Smart Thermostat Evaluation Protocol

Dec 2016 – May 2023

James I. Stewart,[1] Carly Olig,[2] Sepideh Shahinfard,[3] Ken Agnew,[4] Stefanie Wayland,[5] Zachary Horvath,[1] and Jason Lai[2]

1 Cadmus Group
2 Guidehouse
3 Quantum Energy Analytics
4 DNV GL
5 California Energy Commission

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-5R00-86175
June 2023

Smart Thermostat Evaluation Protocol

Dec 2016 – May 2023

James I. Stewart,[1] Carly Olig,[2] Sepideh Shahinfard,[3] Ken Agnew,[4] Stefanie Wayland,[5] Zachary Horvath,[1] Jason Lai[2]

1 Cadmus Group
2 Guidehouse
3 Quantum Energy Analytics
4 DNV GL
5 California Energy Commission

NREL Technical Monitor: Chuck Kurnik

Suggested Citation
Stewart, James I., Carly Olig, Sepideh Shahinfard, Ken Agnew, Stefanie Wayland, Zachary Horvath, Jason Lai. 2023. *Smart Thermostat Evaluation Protocol: Dec 2016 – May 2023.* Golden, CO: National Renewable Energy Laboratory. NREL/SR-5R00-86175. https://www.nrel.gov/docs/fy23osti/86175.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-5R00-86175
June 2023

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 U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Building Technologies Office. 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.

Preface

This document was developed for the U.S. Department of Energy Uniform Methods Project (UMP). The UMP provides model protocols for determining energy savings and demand reductions 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 UMP, visit the website, 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 chapter authors wish to thank and acknowledge the following individuals for their thoughtful comments and suggestions on drafts of this protocol.

- Technical Committee members:
  - Michael Blasnik, Google Nest
  - Abigail Daken, U.S. Environmental Protection Agency
  - Tamara Dzubay, ecobee
  - Miriam Goldberg, DNV-GL
  - Ethan Goldman, Resilient Edge
  - Pace Goodman, Illume Advising
  - Karen Herter, Herter Energy
  - Arnie Meyer, Resideo
  - Bill Provencher, University of Wisconsin–Madison
  - Kahryn Riley, ecobee
- Hossein Haeri, retired
- Dale Hoffmeyer, U.S. Department of Energy
- M. Sami Khawaja, Cadmus
- John Koliner, Apex Analytics
- Michael Li, U.S. Department of Energy
- Joshua D. Rushton, Northwest Power and Conservation Council, Regional Technical Forum
- Michael Siemann, Resideo
- Richard Spellman, GDS Associates

List of Acronyms

AMI	advanced metering infrastructure
API	application programming interface
CDD	cooling degree days
D-in-D	difference-in-difference
DSM	demand-side management
EPA	U.S. Environmental Protection Agency
HDD	heating degree days
HVAC	heating, ventilating, and air conditioning
LDV	lagged dependent variable
RCT	randomized controlled trial
RECS	Residential Energy Consumption Survey
RED	randomized encouragement design
UMP	Uniform Methods Project

Table of Contents

List of Figures

List of Tables

1 Measure Description

A smart thermostat is an internet-connected device that controls home heating, ventilation, and air-conditioning (HVAC) equipment and can automatically adjust temperature set points to optimize performance and achieve energy savings.[1] Smart thermostat features often include two-way communication, occupancy detection (such as geofencing and occupancy sensors), schedule learning, and seasonal optimization algorithms. Smart thermostats can control most conventional HVAC systems including central air conditioners, heat pumps, and forced air furnaces.

Several types of residential utility programs offer smart thermostats as replacements measures. These programs may be delivered as upstream or midstream rebate programs with retail partners, direct-install programs with installations performed by HVAC contractors, or self-install programs that provide utility customers with a thermostat for installation as a stand-alone measure or as part of an efficiency kit. In addition, working with smart thermostat vendors, utilities can offer separate optimization programs to produce energy savings beyond those achieved by installing a smart thermostat.

From an evaluation perspective, smart thermostat programs have several noteworthy features. First, the energy savings from a smart thermostat may change over the life of the device. As a smart thermostat is connected to the internet, original equipment manufacturers can update the thermostat software to improve energy efficiency as implemented by the thermostat. Likewise, users can adjust the thermostat settings and schedules over time in response to changes in weather, thermal comfort, energy prices, or preferences for energy efficiency. Additionally, many thermostat manufacturers offer seasonal optimization programs that recommend changes or make minor, automated adjustments to the thermostat settings to improve energy efficiency. These opt-in programs are now standard offerings for many smart thermostat manufacturers and provided at no additional cost to users. The potential for software updates and continuous optimization and the evolving nature of user interactions mean future energy savings may differ from first-year savings and the energy savings of smart thermostats may need to be evaluated more than once.[2]

Second, smart thermostats often have small unit energy savings relative to a home's total energy consumption, especially in comparison to whole-home retrofit programs. This can make it difficult to detect the smart thermostat savings in billing or advanced metering infrastructure (AMI) meter consumption data. For example, as cooling loads in many regions average about 20% of annual electricity consumption, smart thermostat savings of 10% of cooling energy use would equate to a 2% reduction in home electricity consumption. Evaluators should use regression analysis of whole-home billing consumption or AMI meter consumption data to evaluate smart thermostat savings because, as explained at greater length in Section 2 and

[1] ENERGY STAR® lists requirements for a thermostat to qualify as a connected thermostat. See https://www.energystar.gov/sites/default/files/asset/document/ENERGY%20STAR%20Program%20Requirements%20for%20Connected%20Thermostats%20Version%201.0.pdf

[2] Life cycle savings for traditional energy efficiency measures are often reported and evaluated based on first-year savings estimates and the effective useful life of the measure.

Section 3, these data are usually available to evaluators and regression can control for the impacts of weather and other potentially confounding factors on a home's energy consumption.

Finally, as with other energy efficiency programs, participation in smart thermostat programs is self-selective. As discussed at greater length in Section 3, smart thermostat participants tend to be, among other things, younger, higher-income, and more likely to adopt electric vehicles and internet connected devices than nonparticipants. These differences are often unobservable to the evaluator and correlated with a home's energy consumption, creating the potential for bias in estimating savings. Due to the small unit savings of thermostats, errors and biases from self-selection that may not be very consequential when evaluating whole-home retrofits (e.g., ±2% of home electricity consumption) can have a major impact when evaluating the savings and cost-effectiveness of smart thermostat programs.[3] A percentage point change in the estimated savings could affect the cost-effectiveness of a program.[4] This means it is important for evaluators to assess and to minimize the potential for error from selection bias in estimating smart thermostat program savings.

[3] Errors in smart thermostat savings estimates can arise randomly from sampling or random disturbances in the energy consumption data or from bias introduced in participant self-selection, sampling, the model specification (such as from an omitted variable), or the estimation procedure. Models of home energy consumption for estimating smart thermostat savings are presented in Section 3.1.4.

[4] For example, suppose surveys indicate smart thermostat program participants had higher rates of adopting electric vehicles and adding living space to their homes after installing a smart thermostat than nonparticipants. The surveys also suggest installing a smart thermostat did not cause these changes. As a result of the changes, home electricity consumption across all participants was about one percentage point higher than it otherwise would have been after installing the smart thermostat. Then, the smart thermostat program electricity savings estimate from the billing analysis is likely biased downward by as much as one percentage point. If the threshold for program cost-effectiveness is 1.5% of home electricity consumption and the evaluation based on a matched comparison group finds savings of 1%, the cost-effectiveness threshold is within one percentage point (the amount of potential bias) of the evaluated savings, and the evaluator should not conclude the program is not cost-effective.

2 Application Conditions of Protocol

This protocol applies to evaluating smart thermostat replacement and optimization programs when several conditions are met:

- The smart thermostats were installed in residential buildings.
- The smart thermostats were installed as part of a utility residential thermostat replacement program, or existing smart thermostats were enrolled in an optimization program.
- The replacement smart thermostats are stand-alone measures—they were not bundled with other energy efficiency measures or installed as part of time-varying pricing programs (such as time-of-use rates).
- The objective is to estimate savings of utility-supplied energy (natural gas or electricity). The savings may be measured for a year, a season, specific days, or specific hours of the day.
- For smart thermostat programs, customer billing consumption or AMI meter data are available for the reporting (post-installation) and baseline (pre-installation) periods[5]; for thermostat optimization programs, billing consumption or AMI meter data are available for the baseline (pre- or non-optimization) and reporting (optimization) periods or thermostat telemetry data are available for the baseline and reporting periods.[6]
- The number of program smart thermostats is large enough that there is a high probability of detecting the expected savings through regression analysis of billing consumption or AMI meter data given the unexplained variance in energy consumption in the data.[7]
- The impact of known or likely sources of bias should be small compared to the expected savings. Depending on the study design and methodology, this bias may equal 1% or more of home energy consumption. If a known or likely bias of this magnitude would materially affect the conclusions of the study, the evaluator should redesign the evaluation approach to mitigate the bias (if feasible) and the evaluation report should explain the potential implications of the bias for the study's conclusions and future program and policy design.

While this protocol is applicable to the evaluation of most utility energy efficiency thermostat replacement or optimization programs, it does not address the following objectives or situations:

- The goal is to estimate the demand impacts from smart thermostat demand response programs. Demand response is event-based demand-side management (DSM), meaning it happens in response to specific utility operations contingencies or needs such as high wholesale electricity prices, reliability concerns, or peak load management. While there are many conceptual similarities between estimating demand response savings and estimating hourly energy efficiency savings, there are also important differences that

[5] Billing consumption or AMI meter data are considered accurate because the data are used for billing purposes.
[6] Smart thermostat optimization programs may be evaluated with telemetry data because all treatment and control or comparison group customers will have smart thermostats.
[7] See the Uniform Methods Project (UMP), Chapter 8 for more information (Agnew and Goldberg 2017).

place demand response savings outside the scope of this chapter. Evaluators should consult Goldberg and Agnew (2013) for guidance about estimating smart thermostat demand response program savings.

- The goal is to estimate savings for a thermostat replacement program, and customer billing consumption or AMI meter data are not available. This protocol only recommends nonparticipant comparison group methods for estimating savings from a thermostat replacement program, and construction of nonparticipant comparison groups requires the availability of billing consumption or AMI meter data.[8]
- The smart thermostat measures were installed with other residential measures or coincided with customer participation in other utility DSM programs. Many utility programs bundle installation of smart thermostats with other rebated measures, which can make it challenging to estimate the savings from smart thermostats. This protocol does not address or otherwise prescribe savings estimation methods for these situations.[9]
- The smart thermostats were installed as part of a residential new construction program, which means baseline data for participants are unavailable. This protocol does not prescribe methods to estimate energy savings from smart thermostats in residential new construction or from building energy efficiency codes and standards.

[8] There are three practical challenges with using the telemetry data to evaluate smart thermostat replacement programs. First, when telemetry data are available, they are often anonymized or aggregated to customer groups because of consumer privacy protection rules, and there is not an accepted way to establish the provenance or authenticity of the data. The inability to verify the source and completeness of the data (such as whether vendors provide a complete and accurate rendering of requested data) may limit the trust that program administrators, evaluators, and stakeholders can place in the results. Second, telemetry data for the baseline period (before the installation of smart thermostats) are not available. The absence of such data makes it difficult to establish baseline conditions for smart thermostat replacement program evaluations. Section 5.1 of this protocol discusses potential uses of telemetry data for smart thermostat evaluation in greater detail. Third, it is not possible to construct a comparison group of thermostat program nonparticipants.

[9] When thermostats are installed as part of a bundle, the reliability of savings determined through consumption data analysis depends on the variability of the mix of measures installed across customers, as well as the measures installed being unrelated to the expected savings from the smart thermostats.

3 Energy Savings Calculations

This section presents the recommended approach for estimating energy savings from smart thermostat replacement and optimization programs.

3.1 Thermostat Replacement Programs

This protocol recommends whole-home energy consumption data analysis for evaluating smart thermostat programs. Many features of the whole-home consumption analysis lend themselves well to smart thermostat program evaluation:

- This type of analysis can be applied when the baseline thermostat type (whether it was a manual or programmable unit) is unknown. In general, the baseline thermostat type will be unknown; if such information is available, it is usually self-reported by participants on the rebate application form.
- Whole-home consumption analysis can be applied without knowledge of thermostat settings before and after installation of the smart thermostat. Baseline temperature set points and schedules are usually unknown because most baseline equipment does not record such data. Reporting period temperature set points may also be unknown because of the unavailability of thermostat telemetry data. The unavailability of this information would pose difficulties for many engineering-based evaluation approaches but not whole-home billing or meter consumption approaches.[10]
- Evaluators can detect the energy savings from smart thermostats with statistical analysis of whole-home billing consumption or AMI meter data if the analysis samples are large enough. Submetering of HVAC equipment, which may be prohibitively expensive, is not necessary.
- Highly accurate whole-home billing consumption or AMI meter data are usually available from utilities for program participants and nonparticipants.
- Whole-home consumption analysis captures all nonthermostat energy use changes related to or potentially caused by the thermostats (such as changes in the usage of fans or windows).
- Whole-home consumption analysis of smart thermostat programs (not involving direct installation) accounts for in-service rates by including participant homes that did not install or that removed thermostats during the reporting period in the analysis sample.[11] (Savings from delayed installations may be partially or wholly undercounted in a first year analysis, and savings from program thermostat installations at other [nonparticipating] premises within the territory will be excluded entirely.)

[10] Baseline temperature set points can be collected through surveys, but there are concerns about the accuracy of self-reported set point data. Setpoints and runtimes may also be collected through on-site data collection and metering studies, which are often very expensive.

[11] This assumes participants in the analysis sample are customers who received a smart thermostat rebate or a kit including a smart thermostat and the evaluator cannot be certain whether the customer installed the thermostat. Evaluators using whole-home consumption data analysis should not apply an in-service rate adjustment (estimated from participant surveys) because the smart thermostat savings estimate will reflect the in-service rate.

Given these advantages and that thermostat replacement programs fit the conditions set forth in UMP Chapter 8 for whole-building consumption data analysis (Agnew and Goldberg 2017), this protocol recommends evaluating smart thermostat programs using whole-home energy consumption data analysis.

3.1.1 Challenges

The suitability of smart thermostat programs notwithstanding, there are two principal challenges with applying the UMP Chapter 8 whole-building energy consumption data analysis to the evaluation of smart thermostat programs.

3.1.1.1 Detecting Small Percentage Savings

One challenge of whole-home consumption data analysis is that, as previously noted, energy savings from smart thermostats are often expected to be a small percentage of home energy consumption. In contrast to larger whole-home energy efficiency retrofit projects that involve multiple measures, the "signal"—the expected energy savings from smart thermostats—may be small relative to "the noise"—the variability in the energy consumption data—even in a large panel regression equation that includes control variables for the customer, time period, and weather.

When the expected savings from smart thermostats are a small percentage of home energy consumption, evaluators should have realistic expectations about their ability to detect the energy impacts and the relative precision of savings estimates obtained from a consumption data analysis. As UMP Chapter 8 observes, billing analysis results that have 90% confidence and ±50% relative precision are common and may provide acceptable results for the purposes of some program evaluations (Agnew and Goldberg 2017).

By participating in the smart thermostat program planning stage, evaluators may be able to help program administrators better achieve the evaluation research objectives. Program administrators can increase the probability of detecting the smart thermostat savings by sizing their programs appropriately. Through data analysis simulations or use of statistical power formulas, evaluators can determine the probability of detecting the expected savings for analysis samples of different sizes and adjust the analysis sample size to increase the study's statistical power.[12] UMP Chapter 17 (Stewart and Todd 2020) recommends methods for sizing analysis samples using statistical power analysis. Likewise, for a given level of statistical confidence, evaluators can forecast the absolute precision with which they will be able to estimate the thermostat savings with a sample of given size. It may be possible to increase the statistical power or statistical precision by increasing the number of homes in the analysis sample, the length of the analysis period, the frequency of energy consumption data, or the number of model explanatory variables.

3.1.1.2 Self-Selection in Smart Thermostat Program Participation

Another challenge for evaluators is addressing the potential for bias in estimating savings due to the self-selection of participants into smart thermostat programs. Selection bias can cause the

[12] Evaluators should also consider how known or likely biases in the estimated savings from self-selection in program participation affect the sizing of the analysis sample. For example, if self-selection is likely to bias the savings estimate toward zero, evaluators will need a larger analysis sample than would otherwise be required to detect the savings.

estimated savings to differ from the true savings and arises from the presence of factors not controlled or otherwise accounted for in the energy consumption analysis that make participation in smart thermostat programs more or less likely and that also affect energy consumption.

Evidence for self-selection in smart thermostat program participation comes from analysis of household-level data from the U.S. Department of Energy's Residential Energy Consumption Survey (RECS 2020), which shows adopters of smart thermostats have the following differences compared to other households[13]:

- Smart thermostat households were younger on average by about nine years (58 years of age vs. 49 years) than households without a smart thermostat. Households without a smart thermostat were 2.3 times as likely to be 65 years of age or older.
- Smart thermostat households were 1.6 times as likely to have children under 17 years of age in the home (45% vs. 28%).
- Smart thermostat households were 1.8 times as likely to have an income over $100,000 (59% vs. 32%).
- Smart thermostat households were about one-third as likely to be renters (3% vs. 10%).
- Smart thermostat households were 1.8 times as likely to have someone teleworking in the home (53% vs. 29%).
- Smart thermostat households were four times as likely as households without a smart thermostat to own an electric vehicle (5.2% vs. 1.3%).
- Smart thermostat households had more desktop computers (0.8 vs. 0.6 per home), more laptop computers (2.1 vs. 1.4), tablet computers (1.7 vs. 1.1), and smart speakers (1.8 vs. 0.5).
- Smart thermostat households were 1.7 times more likely to have a new (less than two years old) central air conditioning system (21% vs. 12%) and 1.7 times more likely to have a new central heating system (19% vs. 11%).

Self-selection in smart thermostat program participation may manifest in and lead to biased savings estimates in several ways:

- In comparison to households that do not adopt smart thermostats, households adopting smart thermostats may make other changes to their homes that significantly affect the demand for electricity at or around the same time they adopt the smart thermostat. For example, they may undertake home renovations that coincide with or closely follow the installation of a smart thermostat.[14]

[13] The following statistics were calculated as RECS sample weighted averages for single-family homes with a central heating system (a heat pump or furnace) or central cooling system (n=10,544).

[14] A study of smart thermostat energy savings in the Pacific Northwest (Apex Analytics 2021, p. 24) found that about 60% of households installing smart thermostats made changes to their homes that had significant effects on energy consumption, such as beginning to use an electric vehicle, undertaking a home renovation, installing a new HVAC system, or making changes in home occupancy. DNV-GL (2021, p. 38) found smart thermostat program

- Households adopting smart thermostats may adopt other connected devices that increase their demand for electricity around the same time or after thermostats are adopted.[15] For example, adoption of electric vehicles and/or other connected devices would increase energy consumption and bias the smart thermostat savings estimate toward zero.
- Households adopting smart thermostats may have demographic or economic characteristics such as wealth, youth, increasing size, and tech-savviness that make their participation more likely and affect year-over-year changes in energy consumption.[16]
- Households adopting smart thermostats may adopt other energy efficiency or electrification measures at higher rates compared to naturally occurring adoption in the comparison group.[17] This can cause estimates of smart thermostat savings to be biased downwards.

These behaviors present a challenge for smart thermostat program evaluation because they can generate differences in trend energy consumption between the smart thermostat program treatment group and the comparison group that cannot be easily distinguished from the savings and that can therefore bias the savings estimates.

To see this, consider the intuitive and widely practiced way of estimating smart thermostat savings as a difference-in-differences (D-in-D) in energy consumption between the treatment (participant) group and the comparison group for the reporting and baseline periods. Denoting p as smart thermostat participant customer, np as nonparticipant customer, 1 as the program reporting period, and 0 as the baseline period, a D-in-D of mean consumption per customer $\bar{e}$ is defined as:

participants were more likely to add home floor area (+3 percentage points) and lighting (+9 percentage points). DNV-GL (2022, p. 38) finds smart thermostat rebate homes were more likely than matched nonparticipants to have recently added living space to the home (+4 percentage points) and to use more lighting (+9 percentage points).

[15] Apex Analytics (2021, p. 24) found 35% of smart thermostat homes installed other connected devices before and after the smart thermostat was installed, 25% installed other connected devices only after the smart thermostat installation, and 10% installed other connected devices at the same time or before the smart thermostat installation. Consistent with the Apex Analytics findings about additions to home electricity loads after smart thermostat adoption, Guidehouse (2020a, p. 133) observed a net increase in baseload energy consumption on mild (non-HVAC using) days averaging about 0.24 kWh/day or 78 kWh/year from a future participant comparison group study. DNV-GL (2021, p. 38) found smart thermostat program participants were more likely to add an electric vehicle (+4 percentage points) and a refrigerator (+9 percentage points). DNV-GL (2022, p. 38) finds smart thermostat rebate homes were more likely than matched nonparticipants to have recently added electric vehicle charging to the home (+ 6 percentage points).

[16] DNV-GL (2020, pp. 34–36) found that smart thermostat program participants were more likely than nonparticipants to be homeowners, reside in newer and larger homes, have central air conditioning, and have higher incomes. Adopters were also more likely to have recently experienced an increase in household size and to have added electric vehicle charging to the home. DNV-GL (2021, pp. 36–39) made similar findings about smart thermostat program participants.

[17] For example, in California, smart thermostat program participants were more likely to install water-saving aerators and perform duct sealing through the utility's rebate programs by about 2 percentage points (DNV-GL 2020, p. 34). Evaluators can drop smart thermostat participants who installed other rebated measures from the analysis sample or retain these participants and attempt to adjust the savings estimates for this additional participation. There is more potential for efficiency improvements undertaken outside of utility energy efficiency programs to bias the savings estimates because these improvements are likely to be unobserved.

$$\text{D-in-D savings} = (\bar{e}_{p,1} - \bar{e}_{p,o}) - (\bar{e}_{np,1} - \bar{e}_{np,o}) \quad (1)$$

The first difference equals the smart thermostat program effect plus any other consumption change between the reporting period and the baseline period unrelated to the program for participant customers. The second difference is the change in nonparticipants' consumption between the reporting and baseline periods. If the participant group, absent adoption of the thermostat, and the comparison group would have followed the same reporting-period consumption trend, any time-invariant (pre-existing) level difference in consumption between the groups will be differenced out and the D-in-D calculation will yield an unbiased estimate of the savings. Figure 1 illustrates this "parallel trends" assumption that must hold for the D-in-D calculation to be unbiased.

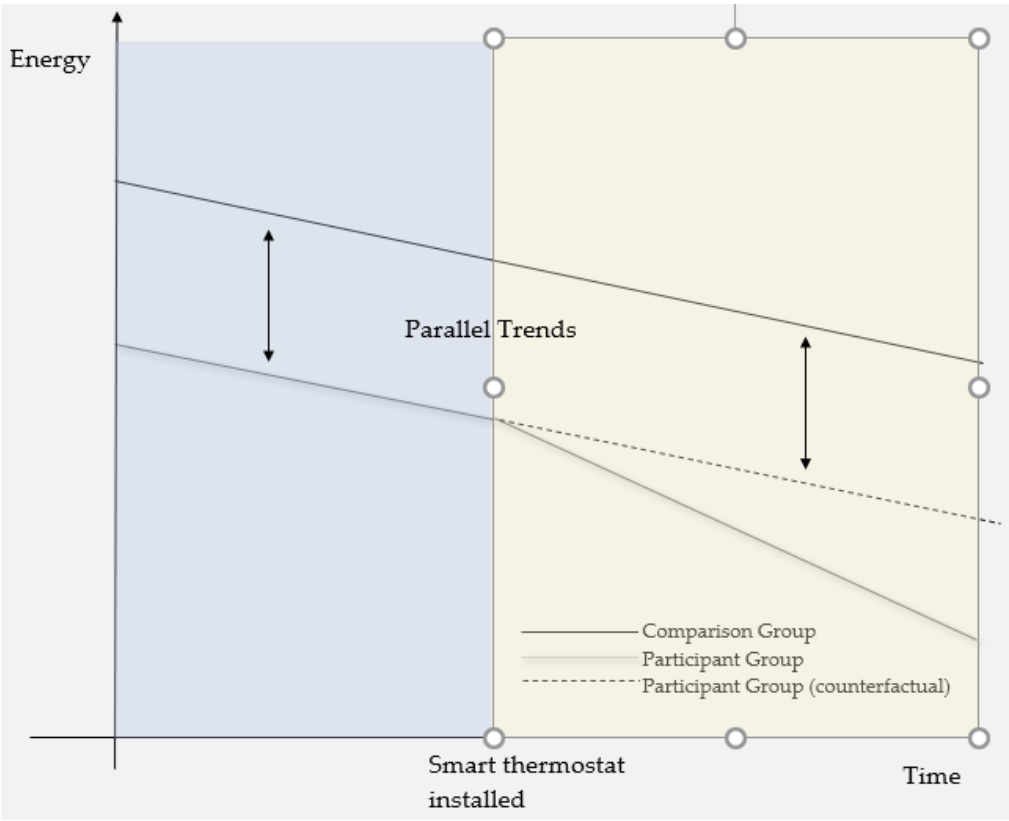

Figure 1. Parallel trends assumption

However, if the parallel trends assumption is not satisfied, perhaps because smart thermostat participants simultaneously install other appliances that raise consumption and these installations are unobserved by the evaluator, the groups would have followed different trends if the participant group had not received the smart thermostat and the D-in-D savings estimate will be biased. In this situation, shown in Figure 2, the D-in-D estimate will yield a downward-biased estimate of the smart thermostat savings because the participant group counterfactual consumption is trending upward relative to the comparison group.[18] In subsequent sections, this

[18] The bias could go in the opposite direction if, for example, homes installing smart thermostats simultaneously replace other home appliances with more energy efficient models.

protocol discusses research designs to minimize the potential for unobservable trend differences to arise and statistical tests for detecting such differences.

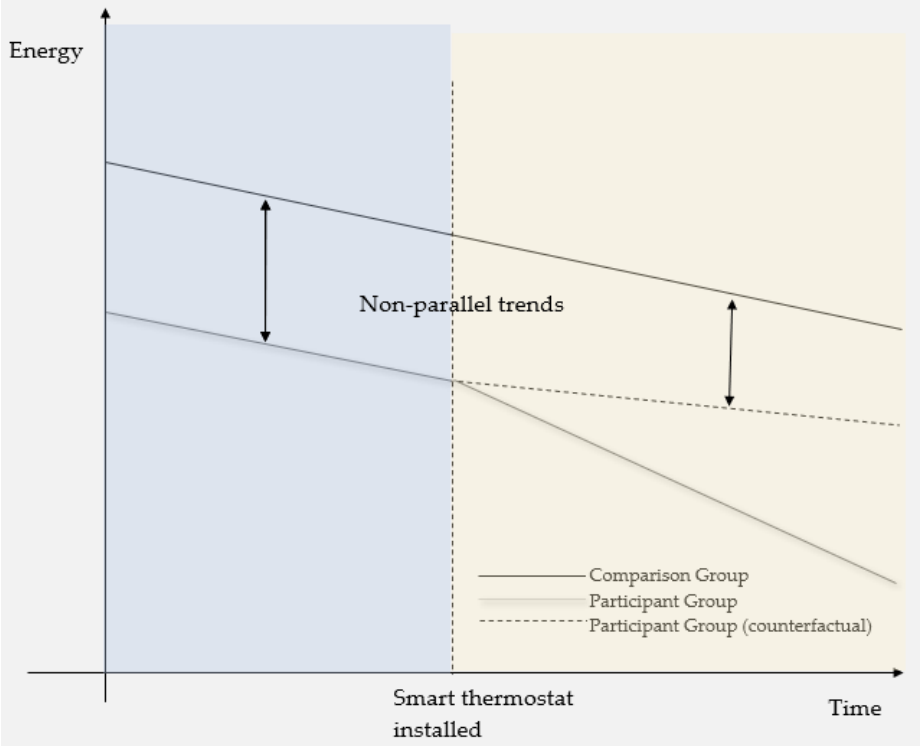

Figure 2. Violation of parallel trends assumption

The self-selection phenomenon is not unique to smart thermostat programs—it affects participation in all utility energy efficiency programs in which customers self-enroll. However, self-selection may present a bigger challenge for evaluation of smart thermostat programs. First, given the newness of smart thermostats as a technology, self-selection in terms of who adopts these products may be quite severe. As previously noted, households that adopt smart thermostats tend to be wealthier, younger, faster-growing, and tech-savvy. Second, any bias from self-selection may be large compared to the typical smart thermostat percentage savings, which are usually less than 5% of home consumption and often less than 3% when evaluating cooling savings in moderate climates.[19] Even if a bias in smart thermostat savings estimates is only 1% to 2% of consumption, this bias may be enough to confound assessments of whether programs save energy and are cost-effective. In contrast, bias of 1% to 2% will be less consequential for assessing the cost-effectiveness of retrofit programs with larger percentage savings.

3.1.2 Evaluation Approaches

The section recommends specific evaluation approaches and methods for minimizing self-selection bias in estimating smart thermostat program savings. The evaluation design recommendations in this protocol align closely with those in UMP Chapter 8 (Agnew and

[19] See Nexant (2017, p. 3), Guidehouse (2018, p. 6), Guidehouse (2020a, p. 11), and DNV-GL (2020, pp. 42 and 44).

Goldberg 2017), UMP Chapter 17 (Stewart and Todd 2020), and the SEE Action Report (State and Local Energy Efficiency Action Network 2012).

All recommended evaluation approaches require a randomized control or quasi-experimental comparison group to adjust for time-varying factors unrelated to the program that affect the electricity consumption of smart thermostat program participants (such as changes in macro-economic conditions like a recession, increasing energy prices, natural occurring efficiency, pandemics, and weather). With quasi-experimental approaches, evaluators can use a comparison group to try to isolate the smart thermostat savings and to avoid a situation in which the savings estimate is dependent on the model specification. When this happens, the savings estimate changes significantly when different variables are included in the regression model, raising doubt about the accuracy of any estimate. By controlling for unexplained non-program-related changes in energy consumption between the baseline and reporting period, the use of a comparison group makes this model dependency less likely to occur.

3.1.2.1 Randomized Field Experiments

Randomized field experiments are the gold standard in energy efficiency program evaluation and are the most reliable way to obtain unbiased savings estimates, particularly when bias from self-selection is a concern (Stewart and Todd 2020). By randomizing who installs a smart thermostat (through a randomized controlled trial [RCT]) or who receives encouragement to install a smart thermostat (through a randomized encouragement design [RED]), these approaches ensure that receipt of treatment or encouragement is uncorrelated with customer characteristics and that an unbiased savings estimate can be obtained by comparing the randomized treatment and control groups. RCTs have been widely used and shown to be effective for evaluating large-scale residential behavior-based programs with small savings (Allcott 2011, 2015). To a lesser degree, evaluators have also used RCTs to evaluate smart thermostat programs due partly to the challenges of determining customer eligibility for a smart thermostat and installing the thermostats in customer homes. See DNV-GL (2015) and Brandon et al. (2021) for examples.

Because randomized experiments are expected to produce unbiased savings estimates, this protocol encourages evaluators of smart thermostat programs to use these methods when possible. Table 1 lists the recommended methods for randomized experiments. An RCT for smart thermostat replacement would involve recruiting customers who are interested in and eligible to install smart thermostats into the experiment, randomly assigning some customers to receive and install a thermostat (these customers become the treatment group), and either denying or delaying the installation of smart thermostats for the rest of eligible customers (these customers become the control group). An RED for smart thermostat replacement would involve identifying eligible customers and randomly assigning some of them to receive direct encouragement to participate in the program. These customers would constitute the treatment group, and some portion of them will comply by opting into the program and receiving a smart thermostat. The control group does not receive the encouragement and provides the baseline for estimating the savings. The encouragement provides an exogenous source of random variation in participation

that can be used to estimate the average savings per complier with the encouragement or the average savings per encouragement group customer who receives a smart thermostat.[20]

More detailed descriptions of RCTs and REDs, and the nuances involved in using them for evaluations, can be found in UMP Chapter 17 (Stewart and Todd 2020), UMP Chapter 8 (Agnew and Goldberg 2017), and the SEE Action Report (State and Local Energy Efficiency Action Network 2012). From a customer experience perspective, REDs are usually the preferred approach because it is unnecessary for program administrators to delay or deny the program participation of interested customers. The main challenge of using an RED is detecting small percentage treatment effects in the population of encouraged customers. Program administrators need to run very large experiments and make sure the encouragement lifts the participation rate relative to the nonencouraged customer (control) group to increase the probability of detecting the savings.

[20] Encouraged customers who opt into smart thermostat programs include customers who opt in due to the encouragement (known as "compliers") and customers who would opt in whether or not they receive the encouragement (known as "always-takers").

Table 1. Randomized Field Experiment Approaches for Smart Thermostat Programs

Approach	Description	Advantages	Potential Challenges
Randomized Controlled Trial (RCT)	• Opt-in recruit-and-delay or recruit-and-deny • Eligible and interested customers are randomly assigned to treatment (receive smart thermostat) or to control (do not receive treatment or receive delayed treatment)	• Yields unbiased estimate of intent-to-treat treatment effect for population studied • Controls for self-selection in program participation	• Potential dissatisfaction from customers whose enrollment was delayed or denied • Verifying customers are eligible (such as having compatible HVAC equipment) to participate in the experiment • Noncompliance with assigned treatment such as not installing or uninstalling the thermostat • Requires effort and coordination to plan the experiment • Requires monitoring of experiment implementation • Results may not be externally valid because of program or experiment eligibility requirements and customer willingness to participate in experiment • Recruitment process can change control group behavior, leading to biased estimates of effects
Randomized Encouragement Design (RED)	• Eligible customers are randomly assigned to receive encouragement (encouragement group) or not to receive encouragement (control group) • Any encouragement or control group customer can participate in the program	• Yields unbiased estimate of local average treatment effect of encouragement, which, depending on whether control group customers participate in the program, equals the net savings for all customers who receive a thermostat or the net savings for compliers with the encouragement (those who receive a thermostat because of the encouragement)[a] • All interested customers in the randomized encouragement and control groups can participate—no need to delay or deny participation to any customer	• Insufficient compliance with the encouragement (such as a small difference in program uptake between the encouragement and control groups) will lead to low precision estimates • Large sample sizes are required to obtain precise savings estimates • Requires effort and coordination to plan experiment • Requires monitoring of experiment implementation • Results may not be externally valid because of program or experiment eligibility requirements

[a] To obtain an estimate of the local average treatment effect, the encouragement group customer savings must be scaled by the difference between the encouragement and control groups in the percentage of customers receiving thermostats. As another option, evaluators can employ instrumental variables using a random assignment of encouragement as an instrument for participation. If control group customers are not allowed to participate in the program, the local average treatment effect will be equal to the average treatment effect for all treated customers. If control group customers can participate, the local average treatment effect equals the average treatment effect for compliers with the encouragement. More details on REDs can be found in UMP Chapter 17 (Stewart and Todd 2020) and UMP Chapter 8 (Agnew and Goldberg 2017).

While RCTs and REDs have many benefits, this protocol recognizes that implementing these approaches can prove challenging, and that few smart thermostat programs have been evaluated this way. Also, regulatory or program policies or program design elements often prevent the use of randomized field experiments. In these cases, the evaluators will need to employ a quasi-experimental design.

3.1.2.2 Quasi-Experimental Approaches

When smart thermostat programs are not designed as randomized experiments, thus precluding evaluation via the methods described in Table 1, evaluators can use quasi-experimental approaches to estimate smart thermostat program savings. The limitations of using quasi-experimental techniques are noted in UMP Chapter 8 (Agnew and Goldberg 2017):

> An observed change in consumption between pre- and post-installation periods includes the effect of the whole-building intervention itself, along with the effects of other factors unrelated to the program that may occur in the same timeframe. These effects could include changes in occupancy, physical changes to structure, behavioral changes, weather, etc. Without special attention, these non-program effects may be conflated with program effects, leading to incorrect estimates of program effects or savings.

Controlling for time-varying, nonprogram consumption effects requires a comparison group of nonparticipants. The resulting baseline will be accurate if the comparison group's energy consumption accurately represents the counterfactual consumption of smart thermostat participants if they had not participated in the program.

As Table 2 shows, this protocol recommends two comparison group approaches and presents a third approach that evaluators can use at their discretion if the data required to implement the first two approaches are not available. However, the savings estimates from the first two recommended approaches are more likely to be closer to the true savings; that is, they are likely to be less biased. The recommended approaches better address the greatest threat to internal validity: the potential for participants to self-select into smart thermostat programs based on characteristics that affect the probability of participation and future energy consumption. Minimizing self-selection bias is important because any bias will likely be large relative to the expected smart thermostat energy savings, which tend to be less than 5% of home energy consumption.

In consideration of the potential for bias in smart thermostat savings estimates obtained from quasi-experimental comparison group studies, this protocol recommends evaluators discuss the following topics in their evaluation reports:

- The potential for unobserved differences between participants and comparison group nonparticipants to bias the savings estimates, with references to the potential biases noted in this chapter as appropriate
- Attempts to test for the presence of these factors, the test results, and any attempts to correct or adjust the savings estimates for bias

- How the potential for any bias in the savings estimates affects the interpretation of the results, particularly regarding program evaluation goals and cost-effectiveness.

When making adjustments to the quasi-experimental modeled savings to account for suspected biases, evaluators should be restrained and transparent. There should be compelling evidence that an adjustment is likely to improve the accuracy of the modeled smart thermostat savings estimate, and evaluators should be fully transparent about how the adjustment was made and its impact on the savings. Ideally, in the evaluation planning stage, evaluators and program administrators would anticipate potential sources of bias and incorporate protocols for addressing such bias in the program evaluation plan. This advance planning will minimize the potential for prior beliefs about the smart thermostat savings to unduly influence the savings adjustments. At a minimum, evaluators should write out the plan for addressing bias in advance, in enough detail to ensure that every step of the approach—statistical tests, survey questions, calculation methods, criteria for pulling the trigger on a potential adjustment—are objective and direction-neutral (e.g., if the comparison group purchased more electric vehicles than participants [the opposite of what is expected], this is where the evaluator would spell out how this would lead to a downward [rather than an upward] adjustment to the thermostat savings).

An important difference between the estimation approaches is that the interpretation of the savings estimate will depend on how the comparison group was constructed and which customers were included. As UMP Chapter 8 explains, a comparison group analysis may yield an estimate of gross savings, net savings, or, more likely, something in between depending on the comparison group used (Agnew and Goldberg 2017). Essentially, if the comparison group has no measure adoption or virtually none of its own, the D-in-D estimate produces gross savings, i.e., the effect of the measure itself on energy consumption. If the comparison group has the same level of natural adoption as participants, the D-in-D estimate produces an estimate of net savings. Often, the D-in-D approach produces a savings estimate somewhere in between because of differences between participants and the matched comparison group in rates of natural adoption.

Future Participants (Variation in Timing of Participation)

The most reliable quasi-experimental approach uses variation in the timing of smart thermostat program participation to estimate savings. This approach usually compares current smart thermostat program enrollees to later program enrollees and therefore uses differences in participant enrollment dates to define the comparison group.[21] An example of smart thermostat evaluations using future participants is provided in Guidehouse (2020a).

As UMP Chapter 8 (Agnew and Goldberg 2017) describes, the future participant comparison can be implemented in two main ways:

- **Static comparison group:** this approach compares smart thermostat participants to the same group of nonparticipants throughout the reporting period. All participants become participants before the reporting period begins, and all nonparticipants do not become participants until after the reporting period ends.

[21] This comparison group can comprise all prior and/or future participants or a matched subsample of prior and/or future participants.

- **Time-varying (rolling) comparison group:** this approach allows nonparticipants to become smart thermostat program participants over the reporting period. In each reporting interval, participants are compared to all nonparticipants in that interval, who may become participants in the next or another future interval.

The future participant approach is attractive because future smart thermostat program participants are likely to have similar energy end uses and to experience similar changes in energy consumption patterns as participants over time. The main difference between participants and the comparison group of future participants is not *whether* they decided to install a smart thermostat but *when* they decided to install one. Provided the program population remains stable over time, this approach can mitigate some aspects of the potential for self-selection bias, specifically long-term trend differences in consumption preceding the thermostat installation. However, this approach may not mitigate self-selection bias related to short-term trend differences associated with the act or timing of installation. For example, if smart thermostat program participants are more likely to undertake home renovations around the time they install a smart thermostat, future participants will also be more likely to undertake those renovations, but the renovations will not occur until the thermostat is installed, possibly a year or more later. In this case, the baseline consumption of future participants would be too low and the smart thermostat savings estimate would be biased downward. Evaluators can use participant surveys to test for such differences.[22] Nonetheless, of all the quasi-experimental methods, this approach is expected to do the best in reducing the potential for bias from unobserved customer attributes and other time-varying changes in consumption.

For this approach to be valid, it is important that the demographic composition of the participant population and the smart thermostat program implementation not change significantly over time in ways that produce unobservable trend differences in consumption between participants and future participants. For example, if the earlier participants tend to be more energy conscious and efficient than later participants, this could bias the estimate of savings upward. When possible, evaluators should inspect and test for the presence of level and trend differences in energy consumption between current participants and future participants in the baseline period.[23] Such differences would suggest the presence of omitted variables that could invalidate the baseline and bias the savings estimates. In general, a minimum of two years of baseline period monthly consumption data are required to test for trend differences. Guidance about what evaluators should do if they detect such differences is provided in Section 3.1.4.3.

Evaluators should enhance the validity of the future participant approach by matching current participants to similar future participants based on observable customer and home characteristics such as energy consumption, age, and income. With matching, participants will only be compared to future participants who have similar energy consumption and attributes.

Requirements for implementing the future participant approach are also discussed in UMP Chapter 8 (Agnew and Goldberg 2017). The most important requirements include having a large enough sample of future participants as well as two years of consumption data. One year of

[22] See Apex Analytics (2021) and Guidehouse (2021) for household survey data analysis showing smart thermostat adoption preceded or coincided with significant changes in household energy consumption.
[23] For an example of testing using a rolling comparison group, see Harding and Hsiaw (2014).

consumption data will be needed for current and future participants for the baseline period and one year of consumption data will be needed for both groups for the reporting period. In addition, evaluators who want to match current participants to future participants on any attributes beyond energy consumption will need customer and home characteristics data.

The variation in timing of the participation approach yields an estimate of the smart thermostat program gross energy savings, as the analysis sample only includes program participants. The estimate is gross because the comparison group can reasonably be assumed not to have naturally occurring levels of smart thermostat adoption due to the proximity of their future adoption. To obtain an estimate of net program savings, a separate freeridership analysis will need to be conducted.

Matched Nonparticipants on Basis of Energy Consumption and Customer Attributes

The second recommended approach is matching smart thermostat program participants to nonparticipants. This approach should be used when it is not possible to implement the future participant approach (that is, when the program population or implementation was not stable over time or the program population changed over time) and data required for matching are available for program participants and nonparticipants.

This approach requires matching participants to nonparticipants based on the customer baseline period energy consumption and observable characteristics, including, most importantly, income and age. As previously discussed, adopters of smart thermostats in recent studies have tended to be younger and have higher incomes, and both characteristics are correlated with growing energy consumption. With data on income and age, evaluators may be able to construct a comparison group that more closely resembles the participant group.[24] While an improvement over only matching on energy consumption, this approach is less attractive than the future participant approach because participants and the matched comparison group may still not be similar in their unobservable characteristics, including motivations to save energy and interest in and purchases of technology such as electric vehicles and "smart" home devices. For example, while smart thermostat participants tend to be young and have high incomes, such households with an interest in adopting smart devices may be the most likely to participate in smart thermostat programs. In general, it is not possible to identify young and high-income nonparticipants who have these interests. As a result, even if the matched comparison group and the participants have similar ages and incomes, the energy consumption of the matched comparison group and the counterfactual consumption of participants may follow different trends, providing a biased estimate of savings.

The matched comparison group approach requires baseline period energy consumption, household demographics, and housing characteristics data to be collected, and participants to be matched to nonparticipants using a matching algorithm (discussed in Section 3.1.3.2, Matched Comparison Group Construction). As with the future participant approach, when feasible, evaluators should test for level and trend differences in energy consumption during the baseline period. However, evaluators should be aware demographic data for matching may be incomplete

[24] Matching on customer demographic and home characteristics in addition to energy consumption can reduce the quality of matches on energy consumption, and therefore evaluators need to use their best judgement about whether to match on these variables if the quality of the match on consumption significantly worsens.

or unavailable for many residential customers, which can significantly limit the pool of smart thermostat participants and nonparticipants available for matching and the consumption analysis. If demographic data are unavailable for many customers, evaluators should carefully weigh the benefits of matching on demographics against the costs of limiting the sample to customers with such data available and the ramifications for the external validity of the savings estimates.

As UMP Chapter 8 explains, the matching of participants to nonparticipants and the estimation of a D-in-D regression usually yields a savings estimate between net and gross savings (Agnew and Goldberg 2017). As smart thermostat participants likely include a higher percentage of would-be natural adopters or freeriders than nonparticipants, the comparison group will account for some but not all freeridership in the participant population.[25] As in DNV-GL (2020), evaluators can adjust the D-in-D savings estimates to obtain estimates of smart thermostat gross and net savings.[26]

Matched Nonparticipants on Basis of Energy Consumption

Without customer demographics data, evaluators will only be able to match participants to nonparticipants based on energy consumption. Given that smart thermostat households tend to have higher incomes, be younger, and have increasing energy consumption, there will be enhanced potential for bias in estimating smart thermostat program savings when evaluators can only match on energy consumption. Accordingly, this matching approach, the last option in Table 2, should only be used when data to implement one of the other two approaches are not available.

[25] Differences between participants and the matched comparison group in their propensity to adopt a smart thermostat in the absence of the program only affect the distinction between gross and net savings if the naturally occurring adoption would have occurred during the evaluation period. Measured against a baseline of comparison group customers who adopted smart thermostats prior to the baseline period, the savings of smart thermostat program participants would be net savings.

[26] To obtain an estimate of smart thermostat gross energy savings per participant, DNV-GL (2020) makes adjustments to the D-in-D savings estimate for (1) differential trends in baseload consumption between the treatment and comparison group; (2) the prevalence of space heating and space cooling in participant and comparison group homes; and (3) the prevalence of smart thermostats in the comparison group. The authors then apply a net-to-gross factor to the adjusted gross savings to estimate the net program savings.

Table 2. Quasi-Experimental Evaluation Approaches for Smart Thermostat Programs

Comparison Group Approach	Description	Potential Advantages	Potential Challenges	Gross or Net Savings
Future participants (variation in timing of participation)	Future or prior participants who may be matched based on energy consumption and demographics[a]	• Comparison group comprises smart thermostat participants • Quasi-experimental approach minimizes potential for self-selection bias	• Requires customers to enroll for two or more years • Time-varying participant characteristics or program implementation can invalidate the comparison group	Gross
Comparison group matched on consumption and demographics/ housing characteristics	Comparison group of nonparticipants matched on energy consumption and other characteristics	• Straightforward conceptually and easy to implement for most smart thermostat programs • Will account for potential bias from matching variables	• Unobservable characteristics from self-selection may lead to varying consumption trends and bias savings • May not produce unbiased savings estimates • Requires collecting data on customer demographics	Likely between net and gross
Matched comparison group with customers matched on consumption only	Comparison group of nonparticipants matched solely on energy consumption	• Only requires billing consumption data to implement	• Least robust of recommended methods for addressing selection bias • Unlikely to produce unbiased savings estimate	Likely between net and gross

[a] Prior participants are customers who participated in the smart thermostat program before the start of the evaluation baseline period. As prior participants, they are expected to be similar to participants whose savings will be evaluated if the program and participant population have remained stable over time. The difference in prior participant consumption between the evaluation reporting period and the evaluation baseline period is the naturally occurring change in energy consumption that participants would have experienced if they had not installed a smart thermostat. Taking the difference in savings in the reporting baseline period between participants and prior participants yields an estimate of the smart thermostat savings. See UMP Chapter 8 (Agnew and Goldberg 2017) for an illustration of this method.

Discouraged Evaluation Approaches

This protocol discourages evaluators of smart thermostat programs from using a random sample of residential nonparticipant customers as a comparison group. Given the high potential for self-selection in smart thermostat program participation, it is likely that the savings estimates from such a comparison will be biased.

Also, this protocol discourages evaluators from using the participant's energy consumption before participation is adjusted for differences in weather to construct the baseline. The validity of this within-subject design depends strongly on the ability to accurately adjust the participant's baseline for differences in weather and other time-varying factors. While it may be possible to

adjust the baseline for differences in weather, it is usually not possible to adjust for the effects of other time-varying factors.

3.1.3 Data Analysis

Data analysis concerns evaluation data requirements and collection, matching of participants to nonparticipants for quasi-experiments, and regression modeling to estimate savings.

3.1.3.1 Data Requirements

Regardless of the experimental or quasi-experimental approach, the following data are needed to evaluate smart thermostat program savings:

- Utility customer consumption data from a utility billing or AMI meter data system
    - This protocol recommends analyzing customer daily or monthly interval consumption data to estimate energy savings and analyzing hourly or subhourly consumption data to estimate subdaily energy impacts. Hourly or subhourly consumption data may also be aggregated to the daily level to estimate energy impacts.[27]
- Program tracking data
    - Program tracking data concern customer enrollment dates, equipment installation dates, measure descriptions, and incentive payments. These data are required to identify smart thermostat program participants and nonparticipants and determine participation start dates.
- Weather data
    - Weather data usually includes hourly or daily outdoor temperature data that can be used to calculate heating or cooling degree days or the average temperature for the analysis time periods (which may be days, calendar months, or customer billing periods).

For additional considerations related to program tracking data and weather data, see Sections 4.5.2.2, 4.5.2.3, and 4.5.2.4 of UMP Chapter 8 (Agnew and Goldberg 2017). If monthly consumption data are used for the evaluation, see Section 4.5.2.1 of UMP Chapter 8. For helpful guidance about assessing data sufficiency and quality and preparing consumption data for analysis, this chapter encourages evaluators to consult Section 2 of the CalTRACK guidelines (CalTRACK 2018).[28]

To implement the recommended quasi-experimental comparison-group matching approaches, evaluators will also need data on customer demographic and home characteristics. When available, the most important such data will be utility customer income and age.

In addition, evaluators may find it useful to collect data on other characteristics such as dwelling type or heating fuel to match or estimate savings.

[27] With the implementation of AMI in more jurisdictions, there is growing availability of hourly data, making it possible to estimate savings for specific event windows or times of the day.
[28] See https://docs.caltrack.org/en/latest/methods.html

3.1.3.2 Matched Comparison Group Construction

Matching of participants to similar comparison customers is a preprocessing stage in the savings analysis designed to minimize baseline period differences between participants and the matched comparison group and to minimize the savings estimates' dependence on model specification choices.[29] As a first step in selecting a comparison group, the potential comparison group population should be limited to utility customers who would be eligible for the program, especially if eligibility requirements are strongly correlated with energy consumption. For example, if utility customers must qualify as low income to participate in the smart thermostat program, in most cases evaluators will want to restrict potential matches to other low-income customers. This first step of excluding ineligible customers constitutes a type of exact matching.

Next, one of several statistical matching methods can be used to select the matched comparison group based on observable customer characteristics. For example, if smart thermostat participants are more likely to use electricity for space heating, participants and comparison customers can be matched based on heating fuel type to increase the similarity of the samples in this respect and reduce the need for modeling electricity consumption as a function of heating fuel. Evaluators should only match on variables correlated with participation and expected to affect energy consumption. Also, to the extent that observable characteristics that are controlled by this preprocessing of the data are themselves good proxies for unobservable characteristics associated with self-selection, matching can also serve to mitigate self-selection bias, though this is not the primary purpose of the matching procedure. In fact, as previously noted, even comparison groups that are closely matched on key observable characteristics such as income and age may not be valid because of unobservable differences affecting participation and trend electricity consumption.[30]

There are several good options for constructing a matched comparison group sample:

- ***Propensity-score matching*** involves modeling the propensity to participate in the smart thermostat program as a function of observable utility customer characteristics and then predicting the propensity of participants and comparison customers to participate. Participants are then matched to comparison customers with the closest propensity scores. This method reduces multidimensional differences between participants and nonparticipants to a single score, which reduces the computational burden of using this method, but also creates potential for matched participants and matched comparison customers to not be well-balanced on some observable characteristics. There are many technicalities to implementing propensity score matching, including whether to employ one-to-one or many-to-one matching, determining which variables to include in the propensity scoring model, and deciding whether to trim poorly matched observations from the analysis sample using a caliper. Imbens and Rubin (2015) is a very good reference for these considerations.

[29] We use "comparison customer" rather than "nonparticipant" when referring generically to matching to avoid confusion about situations when evaluators match participants to future participants. In this context, it is ambiguous whether "nonparticipant" applies to future participants.

[30] Evaluators may be able to employ advanced econometric methods to control for selection bias based on these unobservable characteristics; see DNV-GL (2017) and Goldberg et al. (2017). Some methods require functional form assumptions about the model and distributional assumptions about the error term.

- ***Mahalanobis distance matching*** selects the nearest comparison customer using a measure of distance based on *weighted* differences in observable characteristics between participants and comparison customers. Distance is calculated as a weighted Euclidean distance, using the sample covariance matrix of the matching characteristics. This approach is simple and often very effective at identifying comparison customers with similar observable characteristics and yielding a well-balanced sample of matched comparison customers. DNV-GL (2020) is an example of a smart thermostat evaluation that uses propensity score matching and Mahalanobis distance matching.
- ***Exact and Euclidean distance matching*** involves first conducting exact matching on demographic data (such as income and age ranges) then choosing the best match on energy usage based on the Euclidean distance within the group of exact matches on demographics. This approach is straightforward and effectively identifies comparison customers with similar observable characteristics, yielding a well-balanced sample of matched comparison customers. Guidehouse (2020a) is an example of a smart thermostat evaluation that uses exact and Euclidean distance matching.
- ***Coarsened exact matching*** is a conceptually simple approach for obtaining a matched comparison group that is well balanced on observable characteristics. It involves first stratifying the participant sample using different customer characteristics—such as by ranges of energy consumption, income, and age—then sampling comparison customers from each stratum in proportion to participants in the stratum. This should result in a participant and matched comparison sample that is exactly balanced on the coarsened matching variables. Iacus, King, and Porro (2011) is a good reference for this approach.

Regardless of which method is used, because smart thermostat adoption is highly correlated with both income and age, this protocol recommends matching participants to comparison customers based on these two variables and energy consumption. However, data on income and age may be unavailable for many of the utility's customers, and evaluators should assess the impacts of the data availability on the size and representativeness of the final analysis sample.[31]

In addition, the accuracy of the baseline may be improved by matching on other demographics, housing characteristics, and energy variables that may be correlated with participation and energy consumption:

- Other demographics, such as education and home ownership
- Dwelling unit type (single family, multifamily, other)
- Geography (zip code or census tract if feasible)
- Energy end uses (such as natural gas space heating, electric water heating, or electric vehicle ownership)
- Energy efficiency or other DSM program participation.

[31] If evaluators use a future participant approach and the utility collects data on participant income and age from participants, the availability of income and age data may be limited due to survey nonresponse. If utilities collect data on income and age from third-party data providers, head-of-household age and household income may be missing for many customers or imputed for others and therefore subject to error.

Evaluators can also match on participation status in other energy efficiency or DSM programs or measures. Typically, this is necessary only for large programs or measures that overlap significantly with thermostat replacement. An example would be a home energy reports program for utilities that have a very large portion of their customer base enrolled in such a program.

Note that matching on demographic and housing characteristics will not improve and may even reduce the alignment on baseline energy consumption between participants and matched comparison customers. As the demographic and housing variables are intended to account for trend consumption differences, there may be a trade-off between accounting for future consumption trends and aligning on baseline period energy consumption. Evaluator discretion will be needed to determine which set of matching criteria are most appropriate.

Assessing Match Quality

After completing the matching, this protocol recommends always assessing the quality of the matches between participants and nonparticipants. There are several checks that should be performed, and the findings of the match quality assessment should always be reported so the validity of the baseline can be judged.

First, after matching participants to comparison customers, check for balance on the time-invariant attributes used in matching. If matching worked as intended, it should reduce any differences between participants and matched comparison customers in these matching variables. This can be checked by comparing the mean values and distributions of the matching variables for participants, matched comparison customers, and all comparison customers. The comparison to all comparison customers will show whether matching was an improvement over taking a simple random sample of comparison customers.

Second, check for balance between participants and matched comparison customers in baseline period energy consumption and other time-varying characteristics such as energy efficiency program participation. Use plots and summary statistics to check for level and trend differences.

Third, check for balance on observable attributes not used in matching.[32] This provides yet another check on the quality of matches. However, it is most important that participants and matched comparison customers are balanced for variables that affect both participation and energy consumption.

Fourth, if enough pretreatment energy consumption data are available, compare the energy consumption of participants and matched comparison customers outside the window used in the matching. Such a test can be used to detect the presence of omitted variables that can bias the savings estimate. For example, if 24 months of pretreatment consumption data are available, it may be possible to match participants to comparison customers using the first 12 months and then check the quality of the match using consumption data for the next 12 pretreatment months. This out-of-sample comparison is useful for identifying imbalances or pre-existing trends in energy consumption and can be employed on other time-varying customer characteristics.[33]

[32] See DNV-GL (2021, p. 66) for an example of this check.
[33] See DNV-GL (2021, p. 41) for an example of this validation check.

3.1.4 Savings Estimation

3.1.4.1 Defining Baseline and Reporting Periods

An initial step in estimating smart thermostat program savings is to define the baseline and reporting periods. The reporting period is when savings will be estimated. The baseline period is before the smart thermostats were installed and provides the baseline (pre-thermostat installation) consumption for the reporting period. The baseline and reporting periods may be defined identically or allowed to vary between smart thermostat program participants. The baseline and reporting periods should be defined using the thermostat installation date. For direct install programs, this information should be available from the program administrator. For customer self-install programs, the thermostat installation date will be unknown. However, a reasonable proxy for the thermostat installation date is when the thermostat was first connected to the internet. This information may be available from the smart thermostat vendor.[34] Also, while a participant's thermostat purchase or rebate application date may be known, these dates are unlikely to correspond to the installation date. As a result, evaluators should exercise care in defining the baseline and reporting periods.

This protocol recommends that when the smart thermostat installation dates are known, evaluators should define the reporting period to begin at least 30 days after the installation date. This buffer (1) allows time for the household to program the smart thermostat and for the thermostat to learn the household's space conditioning schedule and behaviors; and (2) eliminates the possibility that the customer bill for the installation month includes consumption for both the preinstallation and post-installation periods in the case that the smart thermostat was installed during the middle of the customer billing cycle. The first consideration means that even when analyzing daily energy consumption data and knowing the installation date, evaluators should exclude the first 30 days after installation from the reporting period.

When installation dates of smart thermostats are unknown, this protocol recommends that evaluators work with the smart thermostat vendor(s) to determine the first-connected dates for the devices and begin the reporting period 30 days after the first-connected date. If both the installation and first-connected dates are unknown, evaluators should consider starting the reporting period 60 days after the rebate application date or thermostat purchase date. This additional time allows for participant delays in installing the smart thermostats in their homes.

3.1.4.2 Regression Modeling

After selecting a matched comparison group and defining the baseline and reporting periods, this protocol recommends using regression analysis to estimate savings. Regression analysis will help to isolate the smart thermostat program savings by controlling for the consumption impacts of time-varying factors such as weather as well as for the impacts of any remaining differences in the matching variables between the treatment and control groups. The remainder of this section discusses considerations for regression model forms. The model forms discussed here assume the availability of monthly billing consumption data and are discussed in greater length in UMP Chapter 8 on whole-building retrofits (Agnew and Goldberg 2017) and UMP Chapter 17 on

[34] Some thermostat vendors require customers to consent to the release of this information so it may not be available for all participants.

behavioral programs (Stewart and Todd 2020). However, the model specifications are valid for daily, weekly, or bi-monthly consumption data with minor redefinitions of the variables.

Smart thermostat program evaluators have three main options for regression modeling of electricity consumption: a two-stage approach, a two-way fixed-effects panel regression, or a lagged dependent variable panel regression. Regardless of the approach, evaluators should report regression coefficient standard errors, regression model fit statistics such as the R^2 and F statistics, and confidence intervals for the savings estimates.

Two-Stage Approach

In the first stage, the evaluator fits separate baseline and reporting period regression models of whole-home energy consumption for individual participants and nonparticipants. The models explain average daily energy consumption for customer *i* in month *t* (E_{it}) as a function of a constant (which represents the customer daily baseload consumption α_i), average daily heating degrees (HDD_{it}), average daily cooling degrees (CDD_{it}), or both HDD and CDD, depending on the fuel being modeled:

$$E_{it} = \alpha_i + \beta_{i1}HDD_{it} + \beta_{i2}CDD_{it} + \varepsilon_{it} \quad (2)$$

where ε_{it} is the model error and the coefficients α_i, β_{i1}, and β_{i2} are the parameters to be estimated for customer *i* in the baseline period or reporting period.

Using data for a normal weather year, the evaluator uses the fitted models to predict each customer's normalized annual consumption for the baseline and reporting periods and to calculate the difference between periods. This change in the customer's normalized energy consumption reflects the effects of the smart thermostat program; time-varying, nonprogrammatic factors during the reporting period for participants; and the effects of time-varying, nonprogrammatic factors for the matched comparison or randomized control group.

In the second stage of the estimation, the evaluator runs a cross-sectional ordinary least squares regression of the customer's change in normalized annual consumption on an intercept and an indicator variable for whether the customer was a smart thermostat program participant. The coefficient on the indicator variable is an estimate of the program's impact on normal weather-year average daily energy savings.

This two-stage approach is versatile and can be implemented with different research designs, including RCTs, REDs, the future participant approach, and matched comparison groups. There are many technical details to be mindful of in implementing the two-stage approach, including the selection of degree day base temperatures, allowing degree day base temperatures to vary between customers and between the pre- and post-periods, sample selection (including the removal of outliers), and steps for addressing imprecisely estimated first-stage regression coefficients. Evaluators should consult UMP Chapter 8 (Agnew and Goldberg 2017) for details about these considerations.

Two-Way Fixed-Effects Panel Regression

The second approach is a standard two-way fixed-effects D-in-D panel regression model of average daily energy consumption:

$$E_{it} = \alpha_i + \tau_t + \beta_0 DD_{it} + \beta_1 DD_{it} * Part_i + \gamma_0 DD_{it} * Post_t + \gamma_1 Part_i * Post_t + \gamma_2 Part_i * Post_t * DD_{it} + \varepsilon_{it} \quad (3)$$

where:

- E_{it} = Average energy consumption of utility customer *i* in month *t*
- α_i = Customer fixed effect to control for time-invariant differences between customers in energy consumption
- τ_t = Time period fixed effect
- DD_{it} = Degree days for customer *i* in month *t*; depending on the fuel and the seasons being analyzed, this model might include cooling degree days, heating degree days, or both variables[35]
- $Part_i$ = Indicator for smart thermostat program participant—this variable equals 1 if the customer was a participant and equals 0 otherwise
- $Post_t$ = Indicator for the reporting period after the thermostat was installed—this variable equals 1 if the period *t* was post-installation and equals 0 otherwise
- ε_{it} = Error term for customer in *i* in period *t*

This two-way fixed-effects panel regression controls for time-invariant differences between customers in their energy consumption (through the customer fixed effect), time period-specific consumption impacts unrelated to weather (through the month-year of sample fixed effect), and weather (through the cooling and heating degree variables). As the two-way fixed-effects model includes separate intercepts for each customer and separate degree day variables for the treatment and control groups and the pretreatment and post-treatment periods, the specification can flexibly model customer consumption and has many parallels to the two-stage approach.

This model can be estimated by ordinary least squares using data on customer energy consumption before and after the thermostat installation for program participants and nonparticipants in the comparison or control group. When estimated by ordinary least squares, the fixed-effects panel model will yield an unbiased estimate of the program savings if the error term is uncorrelated with $Part_i * Post_t$ conditional on the other model variables. This assumption will be satisfied if program participation was determined through a randomized controlled procedure (such as an RCT) or in quasi-experiments if the parallel-trends assumption holds: absent the installation of the smart thermostats and conditional on time-period fixed effects, customer fixed effects, and heating and cooling degree variables, participant and nonparticipant consumption would have followed the same reporting period time trends.

In the two-way fixed effects model, the energy savings for participant customer *i* in post-installation period *t* are equal to $-1 * (\gamma_1 + \gamma_2 DD_{it})$. The coefficient γ_1 represents the energy consumption impact that does not depend on cooling or heating degrees. The coefficient γ_2

[35] Within a month, degree days would vary between customers because of geographic variation in outdoor temperature or differences between customers in the base heating or cooling temperature.

represents the smart thermostat daily energy impacts per cooling or heating degree. For a smart thermostat program, it is expected that most savings will be temperature-related and the coefficient γ_1 will be close to zero. Nonetheless, it is advisable to include the intercept to mitigate any nonlinearity in the relationship between the smart thermostat savings and degree days. The annual savings for a normal weather year or the reporting period can be estimated as -1 * (365.25 * $\gamma_0 + \gamma_1 \text{AnnualDD}_{it}$), where $AnnualDD_{it}$ is the degree days for a normal weather year or for the reporting period.

As with the two-stage model, there are many technical details to keep in mind, and evaluators should consult UMP Chapter 8 (Agnew and Goldberg 2017) to learn more about these details.

Lagged Dependent Variable Panel Model

The lagged dependent variable (LDV) panel model (sometimes referred to as a "post-only model") is a panel regression model of *reporting-period* energy consumption. It is estimated with observations of customer average daily consumption or daily energy consumption during the reporting period for smart thermostat participants and matched nonparticipants. The model gets its name from the inclusion of a lag of the dependent variable—the customer's energy consumption for the same interval during the baseline period—as an explanatory variable.

A typical LDV panel model specification is as follows:

$$E_{it} = \tau_t + \beta_1 * \text{Part}_i + \rho \overline{E_{it}^{pre}} + \varepsilon_{it} \quad (4)$$

where:

E_{it} = Average energy consumption of utility customer *i* in month *t*

τ_t = The time-period fixed effect affecting the consumption of all subjects during month-year *t*; the month-by-year fixed effect can be estimated by including a separate dummy variable for each month-year *t*

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

Part_i = An indicator variable for whether customer *i* participated in the smart thermostat program; the variable equals 1 for participants and equals 0 otherwise

ρ = Coefficient indicating the average effect of consumption during the same interval of the baseline period

$\overline{E_{it}^{pre}}$ = Average consumption during the corresponding interval of the baseline period for customer *i*; for example, if the dependent variable was a customer's average daily consumption in July during the reporting period, $\overline{E_{it}^{pre}}$ would equal the customer's average daily consumption for July in the baseline period

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

Evaluators can estimate slightly different versions of the LDV model by including a participation indicator variable for each interval of the reporting period instead of a single participation indicator variable for the entire reporting period. This specification will produce an estimate of average savings per subject for each interval. Also, evaluators can add other variables to the model, such as weather.

A limitation of the LDV model is that it is generally inappropriate for weather-normalizing savings; that is, for calculating the savings that would occur in a normal weather year. This is because the coefficient on the lag of the dependent variable picks up some weather effects.

3.1.4.3 Detecting Self-Selection Bias and Diagnosing Its Causes

Evaluators should examine the results of the two-stage approach and the two-way fixed-effects panel regression approach for evidence of model misspecification or omitted variables. Specifically, it is highly recommended that smart thermostat program evaluators do the following:

- Test for differences in trend consumption between participants and future participants or the matched comparison group if two or more years of baseline period data are available. This step can also be performed when assessing the match quality. Using baseline period data for two years, run a two-way fixed-effects regression model of customer monthly energy consumption on customer fixed effects, time-period fixed effects, customer heating degrees, customer cooling degrees, and time-period fixed effects interacted with an indicator for whether the customer is a program participant during the reporting period. Plot the estimated coefficients on the interaction variables (time-period interacted with participant indicator) against time to look for evidence of differences in trend consumption. If the per-period trend difference is large relative to the per-period expected smart thermostat program savings, this trend would likely bias the savings estimate. However, this test would only capture savings estimation bias from factors present before the adoption of the smart thermostat; it would not capture any bias from differences in adoption of other energy-intensive devices such as electric vehicles occurring after the thermostat is installed.
- Test for differences between participants and matched nonparticipants regarding changes in the baseload (non-weather-sensitive) energy consumption between the reporting and baseline periods. Many participants and matched nonparticipants will exhibit large negative or positive changes in baseload consumption between the baseline and reporting periods. These changes could reflect their adoption of new appliances or other household durables, permanent changes in household occupancy, or other changes in energy consumption behaviors. However, if the matched nonparticipant group provides an accurate counterfactual for participants, there should not be large differences between the groups in terms of the mean or the distribution of baseload consumption changes (as smart thermostats primarily affect heating and cooling energy consumption loads). Differences in baseload consumption changes between the groups would suggest that one group is adding or reducing baseload consumption more than the other group. Evaluators should test for differences in the mean change in baseload consumption and the distribution of the change in baseload consumption. In the two-stage approach Equation 2, the change in baseload consumption for an individual customer equals $\alpha_{i,Reporting}$ -

$\alpha_{i,Baseline}$. Evaluators should assess the magnitudes of and conduct statistical tests for differences in the mean and distributions of baseload consumption changes between participants and matched nonparticipants. In the two-way fixed-effects regression model, the coefficient γ_1 represents the energy consumption impact that does not depend on weather. If the coefficient γ_1 is large (whether negative or positive) relative to baseload energy consumption, this would suggest a difference in the change in baseload consumption between participants and the matched comparison group.

If there is evidence of significant differences in trend consumption or changes in baseload consumption between participants and matched nonparticipants, the regression model may be mis-specified and the savings estimate may be biased. In such a situation, evaluators should follow two steps:

1. Assess the significance (in terms of magnitude, not just statistical significance) of the difference relative to the expected smart thermostat savings and the potential for bias. If the potential for bias is small, report the assessment results and the estimated savings.
2. If the potential for bias is high, attempt to diagnose the cause. Many smart thermostat program impact evaluations conduct participant surveys. This protocol recommends surveying participants and matched nonparticipants about their recent durable equipment purchases including electric vehicles, changes in household occupancy, and demographics and comparing the responses. This comparison may help the evaluators to diagnose the cause of the difference and to identify potential remedies.

If the potential for bias is high, evaluators have several options:

- Conduct the matching again, attempting to correct for the source of the trend or baseline consumption change differences by incorporating additional co-variates in the matching, and re-estimate the smart thermostat program savings.
- Retain the matched comparison group but adjust the smart thermostat replacement energy savings for the bias from the trend or baseload consumption change differences. For an example of such an adjustment see DNV-GL (2020, p. 44) or Guidehouse (2020a).[36]
- Retain the matched comparison group but include additional variables in the regression to attempt to correct for the source of the trend or baseline consumption change differences.
- If it is not possible to improve the match quality or adjust the savings on the backend, retain the savings estimate but clarify the potential for bias in the savings estimate, and to the extent possible, state the magnitude of the potential bias so that program administrators and policymakers can factor this information into their planning and policy decisions.[37]

[36] DNV-GL (2020) adjusted the smart thermostat savings for three factors: the difference in trend consumption between participants and matched nonparticipants; the fact that not all customers have space heating and space cooling; and the fact that some matched nonparticipants already have smart thermostats. Adjusting the smart thermostat savings for differences in trend or baseload consumption may correct for bias related to non-weather-sensitive drivers of consumption (e.g., adoption of electric vehicles). Other sources of bias may exist in the measurement of differences between participants and matched nonparticipants in weather-sensitive loads.
[37] Guidehouse (2020a) and DNV-GL (2020) are examples of evaluations that are transparent about the potential causes of bias and that adjust the regression-based savings estimates to reduce the potential for bias.

3.1.4.4 Daily or Hourly Energy Consumption Models

The presentations of the two-stage approach and the two-way fixed-effects panel regression model were based on the availability of monthly billing consumption data. However, with some modifications, both modeling approaches may be adapted for use with daily or hourly energy consumption data.[38]

When working with daily or hourly consumption data, evaluators have greater flexibility with their model specifications.[39] The most important consideration is to specify a model (or models) that capture the relevant variation in daily or hourly energy consumption. With daily energy consumption data, evaluators may allow consumption to depend on day of the week. With hourly energy consumption, evaluators may allow consumption to depend on hour of the day and day of the week. Within the contexts of the two-stage, two-way fixed-effects panel regression and LDV panel models, evaluators can estimate separate models for each day of the week or each hour of the day.[40] Alternatively, evaluators can expand the models to include interaction terms that allow the effects of baseload consumption and weather to depend on day of the week or hour of the day.

3.2 Smart Thermostat Optimization Programs

Smart thermostat optimization programs optimize cooling and/or heating schedules to produce incremental energy savings beyond those achieved by the base programming in the smart thermostat. Smart thermostat optimization can also achieve energy savings, load shifting, and bill savings for customers who are on a time-of-use or dynamic rate. The programs use software algorithms to optimize cooling and/or heating schedules through a series of very small adjustments to scheduled set points. Several smart thermostat vendors now provide optimization algorithms to their customers for free on an opt-in basis. Working with vendors, utility program administrators also can offer the programs to their customers. Most often, utility customers receive an offer to participate in the optimization program on the thermostat display or the thermostat app and must opt in to participate.

3.2.1 Evaluation Approaches

This protocol recommends that evaluators work with vendors to implement REDs or RCTs to estimate savings from smart thermostat optimization programs. The optimization savings is the difference in consumption between the smart thermostat with and without the optimization. Implementing randomized experiments to estimate optimization savings is usually straightforward because the optimization algorithms can be switched on and off remotely and at random by the thermostat manufacturer or service provider.

---

[38] If high-frequency data are unavailable, consult the UMP Chapter 10 (Stern and Spencer 2016) for alternative methods that combine an analysis of monthly energy consumption with engineering calculations or end-use load shapes. In general, these alternative methods require stronger assumptions and may not be as accurate as methods that analyze high-frequency consumption or runtime data.

[39] More research is needed about whether matching on hourly or daily data produces significant differences in results than matching on monthly data. While including aspects of hourly data in the matching process theoretically improves match quality with respect to loads across hours of the day, peak loads, and other aspects of hourly loads, it often proves difficult to attain tight matches using only hourly loads due to the noisiness of hourly data. Matching on daily or weekly data may produce better results.

[40] See DNV-GL (2020) for an example.

Table 3 lists the recommended evaluation approaches for smart thermostat optimization programs from most highly recommended to least recommended. Note that two separate approaches for RCTs are provided: an opt-out design and an opt-in design. In all approaches, program nonparticipants (who own smart thermostats) provide the baseline consumption for estimating the optimization impacts.

Table 3. Smart Thermostat Optimization Evaluation Design Options

Approach	Description	Advantages	Potential Challenges	Gross or Net Savings
RED: Utility randomly varies who receives encouragement to participate in the optimization program	• Eligible customers are randomly assigned to receive encouragement to enroll in the optimization program (encouragement group) or not to receive encouragement (control group) • Any encouragement group or control group customer can participate in the program	• Depending on whether control group customers can participate, RED yields an unbiased estimate of average savings per customer for all treated customers (control group participation not allowed) or for all customers who receive encouragement and opt in due to the encouragement (compliers with the encouragement) • All interested customers can participate—no need to delay or deny participation	• Insufficient compliance with the encouragement: not enough encouraged customers opt into the program • Large sample sizes required to obtain precise savings estimate	Net savings for all treated customers or compliers with the encouragement, depending on whether control group customers can opt in
RCT: Utility randomly varies who receives treatment or on which days treatment is given	**Opt-out:** The utility randomly assigns eligible customers to receive the optimization (treatment group) or not to receive treatment (control group). The utility can alternate the groups between receiving treatment and serving as the control	**Opt-out:** Yields unbiased estimate of intent-to-treat treatment effect as some auto-enrolled customers will unenroll from the program; controls for self-selection in participation	**Opt-out:** Potential customer dissatisfaction from auto-enrollment	**Opt-out:** Net savings per treatment group customer (intent-to-treat treatment effect)
	Opt-in: Customers self-enroll in optimization program and the utility either (a) randomly varies who receives treatment on a given day; or (b) randomly varies the days the optimization treatment is applied to all or some customers	**Opt-in:** All eligible and interested customers can participate.	**Opt-in:** Insufficient overlap in distributions of weather or other time-varying factors affecting consumption between treatment and control days; also, spillover of impacts from event to nonevent days can confound savings estimates	**Opt-out:** Net savings per treated customer
Matched comparison group	Opt-in: Comparison group of nonparticipants matched on either energy consumption and other characteristics or only on energy consumption	• Unnecessary to set up randomized experiment at the program start • Less susceptible to selection bias than replacement program evaluation as customers in matched comparison group have smart thermostats	Potential for bias from self-selection still exists	Net savings per enrolled customer

Following are additional descriptions of these approaches:

- ***RED experimental design.*** An RED for an optimization program would require randomly selecting eligible customers with smart thermostats for the encouragement group, who would be encouraged to participate in the optimization program. Customers who are randomly selected for the control group would not receive the offer and provide the baseline for measuring the savings for the encouragement group. Some treatment group customers receiving the offer who would not otherwise enroll will accept (these are referred to as compliers). The RED is expected to produce an unbiased estimate of the net savings for all treated customers (the treatment effect for the treated) through an instrumental variables two-step estimation procedure if none of the customers in the control group receive the optimization or for compliers (the local average treatment effect) if control group customers can enroll in the optimization program.[41] To obtain a precise savings estimate, REDs usually require large treatment and control groups and sufficiently large rates of compliance with the encouragement. See the UMP Chapter 8 (Agnew and Goldberg 2017) and UMP Chapter 17 (Stewart and Todd 2020) for additional details about implementing REDs and see Guidehouse (2019, 2020b) and Blonz et al. (2021) for examples of an RED smart thermostat optimization evaluation.
- ***RCT experimental design.*** RCTs vary who receives treatment or on which days treatment is given and can be implemented in two main ways, depending on whether customers self-enroll or the utility auto-enrolls them:
    - In the first approach, randomly selected customers are defaulted into the optimization program while retaining the ability to opt out. Customers are randomly assigned to two or more groups. To estimate savings, one or more groups receive the optimization intervention, and their consumption is compared to that of a control group who does not receive the optimization.[42]
    - In the second approach, interested customers self-enroll in the optimization program and the optimization treatment is applied to either (a) randomly chosen customers; or (b) on randomly chosen days. In (a), the evaluator randomly assigns enrollees to two or more groups so the groups have similar mixes of geography, consumption, and demographics. Then the evaluator can generate a random sequence of testing periods (multiple consecutive days) during which one group would not receive the optimization and the remaining groups would. For the next period, a different group would serve as the control and the remaining groups would receive treatment, and this procedure would be repeated. In (b), the evaluator uses days when the thermostat optimization is not operational to establish the baseline. For example, a thermostat vendor could probabilistically assign customers to receive or not receive optimization for several consecutive days during the cooling or heating season. In this control-day approach,

[41] While control group customers would not receive the encouragement, the program administrator could allow interested control group customers to participate. In this situation, the RED would yield an estimate of the net energy savings for compliers (customers who participated due to the encouragement).
[42] Customers who opt out of treatment should be retained in the analysis sample for the duration of the experiment. The savings estimate from this RCT would be an intent-to-treat treatment effect (savings) unless the evaluator adjusts the savings estimates to account for opt-outs.

evaluators must ensure that the range of weather is equivalent on optimization and nonoptimization days or use regression analysis to adjust for any differences in weather conditions between optimization and nonoptimization days to minimize potential for bias in the savings estimates. Also, evaluators must rely on data at a daily or more granular level.[43]

- ***Matched comparison group.*** In this approach, evaluators match optimization program participants to smart thermostat customers who do not participate in the optimization program based on baseline period energy consumption and possibly customer demographic variables. Because the matched comparison group only comprises smart thermostat customers, the potential for bias from self-selection is lower for optimization programs than for replacement programs. The thermostat optimization program could be implemented on an opt-in or opt-out basis. Making participation the default option while allowing participants to opt out would reduce self-selection in participation due to customer tendencies to adhere to the status quo (Fowlie et al. 2021) and make it easier for the evaluator to construct a valid comparison group.

In addition, through a statistical power analysis, evaluators should verify that the planned sample sizes for the treatment and control or comparison groups are large enough to detect the expected savings given the unexplained random variation in the energy consumption or vendor telemetry data.[44]

3.2.2 Smart Thermostat Telemetry Data and Whole-Home Consumption Data

To estimate optimization program savings, evaluators can analyze whole-home consumption data from the utility or thermostat runtime telemetry data from the thermostat vendor. This protocol recommends analyzing whole-home hourly or daily consumption data when such data are available because they will account for all optimization program-induced energy impacts, including the use of other appliances (such as fans) to control the home temperature, changes in refrigerator runtimes, and other secondary effects of the optimization. In addition, in contrast to telemetry runtime data analysis, home energy consumption data analysis does not require converting smart thermostat HVAC runtime impacts to energy. Both features mean that home energy consumption data analysis is likely to yield more accurate savings estimates than telemetry runtime data analysis.

However, utility meter data may be unavailable, or the probability of detecting the expected smart thermostat program optimization savings in the meter data may be low. Evaluators of smart thermostat optimization programs should analyze thermostat runtime when one or more of the following conditions are met:

- Optimization program participants and nonparticipants do not have AMI meters, or the utility is otherwise unable to provide interval data from AMI meters.

[43] For both approaches, evaluators should be aware of the potential for changes in thermostat-setting behaviors by participants who become aware of the treatment. For example, participants may become habituated to more energy efficient temperatures and set new, more efficient thermostat set points.
[44] More information is available in UMP Chapter 8 (Agnew and Goldberg 2017).

- Due to customer privacy protections and policies, smart thermostat vendors are unable to reveal the identities of optimization program participants and nonparticipants, preventing evaluators from linking smart thermostat customers to utility meter data.
- The optimization program analysis sample is not large enough to detect the expected optimization energy impacts in whole-home AMI meter data. The consumption data may contain too much noise to pick up the optimization energy impacts. The probability of detecting the savings through statistical analysis may be higher with HVAC runtime data.

Because thermostat telemetry data are usually available for the periods before, during, and after the optimization for both participants and nonparticipants, evaluators can construct accurate baseline HVAC runtimes and estimate the optimization savings as a D-in-D. As an example of a telemetry data analysis, in Massachusetts in summer 2019, several energy efficiency program administrators and a thermostat vendor implemented an RED to test a thermostat optimization program. The evaluator analyzed the impacts of the optimization on thermostat runtimes and HVAC electricity consumption using vendor-supplied telemetry data for customers randomly assigned to the RED encouragement and control groups and a D-in-D fixed effects panel regression (Guidehouse 2020b).

3.2.3 Savings Estimation

3.2.3.1 Whole-Home Consumption Analysis

When estimating optimization program savings by analyzing whole-home consumption data, evaluators should specify a regression model that matches the time granularity of the data and the desired granularity of the savings estimate(s), such as hourly or daily. In addition, evaluators will need to specify a regression model whose coefficients will measure the difference in energy consumption between optimization participants and nonparticipants or between optimization days and nonoptimization days depending on the research design. Evaluators can implement a variant of the two-stage approach, the two-way fixed-effects panel regression, or the LDV panel regression to estimate the optimization savings. These approaches are described in Section 3.1.4.2.

3.2.3.2 Thermostat Telemetry Analysis

Evaluators will need to collect thermostat telemetry data from smart thermostat vendors. If the telemetry data are anonymized, evaluators should start by verifying that the runtimes are associated with smart thermostats within the utility's service territory. Evaluators can compare the zip code locations of customers in the anonymized telemetry data with the utility's or program administrator's service area to confirm the optimization participant receives service from the utility.

Next, the evaluator should select the runtime data based on the season under study (i.e., the winter heating months or summer cooling months). Most telemetry data sets include a field indicating if the home HVAC system is set to off, heating mode, or cooling mode. In the case of programs targeting optimization of auxiliary heat, the analysis should use separate models of

primary and auxiliary heat because the running wattage for the two systems will be different.[45] If separate primary and auxiliary runtime data are unavailable, this protocol discourages the analysis of telemetry runtime data and instead encourages analysis of AMI meter data.[46]

The next step is to run appropriate model specifications to estimate the runtime or energy savings. Evaluators have the option of converting runtime data to energy before modeling energy consumption or modeling runtime and then converting the runtime impact estimate to energy savings. Approaches for calculating runtime-to-energy conversions are presented in the next section.

Conversion From Runtime to Energy Savings

Converting runtime impacts to energy savings requires multiplying runtime or the runtime reduction by an HVAC energy-consumption-per-unit-of-runtime factor.[47] This conversion will be accurate for most single-stage systems.[48] Also, for compressor-based systems, this runtime factor will be a function of outside temperature (Goldman et al. 2017). When converting runtimes to energy, it is best practice to use conversion factors applicable to the smart thermostat program population under study rather than conversion factors applicable to the general population.

Making runtime-to-energy conversions usually requires information about the heating and cooling systems of smart thermostat program participants, such as:

- Cooling equipment type and capacity
- Heating equipment type and capacity
- Cooling and heating efficiency values
- Cooling and heating fuel sources.

The most widely used approach for obtaining runtime-to-energy conversion factors is to draw from region-specific technical reference manuals. In addition, engineering studies or primary

[45] An additional energy-savings source from smart thermostats is their more efficient use of backup/auxiliary heat from heat pump systems. Considering that heat pumps require longer recovery times to bring the house back to a comfort set point due to lower-temperature supply air, a thermostat will sense this delay, automatically turning on auxiliary heat to warm the house more rapidly. Smart thermostats can learn how long it takes for the house to recover from various setback conditions, then automatically adjust the setback amount to maximize the unit efficiency. However, some heat pumps will also sense the delay and switch to a less-efficient, higher-power mode when they do not reach the set point quickly enough. This complicates the interpretation of runtime impacts.

[46] Analysis of smart thermostat runtime data will capture the benefits of auxiliary heating optimization if that system is also controlled by the smart thermostat, but other secondary heat such as portable electric resistance heating that might offset optimization control of the primary heating will not be captured.

[47] Depending on the region, evaluators may have to consider multiple heating fuels. If possible, separate models should be run for thermostats that control natural gas, electric, or oil heat, and for whether the thermostats are controlling furnaces, heat pumps, or boilers. If there is no way to accurately assign thermostats to a given fuel type and equipment type, evaluators can estimate one runtime savings value (or percentage savings) and apply it to the best estimate of the proportions and usage by fuel type and equipment type in the region.

[48] This approach may not work for hydronic systems using water or water-based solutions for heat transfer. Hydronic systems often do not run the burner the whole time a thermostat calls for heat but cycle the unit on based on the hydrostat in the boiler reservoir.

engineering study data (such as baseline studies or participant surveys) may also be used. See Guidehouse (2020b) for an example of converting runtime to power based on analysis of metering data from a baseline study.

Evaluators can also collect nameplate information by visiting a sample of participant sites or asking participants to self-report. See Goldman et al. (2017) for an example of the former approach and how HVAC nameplate data may be used to calculate power. See Guidehouse (2022) for an example of the latter approach in which residential utility customers submitted photographs of home HVAC nameplate information. These approaches yield information about HVAC equipment specific to the program population being evaluated and may be most useful when the smart thermostat program population differs from the average residential utility customer as represented in the TRM. However, if site visits or self-reporting are employed, evaluators should take steps to minimize any bias from self-selection related to who participates in a site visit or self-reports HVAC information.

Evaluators should be mindful of the increasing prevalence of two-stage or three-stage HVAC systems, which will draw different amounts of electric power or natural gas depending on the system's level of operation and will complicate the conversion from runtime to kilowatt-hour or therm impacts. Evaluators will need separate conversion factors for the runtimes for the three stages. A useful reference for making runtime-to-energy conversions for multi-stage systems is Cutler et al. (2013).[49]

[49] Most variable speed HVAC systems including those with more than three heating or cooling capacities are incompatible with smart thermostats.

4 Net-to-Gross Considerations

Consult the UMP Chapter 21 (Violette and Rathbun 2017) for a discussion about determining net program impacts at a general level, including direction on how to assess freeridership. Additional net-to-gross discussion and consideration based on the program conditions are available in Section 3.1.2.2 of this protocol and UMP Chapter 8 (Agnew and Goldberg 2017).

5 Other Evaluation Issues

5.1 Data Availability, Accessibility, and Security

Telemetry data analysis can provide useful insights about the energy impacts of smart thermostats. For example, if AMI data show an increase in energy consumption during certain hours, and thermostat runtime data shows that the HVAC system was not running during the same hours, an evaluator can infer that the thermostat did not contribute to the energy consumption increase in the home during these hours. In addition, thermostat telemetry data may be the only data available to evaluators of optimization programs.

Due to the potential value of thermostat telemetry data, evaluators, program administrators, and smart thermostat vendors should continue to work together to find solutions to several issues related to telemetry data availability, accessibility, and security[50]:

- Vendors have valid concerns about providing smart thermostat data to third-party entities and use various data security and privacy steps (such as anonymization or aggregation of the data) to protect customer data from unauthorized access, destruction, use, modification, or disclosure. At the same time, utility program administrators have obligations to be responsible stewards of rate payer funds and to undertake rigorous, transparent, and replicable evaluations. Steps to protect privacy and ensure security can make it difficult for evaluators to establish the source, completeness, and quality of the telemetry data and to link the telemetry data to program participants.
- To receive customer identifying information in the telemetry data such as name, street address, and utility account number, utilities or evaluators often must obtain permissions for such information from individual customers. Evaluators may not receive all needed permissions or the customers giving permission may not be representative of the population, providing analysis results that are not statistically significant or externally valid. There may be ways for evaluators and vendors to increase the number and representativeness of customers opting to share their thermostat data.
- Even if explicit customer authorizations are obtained, evaluators will need to work with vendors to obtain the data. This may involve verifying with the thermostat vendor that customer authorization procedures will be acceptable to facilitate the data release. In general, it is prudent to have explicit support for customer telemetry data sharing referenced in the incentive agreement between the utility (or program administrator) and each vendor. Also, evaluators should test the data access process on a few sample thermostats early in the process to ensure that all necessary data fields will be available and each customer's data can be matched to program participant records, if necessary. It may also be necessary for the program administrator to negotiate data access requirements as part of the incentive agreement (e.g., as part of the request for proposal for qualified products), and the vendor may charge an additional fee for data access infrastructure (e.g., APIs, dashboards) and technical support.

[50] Apex Analytics (2021) discusses many of these issues and impediments to using telemetry data for the purposes of evaluation.

5.2 ENERGY STAR Connected Thermostat Metric

The U.S. Environmental Protection Agency's (EPA) *ENERGY STAR® Connected Thermostat Method to Demonstrate Field Savings* was originally designed to certify under the ENERGY STAR program that smart thermostats were capable of delivering significant energy savings to most households. ENERGY STAR primarily offers a metric of the relative performance of different thermostats compared with a common baseline (of maintaining a constant indoor temperature) and does not allow for or otherwise incorporate different or varied behaviors prior to installing a smart thermostat. The metric was not intended for use in evaluating energy savings for smart thermostat programs.

The ENERGY STAR metric, method, and specification are expected to be updated every three to four years, as is typical for ENERGY STAR specifications. A strong focus of recent EPA research has been to improve the score to be a more reliable indicator of achieved savings. The EPA would like to understand if the ENERGY STAR metric or a modified version of the metric could be used to estimate energy savings.[51]

Two recent studies implemented modified versions of the ENERGY STAR metric and compared the resulting savings to savings estimated with whole-home consumption data (Guidehouse 2020a; Apex Analytics 2021). In both studies, the modifications primarily involved the use of more realistic assumptions about the baseline thermostat setting behavior of smart thermostat adopters before the smart thermostats were installed. Guidehouse (2020a) found nontrivial differences in the estimated percentage cooling savings between the whole-home consumption analysis (7.8%) and the ENERGY STAR metric (10% to 14%).[52] The study's main conclusion was that the estimated savings were sensitive to the assumption of a household's preferred comfort temperature prior to installing their smart thermostat. The main objective of the Apex Analytics (2021) study was to validate the adjusted ENERGY STAR metric as an indicator of energy savings, but the authors found very weak or no correlation between the ENERGY STAR metric savings and savings from site-level meter monthly consumption data analysis. The study "could not establish a method to use thermostat-derived metrics to estimate these energy savings with sufficient reliability for use by Northwest utilities." The Northwest study's finding of a weak correlation is not unexpected because the ENERGY STAR metric is a measure of the efficiency of a home's thermostat set points, not an indicator of energy savings from the smart thermostat, and site-level estimates of energy savings from the meter data analysis were noisy.

The EPA and other researchers are continuing to refine the ENERGY STAR metric with the hope that in the future it may support smart thermostat program evaluation. The ENERGY STAR metric can complement smart thermostat program evaluations that analyze monthly billing consumption or AMI meter data. For now, however, evaluators should not rely on the metric as their primary evaluation method.

[51] The EPA believes that incorporating locally appropriate baseline temperature data, collected through data logging of actual indoor temperatures in representative homes, into an ENERGY STAR metric will improve the accuracy of the savings estimates of the ENERGY STAR method.

[52] This 10% to 14% represents the range of savings associated with assumed baseline behavior agreed to by the study stakeholders. The baseline assumptions were not verified.

References

Agnew, Ken, and Mimi Goldberg. 2017. "Chapter 8: Whole-Building Retrofit with Consumption Data Analysis Evaluation Protocol." In *The Uniform Methods Project: Methods for Determining Energy Efficiency Savings for Specific Measures.* Golden, CO: National Renewable Energy Laboratory. NREL/SR-7A40-68564. http://www.nrel.gov/docs/fy17osti/68564.pdf

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

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

Apex Analytics and Empower DataWorks. November 16, 2021. *Northwest Smart Thermostat Research Study*. Prepared for the Northwest Energy Efficiency Alliance. Report #E21-324. https://neea.org/resources/northwest-smart-thermostat-research-study

Blonz, Joshua, Karen Palmer, Casey J. Wichman, and Derek C. Wietelman. July 21, 2021. *Smart Thermostats, Automation, and Time-Varying Prices*. Resources for the Future. https://www.rff.org/publications/working-papers/smart-thermostats-automation-and-time-varying-prices/

Brandon, Alec, Christopher M. Clapp, John A. List, Robert D. Metcalfe, and Michael K. Price. 2021. *Smart Tech, Dumb Humans: The Perils of Scaling Household Technologies*. http://dx.doi.org/10.2139/ssrn.3961130

CalTRACK. Last updated 2018. "CalTRACK Methods Version 2.0." Accessed April 4, 2023. https://docs.caltrack.org/en/latest/methods.html

Cutler, D., J. Winkler, N. Kruis, C. Christensen, and M. Brandemuehl. 2013. *Improved Modeling of Residential Air Conditioners and Heat Pumps for Energy Calculations.* Golden, CO: National Renewable Energy Laboratory. NREL/TP-5500-56354. https://www.nrel.gov/docs/fy13osti/56354.pdf

DNV-GL. August 7, 2015. *2014 Web-Enabled Thermostat Impact Evaluation.* Prepared for Puget Sound Energy.

DNV-GL (Goldberg, M. L., M. Fowlie, K. Train, and G. K. Agnew). 2017. *A White Paper: Mitigating Self-Selection Bias in Billing Analysis for Impact Evaluation*. Submitted to Pacific Gas and Electric.

DNV-GL. 2020. *Impact Evaluation of Smart Thermostats: Residential Sector – Program Year 2018.* Prepared for California Public Utilities Commission. CALMAC ID: CPU0205.01. http://www.calmac.org/publications/CPUC_Group_A_Report_Smart_Thermostat_PY_2018_CALMAC.pdf

DNV-GL. 2021. *Impact Evaluation of Smart Thermostats: Residential Sector – Program Year 2019.* Prepared for California Public Utilities Commission. CALMAC ID: CPU0232.01. http://www.calmac.org/publications/ https://www.calmac.org/publications/CPUC_Group_A_Residential_PY2019_SCT_Final_Report_CALMAC.pdf

DNV-GL. 2022. *Impact Evaluation of Residential HVAC Measures Residential Sector – Program Year 2020.* Prepared for California Public Utilities Commission. CALMAC ID: x. https://pda.energydataweb.com/api/view/2620/Group%20A%20Residential%20PY2020_RES%20HVAC%20Final%20Report_forPDA.pdf

Fowlie, Meredith, Catherine Wolfram, Patric Baylis, C. Anna Spurlock, Annika Todd-Blick, and Peter Cappers. 2021. "Default Effects and Follow-on Behavior: Evidence from an Electricity Pricing Program." *The Review of Economic Studies* 88 (6): 2886-2934.

Goldberg, Miriam L., and G. Kennedy Agnew. 2013. *Measurement and Verification for Demand Response*. Prepared for the National Forum on the National Action Plan on Demand Response: Measurement and Verification Working Group. https://www.ferc.gov/sites/default/files/2020-04/napdr-mv.pdf

Goldberg, M. L., M. Fowlie, K. Train, and G. K. Agnew. 2017. "Not Just Another Pretty Formula: Practical Methods for Mitigating Self-Selection Bias in Billing Analysis Regressions." *Proceedings of the International Energy Program Evaluation Conference*, Baltimore, Maryland.

Goldman, Ethan, Abiodun Iwayemi, Jennifer Robinson, Ram Narayanamurthy, Ben Clarin, Robert Ruskamp, and Marc Shkolnick. 2017. *Measuring Demand Savings with Smart Thermostat Data*. Prepared for the International Energy Program Evaluation Conference, Baltimore, Maryland.

Guidehouse. 2018. *ComEd Advanced Thermostat Evaluation Research Report.* Prepared for Commonwealth Edison Company. https://icc.illinois.gov/downloads/public/edocket/487373.PDF

Guidehouse. 2019. *ComEd CY2018 Nest Seasonal Savings Heating Season Impact Evaluation Report*. Presented to Commonwealth Edison Company. https://icc.illinois.gov/docket/P2020-0475/documents/300880/files/524633.pdf

Guidehouse. 2020a. *ComEd Advanced Thermostat Evaluation Report: Final Research Report.* Prepared for Commonwealth Edison Company. https://www.ilsag.info/wp-content/uploads/ComEd-Adv-Thermostat-Research-Report-Final-2020-11-10.pdf

Guidehouse. 2020b. *2019 Massachusetts Summer Thermostat Optimization Evaluation*. Prepared for The Massachusetts Program Administrators. https://ma-eeac.org/wp-content/uploads/MA19R10-E-STO_Summer-TO-Evaluation-Final-Report-2020-03-26.pdf

Guidehouse. 2021. *Residential Wi-Fi and Programmable Thermostat Impacts Res24 Final Report.* Prepared for Massachusetts Program Administrators. https://ma-eeac.org/wp-content/uploads/MARES24-Final-Report-2021-09-29.pdf

Harding, Matthew, and Alice Hsiaw. 2014. "Goal Setting and Energy Conservation." *Journal of Economic Behavior & Organization, Elsevier*, vol. 107 (PA): 209–227.

Iacus, Stefano, Gary King, and Giuseppe Porro. 2011. "Causal Inference without Balance Checking: Coarsened Exact Matching." *Political Analysis* 20 (1): 1–24.

Imbens, Guido W. and Donald B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. New York, NY: Cambridge University Press.

International Performance Measurement and Verification Protocol (IPMVP). (2022). IPMVP Volume 1: Concepts and Options for Determining Energy and Water Savings. EVO 10000 - 1:2022. Washington, D.C.: Efficiency Valuation Organization. https://evo-world.org/en/products-services-mainmenu-en/protocols/ipmvp

Nexant. 2017. *Xcel Energy Colorado Smart Thermostat Pilot – Evaluation Report*. Prepared for Xcel Energy. https://www.xcelenergy.com/staticfiles/xe-responsive/Company/Rates%20&%20Regulations/Regulatory%20Filings/CO-Smart-Thermostat-Pilot-Evaluation.PDF

State and Local Energy Efficiency Action Network. 2012. *Evaluation, Measurement, and Verification (EM&V) of Residential Behavior-Based Energy Efficiency Programs: Issues and Recommendations.* Prepared by A. Todd, E. Stuart, S. Schiller, and C. Goldman. Berkeley, CA: Lawrence Berkeley National Laboratory. https://www.energy.gov/sites/default/files/2021-08/emv_behaviorbased_eeprograms.pdf

Stern, Frank, and Justin Spencer. 2016. "Chapter 10: Peak Demand and Time-Differentiated Energy Savings Cross-Cutting Protocol." In *The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures.* Golden, CO: National Renewable Energy Laboratory. NREL/ SR-7A40-68566. https://www.nrel.gov/docs/fy17osti/68566.pdf

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

U.S. Environmental Protection Agency. n.d. "Connected Thermostats Specification Version 1.0." Accessed April 4, 2023. https://www.energystar.gov/products/spec/connected_thermostats_specification_v1_0_pd

Violette, Daniel M., and Pamela Rathbun. 2017. "Chapter 21: Estimating Net Savings – Common Practices." In *The Uniform Methods Project: Methods for Determining Energy-Efficiency Savings for Specific Measures.* Golden, CO; National Renewable Energy Laboratory. NREL/ SR-7A40-68578. https://www.nrel.gov/docs/fy17osti/68578.pdf