discontinued customers. Also, the suspension of treatment may not result in a commensurate reduction in program administration and implementation costs, so that the program's cost-effectiveness may be adversely affected. It is also possible that suspending reports or treatment may dissatisfy some utility customers grown accustomed to receiving treatment.

Program administrators not wanting to conduct their own experiments can use findings about savings persistence from other studies but should borrow from studies that are valid for their own programs. As the rate of BB savings persistence may depend on climate; presence of other efficiency programs; BB program implementation strategies, including the frequency of prior treatment (e.g., quarterly vs. monthly); duration of prior treatment (number of years of treatment); and the form of the treatment (e.g., electronic or paper HERs); savings persistence estimates for one group of utility customers may not apply to other groups.

Finally, Eq. 10 estimates savings for continued and discontinued treatment group customers for the actual weather during the analysis period, but evaluators may want to normalize the persistence savings estimates for year-to-year variation in weather. To obtain weather-normalized savings, evaluators can estimate savings as a function of cooling and heating degrees by adding stand-alone heating and cooling degree variables and three-way interaction variables of degrees with each of $P_{1,t}*Tc_i$, $P_{1,t}*Td_i$, $P_{2,t}*Tc_i$, and $P_{2,t}*Td_i$ to the right side of Eq. 10. All independent variables in Eq. 10 would also remain in this enhanced specification. For example, in a savings model estimated with monthly billing data, the coefficients on the interaction terms would indicate how savings before and after discontinuation of treatment depended on HDDs and CDDs. The coefficients on the two-way interaction variables $P_{1,t}*Tc_i$, $P_{1,t}*Td_i$, $P_{2,t}*Tc_i$, and $P_{2,t}*Td_i$ would indicate the average savings unrelated to weather. Estimating this specification requires within-time period (e.g., a day or month) variation between customers in heating and cooling degrees. Without such variation, it is not possible to isolate the consumption's impact of weather from the impacts of other time-specific factors, which the time-period fixed effects account for.

5 Reporting

BB program evaluators should carefully document the research design; data collection and processing steps; and 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. This should also include a careful description of the impacts measured by the experiment.
- The sample design and sampling process
- The 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
- Results of the savings estimation, including point estimates of savings and standard errors and full results of regressions used to estimate savings. Evaluators should clearly state the time periods to which the savings estimates pertain.
- Assumptions about measure life and savings persistence. If a behavior-based measure has a multiyear measure life, evaluators should describe the calculation of persistence savings and new savings.

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

6 Looking Forward

Evaluators and program administrators should employ randomized experiments for evaluating BB programs whenever possible. However, some BB programs may be difficult or costly to evaluate using these types of experiments. In these cases, evaluators have explored quasi-experiments that rely on random but uncontrolled variation in who participates.

An important question concerns the accuracy of quasi-experimental methods, such as propensity-score matching, regression discontinuity, and DiD estimation for evaluating BB programs. Evaluators of BB programs have employed and will continue to employ these methods. Although this protocol has cited several studies comparing the accuracy of randomized experiments and quasi-experiments, more research will be needed to draw firm conclusions about the accuracy of quasi-experiments.

Depending on the outcome of this research and acceptance by regulators and program administrators of savings estimates from quasi-experiments, evaluators should consider updating this protocol to include quasi-experimental methods.

7 References

ADM Associates, Inc. 2018. *Evaluation of 2016-2017 Home Energy Reports Program*. Report Submitted to Pacific Power. https://www.pacificorp.com/content/dam/pcorp/documents/en/pacificorp/environment/dsm/washington/Home_Energy_Reports_2016-2017.pdf.

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

Allcott, H., and T. Rodgers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10), pp. 3003-3037.

Allcott, H. 2015. "Site Selection Bias in Program Evaluation." *The Quarterly Journal of Economics* 130 (3), pp. 1117-1165.

Baylis, P., P. Cappers, L. Jin, A. Spurlock, A. Todd. 2016. "Go for the Silver? Evidence from field studies quantifying the difference in evaluation results between "gold standard" randomized controlled trial methods versus quasi-experimental methods." ACEEE Summer Study on Energy Efficiency in Buildings 2016.

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

Brandon, A., P. J. Ferraro, J. A. List, R. D. Metcalfe, M. K. Price, F. Rundhammer. 2017. "Do the Effects of Social Nudges Persist? Theory and Evidence from 38 Natural Field Experiments." National Bureau of Economic Research working paper 23277. http://www.nber.org/papers/w23277.

Brandon, A., J. A. List, R. D. Metcalfe, M. K. Price, and F. Rundhammer. 2019. "Testing for Crowd Out in Social Nudges: Evidence from a Natural Field Experiment in the Market for Electricity." *Proceedings of the National Academy of Sciences* 116 (12), 5293-5298.

Bruhn, M., and D. McKenzie. 2009. "In Pursuit of Balance: Randomization in Practice in Development Field Experiments." *American Economic Journal: Applied Economics* 1(4): 200-232.

Burlig, F., L. Preonas, M. Woerman. 2017. "Panel Data and Experimental Design." *Energy Institute at Haas working paper.* https://ei.haas.berkeley.edu/research/papers/WP277.pdf.

Cadmus. 2018. *Residential Customer Behavioral Savings Pilot Evaluation*. Prepared by Cadmus for the Vermont Public Service Department. https://publicservice.vermont.gov/sites/dps/files/documents/VT%20PSD%20RCBS%20Y3%20Evaluation%20Report%20FINAL.pdf.

Cappers, P., A. Todd, M. Perry, B. Neenan, R. Boisvert, R. 2013. *Quantifying the Impacts of Time-based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies: Protocols and Guidelines.* Lawrence Berkeley National Laboratory, Berkeley, CA and the

Electric Power Research Institute, Palo Alto, CA. LBNL-6301E. https://emp.lbl.gov/sites/default/files/lbnl-6301e.pdf.

Consortium for Energy Efficiency Database. 2018. https://library.cee1.org/content/2018-behavior-program-summary-public-version.

Costa, D. L., M. E. Kahn. 2010. *Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment.* NBER Working Paper 15939. http://www.nber.org/papers/w15939.

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

DNV GL. 2016. *Puget Sound Energy: 2015 Home Energy Reports Impact Evaluation*. http://www.oracle.com/us/industries/utilities/home-energy-reports-err-2015-3697558.pdf

DNV GL. 2018. *Puget Sound Energy: 2017 Home Energy Reports Impact Evaluation-Final Report*. https://conduitnw.org/Pages/File.aspx?rid=4415

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

Fowlie, M. 2010. *U.S. Department of Energy Smart Grid Investment Grant Technical Advisory Group Guidance Document #7, Topic: Design and Implementation of Program Evaluations that Utilize Randomized Experimental Approaches.* November 8, 2010.

Fowlie M., Michael Greenstone, and Catherine Wolfram. 2018. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *The Quarterly Journal of Economics* 133 (3), 1597-1644.

Frison, L., and S. J. Pocock. 1992. "Repeated measures in clinical trials: Analysis using mean summary statistics and its implications for design." *Statistics in Medicine* 11.13 (1992): pp. 1685-704.

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

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

Hoffman, Ian, Steven R. Schiller, et al. 2015. *Energy Savings Lifetimes and Persistence: Practices, Issues, and Data*. Lawrence Berkeley National Laboratory Technical Brief. https://emp.lbl.gov/sites/default/files/savings-lifetime-persistence-brief.pdf

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

Illinois Statewide Technical Reference Manual. 2018. *2019 Illinois Statewide Technical Reference Manual for Energy Efficiency Version 7.0, Volume 4: Cross-Cutting Measures and*

Attachments. https://www.icc.illinois.gov/downloads/public/IL-TRM_Effective_010119_v7.0_Vol_4_X-Cutting_Measures_and_Attach_092818_Final.pdf

Integral Analytics. 2011. *Sacramento Municipal Utility District Home Energy Report Program Impact and Persistence Evaluation Report, Years 2008-2011.* Prepared by May Wu, Integral Analytics.

Jenkins, Cheryl et al. 2017. *Accounting for Behavioral Persistence-A Protocol and a Call for Discussion*. Presented at the 2017 International Energy Program Evaluation Conference. https://www.iepec.org/?attachment_id=13250.

Khawaja, M. S. and J. Stewart. 2014. *Long-Run Savings and Cost Effectiveness of Home Energy Report Programs.* Cadmus Research Report. http://www.cadmusgroup.com/wp-content/uploads/2014/11/Cadmus_Home_Energy_Reports_Winter2014.pdf.

List, J. A. 2011. "Why Economists Should Conduct Field Experiments and 14 Tips for Pulling One Off." *Journal of Economic Perspectives* (25), pp. 3–16.

List, J. A., S. Sadoff, S., and M. 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.

Mazur-Strommen, S., and K. Farley. 2013. *ACEEE Field Guide to Utility-Run Behavior Programs.* American Council for an Energy Efficient Economy, Report Number B132.

Minnesota Department of Commerce, Division of Energy Resources. 2015. *Energy Efficiency Behavioral Programs: Literature Review, Benchmarking Analysis, and Evaluation Guidelines.* Prepared by Illume Advising with subcontractors Indica Consulting and Dr. Edward Vine. http://mn.gov/commerce-stat/pdfs/card-report-energy-efficiency-behavorial-prog.pdf.

National Renewable Energy Laboratory. 2017. *Strategic Energy Management Protocol: The Uniform Methods Project.* Prepared by James Stewart. https://www.nrel.gov/docs/fy17osti/68316.pdf.

Navigant Consulting. 2016. *ComEd Home Energy Report Program Decay Rate and Persistence Study – Year Two.* Prepared for Commonwealth Edison Company by C. Olig and E. Young.

Navigant Consulting. 2017. *ComEd Home Energy Report Program Decay Rate and Persistence Study – Year Three*. Prepared for Commonwealth Edison Company by W. Sierzchula and D. Dinsmoor.

NMR Group. 2017. *1606 Eversource Behavior Persistence Evaluation*. Submitted to Energy Efficiency Board Evaluation Administrator. https://www.energizect.com/sites/default/files/R1606_Eversource%20Behavior%20Persistence%20Evaluation_FINAL_10.15.17.pdf.

Pennsylvania Act 129 Phase IV TRM. 2019. http://www.puc.pa.gov/filing_resources/issues_laws_regulations/act_129_information/technical_reference_manual.aspx

Pennsylvania Statewide Evaluation Team. 2018. *Residential Behavioral Program Persistence Study*. Prepared by A. Ciccone and J. Smith. http://www.puc.state.pa.us/Electric/pdf/Act129/SWE_Res_Behavioral_Program-Persistence_Study_Addendum2018.pdf.

Pacific Gas & Electric. 2013. *Evaluation of Pacific Gas and Electric Company's Home Energy Report Initiative for the 2010-2012 Program.* Prepared by Freeman, Sullivan & Co. http://www.oracle.com/us/industries/utilities/evaluation-pacific-gas-company-3631944.pdf.

Rosenberg, M., G. Kennedy Agnew, K. Gaffney. 2013. "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.

SEE Action. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations.* State and Local Energy Efficiency Action Network. Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman, Lawrence Berkeley National Laboratory. https://www4.eere.energy.gov/seeaction/system/files/documents/emv_behaviorbased_eeprograms.pdf.

Seelig, M. J. 2013. *Business Energy Reports*. Presentation at BECC 2013. http://beccconference.org/wp-content/uploads/2013/12/BECC_PGE_BER_11-19-13_seelig-.pdf.

Shadish, W. R., T. D. Cook, D. T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference.* Belmont, CA: Wadsworth Cengage Learning.

Skumatz, L. 2016. "Persistence of Behavioral Programs: New Information and Implications for Program Optimization." *The Electricity Journal* 29 (5): pp. 27-32.

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

Stewart, J. 2013a. *Peak-Coincident Demand Savings from Behavior-Based Programs: Evidence from PPL Electric's Behavior and Education Program.* UC Berkeley: Behavior, Energy and Climate Change Conference. http://escholarship.org/uc/item/3cc9b30t.

Stewart, J. 2013b. *Energy Savings from Business Energy Feedback*. Presentation at BECC 2013. http://beccconference.org/wp-content/uploads/2015/10/presentation_stewart.pdf.

Todd, A. 2014. *Insights from Smart Meters: The Potential for Peak-Hour Savings from Behavior-Based Programs.* Lawrence Berkeley National Laboratory. LBNL Paper LBNL-6598E. http://escholarship.org/uc/item/2nv5q42n.