

The Human Perils of Scaling Smart Thermostats: Evidence from Field Experiments

By ALEC BRANDON, CHRISTOPHER M. CLAPP, JOHN A. LIST, ROBERT D. METCALFE, MICHAEL K. PRICE*

We experimentally estimate the causal effect of smart thermostats on energy consumption with two field experiments where over 1,300 households are randomized into eligibility to receive a smart thermostat. Examining 18 months of high-frequency energy consumption data, we find little evidence that smart thermostats have statistically or economically significant effects on energy use. Using data on adopter interactions with their smart thermostat, we find that human behavior dampens energy savings and explains the discrepancy between estimates from engineering models, which make strong assumptions about how households prioritize energy conservation, and actual households, who make tradeoffs that result in comfortable homes. (JEL C93, Q48, Q55)

Due to the high private and social costs of energy production, there is substantial interest among policymakers in the United States (U.S.) and abroad in leveraging recent advances in information technology to conserve energy without reducing consumer welfare by increasing efficiency. Towards this goal, for example, the U.S. Department of Energy has allocated more than \$8 billion since 2007 through its Smart Grid Investment Grant program to support the development, testing, and installation of advanced metering and information, or smart, technologies (Joskow, 2012). More recently in the U.S., a similar set of smart technologies had billions in subsidies allocated in the Inflation Reduction Act of 2022 (Department of Treasury, 2024). Abroad, policymakers in the European Union and United Kingdom have similarly undertaken significant public investments.

In this article, we study the energy conservation effects of one of these technologies: smart thermostats. Like other information technologies, smart thermostats can communicate with other devices via an internet connection. This

* Brandon: Johns Hopkins University (alec.brandon@jhu.edu); Clapp: University of Chicago (cclapp@uchicago.edu); List: University of Chicago (jlist@uchicago.edu); Metcalfe: Columbia University (rdm2176@columbia.edu); Price: University of Alabama (mkprice2@ua.edu). The order of authors is alphabetical. We thank Dylan Brewer, Fiona Burlig, Ken Gillingham, Michael Greenstone, Koichiro Ito, David Jimenez-Gomez, Juanna Joensen, David Novgorodsky, Mar Reguant, Florian Rundhammer, Casey Wichman, and seminar participants at the Advances with Field Experiments Conference, Heartland Environmental and Resource Economics Workshop, Federal Communications Commission, EPIC Faculty Workshop, NBER Environmental and Energy Economics Summer Institute, and Harvard Seminar in Environmental Economics and Policy for helpful comments and suggestions. Lillian Bartholomew, Ariel Listo, David Liu, Udit Karna, Jack Ogle, and Juliana Lalinde provided excellent research assistance. Opower provided the data analyzed in this paper to the authors under a nondisclosure agreement. Research was conducted with an Institutional Review Board (IRB) exemption from Georgia State University.

ability to communicate allows smart thermostats to have features that manual and programmable thermostats cannot, such as remotely scheduling and overriding scheduled temperature setpoints in a home with a smartphone or computer. These types of features have proven popular with consumers. Between 2015 and 2020, adoption of smart thermostats more than tripled in the United States from 3.4% of households to 10.3% (Energy Information Administration, 2018, 2025). Also like other smart technologies, to promote their adoption, smart thermostats are heavily subsidized. For example, in 2019, 170 U.S. energy providers subsidized the purchase of smart thermostats (Environmental Protection Agency, 2019) and, as of 2026, there are more subsidy programs offered to U.S. consumers adopting smart thermostats than any other energy efficiency product (Environmental Protection Agency, 2026).

Energy providers provide these subsidies for smart thermostats because they expect that households that adopt them will conserve energy. Underlying this expectation is an engineering model of energy consumption, thermostat type, and the indoor temperatures that users set via their thermostat. This model assumes that a smart thermostat causes adopters to schedule less energy intensive temperature settings, or henceforth setpoints, when they are away from the home or when they are asleep (Department of Energy, 2020). These less energy intensive setpoints are referred to as setbacks and a rule of thumb from the engineering literature is that over an 8 hour period each degree Fahrenheit ($^{\circ}\text{F}$) setback reduces energy consumption by approximately 1% (Nelson and MacArthur, 1978). Therefore, for example, an 8 hour setback of 4 $^{\circ}\text{F}$ over the course of the day equates to an approximately 4% reduction in energy consumption.

While this predicted reduction in energy consumption relies on strong assumptions about the way that adopters will use their smart thermostats, the model it is based on set the prior for subsequent studies and appears to have had a persistent effect on the literature. Energy providers and thermostat producers tout impressive savings on heating and cooling costs on the basis of empirical analyses in industry reports and white papers. While their findings are often consistent with the long-held rule of thumb, many of these analyses fail to meet the standards for credibility in ways that upwardly bias estimates of energy savings (see Supplemental Materials Section A for a review of this literature). Ultimately, a real-world comparison of the energy consumed by otherwise similar households with and without a smart thermostat is needed to determine their true impact on energy use.

To do so, we experimentally estimate the effects of smart thermostat adoption on energy consumption in the field. Partnering with Opower, Honeywell, and Pacific Gas and Electric (PG&E), the second largest residential energy provider in the United States, we examine data from two framed field experiments, wherein the 1,385 households in Northern and Central California were randomized into either a treatment group that was eligible to receive free installation of a Honeywell smart thermostat or a control group that did not receive a smart thermostat. In

conjunction with the randomly assigned eligibility for smart thermostat adoption, we observe high-frequency energy consumption over an 18-month period that includes more than 16 million observations of hourly electricity use and almost 700 thousand daily observations of natural gas consumption.

Engineering models predict that adoption of the smart thermostat in our experiments will conserve significant quantities of energy. The most pertinent engineering prediction for policy come from annual reports produced by California’s energy providers and regulators. The predictions in these reports rely on simulations of energy use with and without a smart thermostat in different types of homes and in different types of climates. These predictions are then used by energy providers to justify subsidies for technologies that conserve energy. Mapping these predictions to our experimental samples, we find that savings of 1.3% and 4.0% are respectively predicted for overall electricity and natural gas consumption (California Municipal Utilities Association, 2017), which are consistent with the Nelson and MacArthur (1978) rule of thumb when households use an 8 hour setback of 4°F each day.

In contrast to these predictions, we find that smart thermostats fail to conserve energy. Using treatment status as an instrument for adoption, we find that the effect of the smart thermostat on the electricity and natural gas consumption of would be adopters (i.e., the average treatment effect on the treated or ATT) is statistically and economically indistinguishable from zero. For example, using a specification that pools the two experiments and includes household and time effects, as well as controls for weather, our estimates suggest that smart thermostats decrease average electricity consumption by 0.09% and increase average natural gas consumption by 1.70%. Across specifications, our estimates fall short of the engineering predictions that the smart thermostat respectively reduces electricity and natural gas consumption by 1.3% and 4.0%. While estimates that pool across the two experiments for electricity are not precise enough to reject the engineering prediction, all but one of our pooled estimates for natural gas reject the engineering prediction.

While the smart thermostat may still produce private benefits outside of energy conservation that explain their growing popularity (e.g., more comfortable temperatures in the home), our results suggest that energy conservation is a poor basis upon which to subsidize the adoption of smart thermostats. Furthermore, the failure of engineering estimates to accurately predict measured responses is broadly consistent with a growing body of research that documents real-world effects of energy efficient technology that pale in comparison to the effects predicted by engineers (Davis, Fuchs and Gertler, 2014; Levinson, 2016; Zivin and Novan, 2016; Houde and Aldy, 2017; Fowlie, Greenstone and Wolfram, 2018; Alpízar, Bernedo and Ferraro, 2024; Davis, Martinez and Taboada, 2020; Christensen et al., 2021).

Additionally, we find almost no evidence of significant heterogeneity in the effects of a smart thermostat. Across models estimated on different subsamples

that condition our data to specific days of the week, hours of the day, and types of weather, we consistently find null effects of the smart thermostat on energy consumption. The exceptions are during periods of high humidity, but not necessarily high temperatures, and during PG&E issued peak alerts, but not similar alerts issued by the operator of California's electrical grid. Nonetheless, the overall pattern across all our results indicates that smart thermostats under-deliver on the conservation effects predicted by engineers.

We next turn to mechanisms. Using almost four million observations of heating, ventilation, and air conditioning (HVAC) system activity in the treatment group homes that adopted the smart thermostat, three insights emerge related to setpoints, overrides, and setbacks. First, adopters quickly schedule setpoints after the smart thermostat is installed in their home, do so at times of the day indicating that they are scheduling setpoints as intended, and schedule temperatures that are consistent with energy conservation, although there is substantial heterogeneity in those temperature settings. Second, adopters frequently override their setpoints, with an average of nearly 1.7 overrides per day, and these overrides are typically in the direction of more energy intensive setpoints than the previously scheduled setpoint. Third, most adopters setback their heating setpoints by more than the 4 °F that engineers assume, however these heating setbacks typically last less than the 8 hours assumed by engineers. In contrast, cooling setbacks are typically just 1 to 2 °F, and there is substantial heterogeneity in setback behavior across the sample.

Based on these insights, we combine the setpoint, override, and setback information with our energy consumption data, and fit models designed to test whether we can replicate the engineering model predictions in real-world data by conditioning on households whose actions are consistent with the engineering model assumptions. We reproduce the engineering predictions for natural gas when the control group is compared to a treated group comprised of households that schedule efficient setpoints or schedule setbacks that last at least 8 hours. As these thermostat settings are post-treatment outcomes, we caution against interpreting this analysis as causal. Instead, this analysis illustrates the mechanism that causes the discrepancy between the existing literature and our experimental estimates: engineering models accurately predict the effectiveness of the smart thermostat when we omit adopters who do not use the smart thermostat as assumed. Collectively, these results suggest that engineering predictions fail because of households who adopt the smart thermostat to use its features in ways that undo the purported benefits, suggesting that human behavior is a peril to the real-world effects of smart technologies like the smart thermostat.

We view our results as speaking to several literatures and stakeholders. Regarding energy-efficiency policy, our results offer two insights. First, that the effects of energy-efficient technology predicted by engineers should be discounted when underlying the predictions are strong assumptions about how households will use the technology. In our case, engineering studies assume a smart ther-

mostat optimally programmed for energy savings and a traditional thermostat set to maintain a fixed temperature. These are strong assumptions because individuals make tradeoffs between their thermal comfort and the associated energy costs of achieving that comfort. Second, and more specific to smart technologies, our results bolster the case for smart thermostats that automate the behaviors engineers assume. While this is merely an implication of our findings, more direct evidence comes from Blonz et al. (2026). They study the effects of automating cooling setbacks on a smart thermostat when the price of electricity is highest and find that it increases setbacks and reduces HVAC system runtimes when prices are peaking. Similar types of automation in other types of smart technologies may help them achieve the conservation effects predicted by engineers.

Beyond energy-efficiency policy, the underlying mechanisms that drive our results offer contributions to the broader literature on predicting the effects of interventions at scale. For example, consider the literature in the the medical and economic sciences regarding decreased impacts when interventions are rolled out at a larger scale (see, e.g., Kilbourne et al., 2007; Weiss, Bloom and Brock, 2014; Supplee and Meyer, 2015; Supplee and Metz, 2015; Gottfredson et al., 2015; Cheng et al., 2017; and in economics, Al-Ubaydli et al., 2017, Al-Ubaydli, List and Suskind, 2020). A leading explanation for these drops in effect sizes is non-adherence: that people are less likely to use a treatment as intended at scale. We provide evidence that failing to account for the fact that humans do not always adhere to using smart thermostats as envisioned is the key hindrance to generalizing engineering estimates of the technology’s effectiveness.

The remainder of this study is organized as follows. In Section I we describe the details of the field experiment, the sample of households in the study, and our data. The following section formalizes our empirical specification. Section II presents our model estimates, and Section III explores the mechanisms that drive our findings. The final section concludes.

I. Experimental Design

In this section we describe the experiments that we use to evaluate the effectiveness of smart thermostats. We start by describing the smart thermostat that was installed in treatment group homes. Then we describe how treatment and control group homes were recruited into the experiments, the randomization to the treatment or control group, and installation rates of the smart thermostat in the treatment group. After, we describe the data that we observe on the households recruited into the experiments. We conclude by statistically summarizing our sample, the statistical balance of pre-recruitment outcomes across the treatment and control groups, and provide an event study analysis of pre- and post-recruitment energy consumption to motivate the empirical specifications used to evaluate the smart thermostat.

A. *Smart Thermostat*

The smart thermostat studied in our experiments was produced by Honeywell. It has three features common to smart thermostats. First, a household could use their smartphone or computer to program an extensive schedule of temperature setpoints for each day of the week. Second, households could remotely override scheduled setpoints from their smartphone or computer. Third, the smart thermostat purported to learn how to efficiently transition from scheduled temperature setpoints. For more on the scheduling, override, and user interface can be found in Supplemental Materials Section B.

Of course, the smart thermostat we study also had specific bundle of features that might limit the generalizability of our experimental results. One such feature is the user interface, which was designed by Opower. In particular, the smartphone application that remotely controlled the smart thermostat featured a page that presented a comparison of a household's current temperature setpoint to the average of neighbors and efficient neighbors. This nudge to encourage less energy intensive temperature setpoints is not entirely unique to the thermostat we study. For example, Nest smart thermostats use a leaf that gains its green color as adopters schedule less energy intensive setpoints (Nest, 2026). At the same time, nudges to use less energy intensive setpoints are not a universal feature of smart thermostats. While this may limit the generalizability of our results to other smart thermostats, it does so in the direction of making our experimental results more likely to overstate effectiveness of the smart thermostat.

Another more general example of the specific features of the smart thermostat we study stems from the fact that it was produced in 2012 and smart thermostats have developed as a technology since then. Most notable are features in more recent smart thermostats that connect to motion sensors in the home or the smartphones of residents to automatically adjust setpoints when a home is empty. Automation features like these are likely to help smart thermostats achieve energy savings closer to the predictions of engineers.

B. *Recruitment, Randomization, and Installation*

We evaluate the effects of a smart thermostat with data from two framed field experiments (Harrison and List, 2004). Subjects in the two experiments were recruited to participate in the experiment at malls, farmers markets, and festivals. Recruiters were instructed to emphasize the convenience, energy savings, and additional comfort offered by the smart thermostat (see Supplemental Materials Section C for more on recruitment). Across both of the experiments, eligibility for the experiment required households to own their home, have just one thermostat, use forced air heating and cooling, use an iPhone or Android smartphone, have a high-speed internet connection, and have no plans to move homes in the next year.

Recruitment for the first field experiment took place across four counties in Northern California from July through October of 2012. Subjects in the second experiment were recruited from December of 2012 to February of 2013 in three Central California counties. At this time, adoption rates of smart thermostats were low. According to data from the EIA RECS, two to three years after our experiment only 4.09% of all households in the survey and 10.58% of households eligible for the framed field experiments owned a smart thermostat (Energy Information Administration, 2018). For more on where the households in the treatment and control group of the experiments resided, see Supplemental Materials Section D.

Across both experiments, our final sample is comprised of 1,385 households. There are six households for which we do not observe electricity consumption and another 16 households with no natural gas information. Thus, there are a total of 1,379 unique households across the electricity samples and a total of 1,369 unique households across the natural gas samples. Figure 1 illustrates how the samples in each experiment were divided into the treatment and control group, as well as the number of households in the treatment group that successfully had the smart thermostat installed, declined installation, or had a failed attempt at installation. Across the two treatment groups, installation rates of the smart thermostat were 73%. Of the remaining treatment homes, 19% declined and 8% had complications that prevented installation (e.g., compatibility issues). Among the treatment group homes in which the smart thermostat was successfully installed, installations happened shortly after assignment to the treatment group: 50% of households had their thermostat installed within 5 days of assignment to the treatment group and 95% had it installed within 30 days. For more on the timing of installations, see Supplemental Materials Section D.

C. Energy Data

All households in the experiments were equipped with smart meters that enabled PG&E to record household-level data on hourly electricity use and daily natural gas consumption. The quantity of electricity consumed is measured in kilowatt hours (kWh), and the unit of measurement for natural gas is a therm (thm). As we cannot observe temperature setpoints directly for households that do not install our smart thermostat and energy is the policy-relevant outcome, the measures of energy consumption serve as the outcomes of interest for our analyses. In total, we observe an average of 11,908 hourly electricity use decisions for the 1,379 households in electricity sample and 495 natural gas use decisions for the 1,369 households in the natural gas sample over an 18 month period from July 2012 through December 2013. That is, even though recruitment occurred first for the Northern California experiment and second for the Central California experiment, in both of the experiments we observe energy data once recruitment began in Northern California and stop observing energy data about ten months after recruitment ended in the Central California experiment. For more on the

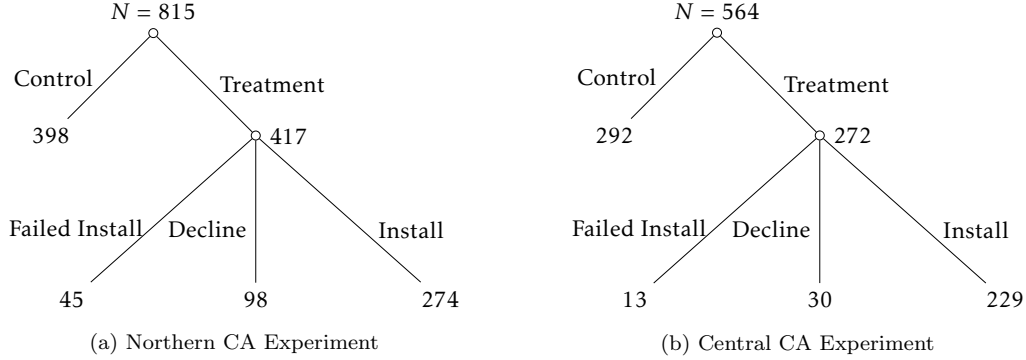


FIGURE 1. RANDOMIZATION OF SAMPLE

Note: This figure presents the number of households in treatment and control for each experiment based on the households for which we observe electricity consumption. Treatment group installations of the smart thermostat, failed installations, and declined installations are also reported. Panel (a) reports the sample sizes for the Northern California experiment and panel (b) reports the sample sizes for the Central California experiment.

timing of the outcomes we observe, see Supplemental Materials Section D.

D. Additional Sources of Data

EXTERNAL DATA . — We supplement our outcome of interest with information from several external sources and additional data collected as part of the experiment. First, we compile hourly temperature, humidity, and heat index readings for each county in the study from the National Oceanic and Atmospheric Administration (NOAA). This data reveals that while the average outside temperature is a moderate 63.7° F, there is substantial variation in conditions across and within the two experiments. That is, summers are hot, humid, and likely to require the use of air conditioning to ensure comfortable indoor temperatures. While the rest of the year is more moderate, there are many days cold enough to create demand home heating.

Second, we use data from two sources that allow us to test whether smart thermostats have a differential effect on usage when demand for electricity peaks. One resource we use is average hourly real-time price of electricity from the California Independent System Operator (CAISO), with the highest prices indicating that demand is likely at an unusually high level. The other resource we use as a proxy for high demand events are data on system-wide peak-alert messages from CAISO and utility-wide alerts from PG&E. The latter of these alerts (referred to as “SmartDays” by the utility) are issued at a finer spatial scale, but a more granular temporal level (daily) than the former. In contrast, the CAISO alerts are issued hourly, but apply to a broader area. Since system-wide alerts may occur on days when it is less obvious that there is a need to reduce demand to

avoid brownouts based on local conditions, we also determine whether CAISO alerts were broadcast by local media outlets (e.g., the *Fresno Bee* or the *Bakersfield Californian*) to ensure that they reach a reasonable level of publicity to be salient to the households in our experiments. Conditioning on the additional information from these sources allows us to examine whether smart thermostats have a pronounced effect when the cost of electricity production is the greatest.

For more on the external data sources described here, see Supplemental Materials Section E.

INTERNAL DATA . — In addition to the external data we collect, we also observe a high-frequency, exact-time log of 3,967,558 HVAC system events, including user interactions with their smart thermostat, from 372 of the roughly 500 households that had the smart thermostat installed. The dataset is truncated in January of 2013. As a result of the timing of the recruitment in the two experiments, the interactions data is heavily skewed towards the Northern California experiment. That is, approximately 95% of the observations of interactions with the HVAC system come from the Northern California experiment and only about 5% from the Central California experiment.

The system events and user interactions we observe include ambient temperature, HVAC state, and heating and cooling setpoints. We infer whether a setpoint is scheduled or an override based on the second at which it occurred, with scheduled setpoints occurring within seconds of the start of a minute and overrides occurring otherwise.

With our inferred observations of scheduled setpoints, we can address the prevalence of setbacks in the treatment group. To do so, we define a setback as occurring when a scheduled setpoint changes the temperature in a less energy intensive direction. For instance, if a setpoint for heating changes the temperature from 70 to 65 °F in a given hour, then a 5 degree setback would have occurred in that hour. Alternatively, if a scheduled setpoint for cooling changes the temperature from 75 to 80 °F, then a 5 degree setback would have occurred in that hour. This definition of setbacks allows us to create a dataset of setback spells corresponding to the household’s scheduled temperature changes over time. For each spell, we calculate three measures of potential energy savings: the setback temperature change (°F), the duration of the setback (hours), and the interaction of the two as measure of overall energy intensity (degree-hours).

Finally, some data on characteristics of both treatment and control households was collected during the recruitment phase of the experiments. We do not use these time-invariant household characteristics in our main analysis because they are redundant to household fixed effects, but next make use of these characteristics to test for balance across the treatment and control groups in the experiments. Before presenting our balance tests, though, we point readers to Supplemental Materials Section F for more on the HVAC system data.

E. Balance

To scrutinize whether the experiment successfully randomized participants into treatment and control groups, we regress a treatment indicator on household and home characteristics, as well as pre-recruitment energy use. Across each of these regressions, we neither find any statistically significant relationships between the home and household characteristics nor the pre-recruitment energy use (see Supplemental Materials Section G for the estimates and a longer discussion of the estimates). This failure to find statistically significant imbalance is consistent with an appropriate randomization process.

F. Time-Trend and Event-Study Analyses

To illustrate basic temporal patterns in the data and the effect of experimental assignment on energy consumption Figure 2 plots daily mean energy consumption by event time for each experiment after projecting out household fixed effects. Event time is measured relative to assignment to the treatment or control group. Panel (a) plots electricity use, while Panel (b) plots natural gas consumption. The figure shows that average energy consumption is not clearly different across the treatment and control groups before or after recruitment into the experiment. However, the raw energy consumption data may be too variable to indicate a clear difference between group trends in energy consumption even if their were one.

To provide a clearer picture of whether there differences in the trends, Figure 3 plots the coefficient estimates and 95% confidence intervals from event studies of the effect of assignment to treatment on daily energy consumption. Each point represents treatment and control group differences in average daily energy consumption (relative to that difference one week before the experiment) over the same event time as Figure 2 after controlling for household fixed effects and daily time effects. To mitigate the visual effects of noisy coefficient estimates resulting from unbalanced lags and leads at the endpoints of the time window, we bin all lags and leads that are based on fewer than 30 observations (Schmidheiny and Siegloch, 2023; Clarke and Tapia-Schyte, 2021). Panels (a) and (b) plot estimates for electricity and natural gas consumption, respectively, based on data from the Northern California experiment. Panels (c) and (d) plot the analogous figures for the Central California experiment. Consistent with Figure 2, the event study plots show evidence of parallel pre-trends, but do not indicate large, persistent effects of being assigned to treatment on energy use.

We note that these figures do not account for incomplete take-up of the treatment, and they are based on temporally aggregated, day-level data. For these reasons, in the next section, we outline empirical models that allow us to instrument for smart thermostat installation and take advantage of the high-frequency nature of the electricity consumption data.

II. Results

In this section we describe the specification of the empirical model that we use to estimate the effect of the smart thermostat on its adopters. Then we report estimates of this effect in the two experiments, after which we report estimates of the effect that are obtained with specific subsamples to address heterogeneity in the effectiveness of the smart thermostat.

A. Model Specification

We model the effect of a smart thermostat on household i 's consumption of energy type j in time period t (e_{it}^j) using a difference-in-differences instrumental variables (DDIV) model:

$$(1) \quad e_{it}^j = \alpha_i^j + \beta_t^j + \gamma^j S_i P_{it} + X_{it} \delta^j + u_{it}^j,$$

where S_i is an indicator equal to one if household i adopts the smart thermostat (i.e., is assigned to the treatment group and has the thermostat installed), P_{it} is an indicator for household i 's post-assignment status in time period t , X_{it} is a vector of controls, α_i^j is a household fixed effect, β_t^j is a time effect, and u_{it}^j is a household and time varying unobservable. We estimate the model separately for each energy type and conduct inference on the estimated parameters with standard errors clustered by household.

The coefficient of interest in Equation (1) is γ^j , which measures the difference in average energy use across the pre- and post-intervention periods for smart relative to conventional thermostat households. Since households in the treatment group could decline installation of the smart thermostat, we instrument for installation with random assignment to the treatment group. Formally, we estimate γ^j using two-stage least squares (2SLS). This requires $E[Z_{it}^j u_{it}^j] = 0$, where $Z_{it}^j = (\alpha_i^j, \beta_t^j, T_i P_{it}, X_{it})'$, and T_i is an indicator for household i 's treatment status in our experiment. That is, Equation (1) is the second-stage equation, and the first stage is modeled as

$$(2) \quad S_i P_{it} = \theta_i^j + \kappa_t^j + \lambda^j T_i P_{it} + X_{it} \pi^j + w_{it}^j.$$

Under plausible conditions, our coefficient of interest, γ^j , identifies the average effect of the smart thermostat on adopters or ATT of a smart thermostat. These conditions require that assignment to the treatment and control group was randomized, assignment to the treatment group was relevant for smart thermostat adoption, and assignment to the control group did not increase the likelihood of installing the smart thermostat (i.e., no defiers) (Cornelissen et al., 2016). While this last condition is impossible for us to examine empirically, the others are not.

Consistent with the randomization assumption, fixed and pre-experimental outcomes are balanced across the treatment and control groups (see Section I.E and Figure 3). Furthermore, assignment to the treatment group was relevant, with Figure 1 showing that across the two experiments 66% to 84% of the households assigned to the treatment group installed the smart thermostat. Finally, while we cannot formally test whether there were any defiers in the experiments, it seems unlikely since so few households in the treatment group declined installation of the smart thermostat. For a more patient discussion of the conditions needed for identification, see Supplemental Materials Section H.

B. Main Estimates

Table 1 presents estimates of the average effect of the smart thermostat on adopters. Panel A reports estimated effects on hourly electricity consumption and Panel B reports analogous estimates based on daily consumption of natural gas. Within each panel, we report the estimated effect in each experiment as well as an estimate obtained from a sample that pools the two experiments. Each column of Table 1 corresponds to a specification that includes a different set of control variables. Column (1) reports estimates of a basic version of the DDIV model without any fixed effects, time effects, or other additional controls. Column (2) reports estimates from a similar model that adds linear and quadratic county temperature and humidity readings to control for ambient weather conditions, as well as indicator for the Northern California experiment in the combined sample to control for potential differences across the two experiments. In Column (3), we add household (HH) fixed effects to the previous model that control for all of the time-invariant, unobserved characteristics of the home and household (e.g., age and square footage of the home, number of family members). Column (4) reports estimates from a model that adds month-of-year (MOY) effects to the previous specification in order to control for aggregate, time-varying effects such as seasonal variation in weather patterns. In Column (5), we add day-of-week (DOW) effects to the previous model in order to control for variation in daily usage patterns due to occupant work and schooling schedules. Finally, in Column (6), we replace the time effects in the previous model with hour or day effects depending on which type of energy use is being modeled.

Each estimate reported in Table 1 shows that the ATT is statistically and economically indistinguishable from zero. For example, the coefficient estimate of -0.001 reported in Column (6) of the third row in Panel A indicates that, across the two experiments, a smart thermostat causes a 0.001 kWh decrease in electricity usage per hour. The estimate of the standard error, which is clustered by household, of 0.022 reported in parentheses indicates that this estimate is statistically indistinguishable from zero. To put the magnitudes of these effects in context, Column (7) reports mean energy use in the control group in the corresponding sample. The estimated effect is equivalent to 0.09% of the control group energy use of 1.140 kWhs per hour. The corresponding natural gas coeffi-

cient estimate in Panel B is equivalent to an increase of 1.70% of control energy use. Across all specifications in both panels, the lack of economic or statistical significance indicates that adopting the smart thermostat did not reduce energy usage.¹

The estimated coefficients in Table 1 also cast doubt on the validity of the savings predicted by engineers. Consider the prediction for policymakers in California reported in the state’s TRM reports, which respectively predicted a 1.3% and 4.0% savings in electricity and natural gas. Normalizing the results from both experiments in Table 1 with the mean in the final column, we see that once household fixed effects are included the smart thermostat saved 0.09% to 0.26% of electricity consumption and increased natural gas consumption by 1.62% to 1.85%. While the electricity results are not precise enough to statistically significantly reject the TRM prediction, all natural gas estimates except for the result in the first column reject the TRM prediction.

Of course, the results of our experiment may not generalize. The smart thermostat we study lacks some of the automated features of more recent smart thermostats. For example, some newer models automatically setback temperatures when motion detectors or residents’ smartphones suggest the home is empty. This functionality is likely to increase energy savings relative to that in our experiments by removing human decision-making from the energy conservation process. At the same time, the smart thermostat in our experiment also has a distinctive feature that is likely to decrease energy use. As we discuss in Section I.A, the smart thermostat we study has a user interface partially designed by Opower to nudge households into selecting less energy intensive setpoints. Although some popular smart thermostats like the Nest feature similar nudges (Nest, 2026), not all do. That we do not find significant effects despite this energy-saving feature strengthens our interpretation of the results as a rejection of engineering predictions that is likely to generalize beyond our experiments’ smart thermostat.

C. Effect Heterogeneity

In this subsection, we investigate heterogeneity in the ATT of the smart thermostat. Our investigation focuses on whether the smart thermostat installed in treatment group homes produced significant energy savings under certain conditions or at specific times. We conduct our analyses by pooling the data across the two experiments and estimating versions of Equations (1) and (2) on subsamples

¹We note that in some specifications, the “Both Experiments” estimate is not bounded by the individual experiment estimates (e.g., Column (1) of Panel B). This occurs because we model the Northern and Central California samples as two waves of the same experiment. In practice, this means we include minimal (if any) controls for differences between the two waves when estimating the combined sample models. Since this “single-experiment specification” of the “Both Experiments” models does not guarantee that the combined result is a convex combination of the individual Northern and Central California estimates, we also estimate a “multiple-experiments specification” that includes an indicator for the Northern California experiment and interacts it with the relevant DDIV control variables in each column. See Supplemental Materials Section I for the results of this specification, which are qualitatively similar.

of the data. The subsamples are defined by conditioning on weather conditions, different periods of the week, and during peak load events. Throughout our heterogeneity analysis, when possible, our estimating equation includes household, month-of-year, and day-of-week effects. These controls are analogous to those used in Column (5) of Table 1, but vary as the subsample requires.

Figure 4 plots estimates of the smart thermostat's ATT given different conditions and brackets that indicate the 95% confidence interval. The left panel plots the ATT effect of a smart thermostat on hourly electricity consumption, and the right panel plots the effect on daily natural gas consumption. Each estimate is obtained by fitting the model on a specific subsample that is indicated on the legend under each panel. For example, the first point plotted on the left panel corresponds to conditioning the sample of observations to those that occurred when PG&E issued an alert for a peak load event. The full set of subsamples considered are: the quintiles of the ambient outdoor temperature, the quintiles of ambient outdoor humidity, the quintiles of the heat index, weekday or weekend, and peak load alerts issued by the CAISO and PG&E. Since peak load alerts cover a small number of hours and are specific to electricity, we do not report estimates related to these events for natural gas.

Broadly speaking, the estimates reported in Figure 4 are small in magnitude and statistically indistinguishable from zero. Overall, 20% of the results for electricity are positive, whereas the results for natural gas are more balanced around zero with 41% of the estimates being positive. Each panel of Figure 4 also plots the predicted engineering savings from the California TRM. While all of the 95% confidence intervals overlap or fall below the dotted line that indicates the engineering prediction for electricity savings, 10 of the 17 results for natural gas are statistically distinguishable from the engineering prediction.

However, there are two important sets of exceptions that are worth noting. First, there is marginally statistically significant evidence that the smart thermostat reduces electricity consumption during periods of high demand that strain the grid. When PG&E issues an alert encouraging customers to conserve electricity, the smart thermostat has a -0.12 kWh/hour effect during peak hours that is statistically significant at the 10% level. The analogous effect with CAISO peak-load alerts is -0.05 kWh/hour, but not statistically significant. Collectively, these estimates suggest that smart thermostats could produce energy savings during the peak-load events when conservation is most valuable, potentially because smart thermostats make it easier for households to more easily change setpoints in response to the alerts. However, this suggestion is not unambiguous because it depends on the type of alert.

Second, there is statistically significant evidence that the smart thermostat reduces electricity consumption when humidity is high. With subsamples of hours during the fourth or fifth quintiles of humidity, the smart thermostat respectively has a -0.04 and -0.07 kWh/hour effect. Both estimates are statistically significant at the 5% level. One plausible explanation for these statistically significant

reductions is that the smart thermostat reduces the level of humidity in treated homes without deviating from a pre-programmed schedule, but individuals in the control group are prone to over-adjusting their more conventional thermostats to more energy-intensive setpoints in order to mitigate the discomfort caused by high humidity.

We conclude by noting that the smart thermostat produced no energy savings for the vast majority of subsamples. With each quintile of ambient outdoor temperature and the heat index, the smart thermostat had a null effect. Similarly, the smart thermostat did not produce statistically significant reductions in energy consumption across the days of the week, hours of the day, or their intersection. Collectively, these results suggest that with the exceptions of peak load events and high humidity weather, the smart thermostat we study consistently failed to produce the energy savings predicted by engineers. In the next section, we consider why this was the case.

III. Why Do the Predictions of Engineers Fail?

In this section, we consider why the engineering model predictions (of decreases of 1.3% in electricity use and 4.0% in natural gas consumption) overstate the effects we find in our experiments. To do so, we supplement our analysis of the smart thermostat’s effectiveness with data on adopter interactions with their thermostat. We start by presenting summaries of adopter setpoint schedules, overrides, and setbacks. Then we estimate variants of our empirical model that partition adopters by their setpoint schedules, overrides, and setbacks to see whether samples of adopters who use their thermostat as engineers assume can recover the engineer’s predictions. Throughout this section, our analysis is of adopters in the Northern California experiment because, as discussed in Section I.D, the data on thermostat interactions is much more prevalent in the Northern California experiment than it is with the adopters from the Central California experiment.

A. Do Adopters Schedule Setpoints?

One possible explanation for the null effect of the smart thermostat is that the adopters in the treatment group did not use the thermostat’s setpoint scheduling functionality as intended. Alternatively, if they did schedule setpoints, they might have programmed energy-intensive setpoints. In this subsection, we investigate these two possibilities using data on the interactions that adopters had with their smart thermostat.

First, we find little evidence that adopters fail to use the smart thermostat’s scheduling functionality. Comparing the time between installation of the smart thermostat and the first scheduled adjustment to a setpoint, we find that adopters quickly scheduled setpoint changes. The median difference between installation and the first scheduled setpoint change is zero days and the mean difference is 4.6 days. We also find that adopters schedule multiple setpoint changes at times

of the day that are consistent with reduced energy use. On average, adopters schedule 3.8 changes to their setpoints per day. Figure 5 plots a measure of the frequency of setpoint changes by hour of the day. Consistent with users reducing energy consumption when they are most likely to be away from the home or asleep (Department of Energy, 2020), setpoints change most frequently in the early morning and late evening hours of the day, with a smaller mass of setpoints in the late afternoon/early evening period. Collectively, these two results suggest that it is unlikely that the null effect of the smart thermostat is attributable to adopters failing to schedule setpoints as intended.

Second, we find mixed evidence on whether adopters schedule efficient or inefficient setpoints changes. Figure 6 presents a box and whisker plot of the temperature settings associated with hourly cooling and heating setpoint changes, with a dashed line indicating the DOE's recommendation (Department of Energy, 2020). The recommendation for cooling setpoints is 78 °F and 68 °F for heating setpoints. The figure shows that the median setpoint schedule changes are consistent with DOE recommendations, as are the average heating and cooling setpoint adjustments of 64 °F and 79 °F. However, Figure 6 also shows that there is substantial variation in setpoint change temperatures in any given hour. The 25th percentile of cooling setpoint changes is frequently lower than the DOE's recommendation and the 75th percentile of adjustments to heating setpoints is often higher than the DOE's recommendation.

Overall, the data on the setpoint changes of adopters suggests that setpoints are quickly scheduled at times consistent with engineering models and the temperatures selected are broadly consistent with DOE recommendations. At the same time, there is evidence of substantial heterogeneity in the efficiency of the setpoints that may have limited the overall effectiveness of the smart thermostat.

B. Do Adopters Deviate from Their Scheduled Setpoints?

Even though many adopters scheduled setpoints consistent with the DOE's recommendation, another potential explanation for the null effect of the smart thermostat is that adopters remotely override their efficient setpoints so frequently that the thermostat's expected energy savings are not realized. To explore this possibility, Figure 7 plots a measure of the frequency of setpoint overrides by time of the day. Across both heating and cooling, the figure shows that overrides are relatively rare during the nighttime hours of 1:00 AM to 5:00 AM and relatively common from 6:00 AM to 11:00 PM. Heating overrides peak in the morning and early evening, while cooling overrides rise throughout the day until about 6:00 PM. More importantly given our focus, the figure illustrates that users often override their scheduled setpoints both when heating and cooling their homes. Compared to the previously noted 3.8 setpoints per day, users in our data temporarily override their scheduled setpoints an average of 1.7 times per day.

When adopters override their scheduled setpoints, they typically do so towards more energy intensive setpoints. Figure 8 plots kernel densities of the difference

between the override temperature and the scheduled setpoint for heating and cooling. The figure shows that when adopters override their scheduled setpoints, they generally do so in ways that use more energy: when cooling, they set temperatures colder and when heating, they set them warmer. Collectively, Figures 7 and 8 suggest that individuals undo some of the benefits of their scheduled setpoints with the remote override functionality of the smart thermostat.

C. Do Adopters Use Setbacks as Engineers Assume?

While much of the case for smart thermostats conserving energy relies on adopters scheduling and not overriding their setpoints, the schedule must entail appropriate temperature changes for extended intervals, such as during weekdays when residents are often not present in the home or overnight when they are asleep (Department of Energy, 2020).² To further examine how smart thermostats are actually used, we define a setback as an event that occurs whenever an adopter schedules a less energy intensive setpoint (i.e., a higher temperature setting for cooling and a lower setpoint for heating). We characterize setbacks based on (i) the magnitude of the temperature change they induce and (ii) the duration of the change. Figure 9 displays a plot of the marginal distribution of each characteristic separately for both cooling and heating setbacks.

The marginal distributions of setback-types illustrate that there is significant heterogeneity in adopter setback behavior. Panel (a) of Figure 9 plots the magnitude of temperature changes for heating and cooling setbacks. Starting with cooling, we see that most adopters setback their temperature by 1 to 2 °F, although there is another mass point around 6 to 7 °F as well. Moving to heating, there is more mass around 6 to 8 °F, but a significant proportion of adopters schedule larger and smaller temperature changes as well. Panel (b) plots the duration of setbacks in terms of hours. Across both heating and cooling, the most frequent setback duration is 7 to 8 hours, with additional significant mass points around 1 to 2 hours and 17 to 18 hours.

As we discussed in the introduction, engineering models of smart thermostat effectiveness predict savings because adopters are assumed to setback the temperature in their home for at least 8 hours a day, with each degree of temperature change reducing energy use by 1%. Relating this assumption to Figure 9, in Panel (a) we see that most adopters setback the temperature by several degrees, especially for heating where the median temperature change is 8 °F. However, in Panel (b) we see that the length of setbacks is often less than 8 hours. For example, only 25% of adopters setback their heating for 8 hours or longer. Thus, the marginal distributions of setbacks suggest that while adopters often adjust their setpoints by the envisioned magnitudes of degrees and hours, a non-trivial

²It is worth noting that Figure 6 displays evidence consistent with these energy-saving setback behaviors. The figure shows that cooling setpoints increase slightly starting at around 9:00 AM and drop back down around 3:00 PM. Heating setpoints follow a similar, but opposite pattern with a more pronounced discrepancy between daytime and evening temperature setpoints.

number of setbacks are for shorter lengths of time than the engineering models assume.

D. Do Setpoints, Overrides, and Setbacks Explain Why Engineers Overstate the Energy Savings of Smart Thermostats?

Next, we use the data on adopter setpoint, override, and setback behavior to define energy-conservation types and link these measures to the energy consumption data observed in the Northern California experiment. This allows us to consider whether engineering estimates of the impact of a smart thermostats are overstated, in part, because they rely on unrealistic assumptions about how adopters use the thermostat's functionality. To evaluate this explanation, we compare the energy consumption of the control group to different subsamples of households in the treatment group. These subsamples are selected to split households into those who use the smart thermostat as engineers assume (e.g., those who schedule 8 hours of setbacks per day) and those who do not. We then relate our estimates in the "uses-as-intended" subsample to the effects predicted by engineers, with similar magnitudes suggesting that the assumed behavior partially explains why the benefits of the smart thermostat are overstated.

We classify the adopters according to how they use the smart thermostat. We do so by defining three types that are likely related to energy conservation: high (H), low (L), and unknown types (?). The unknown type is necessary both because of non-adopters in the treatment group and because we do not observe all households who install a smart thermostat in the HVAC events data.

The high and low types are based on the distributions of five measures of how adopters use the smart thermostat: (i) the average number of scheduled setpoint changes observed per hour, (ii) the average number of overrides per hour, (iii) the median setback temperature change ($^{\circ}F$), (iv) the median setback duration (hours), and (v) median of the interaction of the temperature change and duration (degree-hours). For each of these measures, we use the 10th, 25th, 50th, 75th, and 90th percentile as cutpoints that split the adopters into high and low types. For all of the measures aside from overrides, we define high-type households as those above the cutpoint and low types as those below the cutpoint (e.g., above and below the median). In contrast, for overrides, we define high types as those below the cutpoint and low types as those above the cutpoint. We do not interpret the high types determined by these cutpoints as precise measures of energy conservation. Rather, we view them as proxies that are likely to be broadly consistent with energy conservation for our sample of adopters.

Given these classifications, we interact indicators for type with treatment and estimate the following model,

$$(3) \quad e_{it}^j = \alpha_i^j + \beta_t^j + \sum_k \gamma_k^j R_i^k T_i P_{it} + X_{it} \beta_X^j + u_{it}^j,$$

where $k \in \{H, L, ?\}$ index the three types, R_i^k is an indicator for household i being of type k and all other indexes, variables, and parameters are defined as in Equations (1) and (2). The parameters of interest in this model are the γ_k^j which measure the average difference in energy type j between treatment group households of type k and the control group. We interpret these parameters as associations that are unlikely to be causal because we are conditioning on post-treatment outcomes. We do so to explore whether certain associations can reproduce the engineering predictions. If conditioning on post-treatment outcomes like setback durations reproduces the estimates predicted by engineers, then it suggests that this error in causal reasoning helps explain why engineering estimates overstated the expected benefits of the smart thermostat.

Table 2 reports estimates of the γ_k^{thm} parameters for natural gas consumption based on this specification. To provide context for the other estimates, Panel A reports estimates from a baseline DDITT model that does not differentiate by type. Consistent with the DDIV model estimates reported in Table 1, the effect is not statistically significant. Panel B reports estimates for high- and low-type households from a model that uses scheduled setpoints to classify type; Panel C reports analogous estimates using overrides to classify types; and Panels D, E, and F report estimates from models that define type based on setback behaviors. Columns (2) through (6) report estimates based on varying definitions of the high- vs. low-type percentile cutpoint. For instance, the estimates reported in Column (4) of Panel B define high-types as those with more than the median number of setpoint changes per hour and low-types as those below the median. Intuitively, these estimates are comparisons of trends in type-specific treatment group subsamples to the overall control group. The high-type estimate in Column (4) of Panel B indicates that being a household with a smart thermostat that sets above the 50th percentile of average scheduled setpoint changes per hour is associated with consuming statistically significantly less natural gas compared to the control group. Additionally, high-type adopters with above the 90th percentile of setpoint changes are associated with the greatest difference in their natural gas use and the control group's. In contrast, low-type adopters who program relatively few setpoint adjustments consumed a statistically indistinguishable quantity of natural gas relative to the control group.

Interestingly, estimates for high-type adopters above the median of scheduled setpoint changes are broadly in line with engineering predictions. For example, Column (3) of Panel B shows the high types are associated with consuming 0.107 fewer thm per day than the control group, which consumes an average of 1.422 thm per day (see Table 1). This implies a 7.5% difference that is statistically significantly different from a null association. Coupled with the subsequent estimates in Columns (4), (5), and (6), the associations in Panel B suggest an association in setpoint types that is broadly consistent with the TRM predictions of 4.0% savings for natural gas discussed in the introduction.

Our analysis of overrides in Panel C explains little of the divergence between

our experimental estimates and the engineer’s predictions. In particular, we see a broadly similar association between the differences in the energy use of high-type adopters and the control group. We suspect this is because overrides alter energy use less frequently and over shorter windows of time than scheduled setpoints.

Moving to the setback types in Panels D, E, and F, we see that setback duration, but neither setback temperature change nor the product of the two, reproduces associations that are broadly similar to the energy savings predicted by engineers. We start with Panel D, which defines type based on the median number of degrees that an adopter sets back their temperatures. We see that regardless of how aggressively high-type adopters adjust their temperature settings in an energy-conserving direction, they consume similar quantities of natural gas as the control group.

In contrast, in Panel E, which divides adopter types by the duration of their setbacks, we find associations that are broadly consistent with engineering predictions. Compared to the control group, high-type adopters above the 75th percentile in this panel are associated with a difference in natural gas consumption that is statistically significantly different from zero. In particular, high-type adopters consume 0.201 (14.1%) fewer thm per day than the control group. The 75th percentile corresponds to setback duration of eight hours, which is exactly the setback duration assumed by engineers (Department of Energy, 2020).

Moving to Panel F, we see that high-types characterized by the product of setback temperature change and duration length are not associated with statistically significantly less natural gas consumption, but the magnitudes of the coefficient estimates are broadly consistent with an association between reduced natural gas consumption and households that have larger temperature setbacks for longer durations.

Collectively, these results suggest that engineers overstate the benefits of smart thermostats, in part, because they assume adopters uniformly use the setpoint scheduling to select energy-conserving setpoints and to setback the setpoints for at least 8 hours a day. Instead, a significant share of the treatment group used their setpoint schedules in ways that are more easily explained by a desire for a more comfortable home than a desire to save energy. Engineering estimates effectively assume these types of adopters away, as evidenced by our ability to reproduce the benefits purported by engineers with subsamples of the treatment group that behave as the engineers assume.

IV. Conclusion

This study experimentally estimates the effects of smart thermostat adoption on energy consumption. Across two framed field experiments that randomize eligibility for the free installation of a smart thermostat, we find no evidence of aggregate savings in electricity or natural gas consumption. Furthermore, with the exception of peak load and high humidity events, we find no evidence of significant savings from the smart thermostat across a battery of subsamples selected to

address heterogeneity in effectiveness. Using data from the interactions that adopters have with their smart thermostat, we explore why the energy savings predicted by engineers are not realized in our experiments. We find that adopters who use the smart thermostat in ways that are consistent with the behavior assumed by engineers use significantly less energy than control group households. Broadly speaking, these results suggest that engineering predictions fail because they do not account for the extent to which adopters are motivated to use the smart thermostat to accomplish something other than energy conservation.

There are many ways to extend our research. More than 90% of the households in our experiment faced a tiered energy pricing tariff. Although millions of Americans face tiered tariffs, they have been shown to be sub-optimal from a welfare point of view (Borenstein, 2012) and lead to the use of heuristics that dampen the response to changes in the marginal price of energy (Shin, 1985; Kahn and Wolak, 2013; Ito, 2014). Evidence from Jessoe and Rapson (2014), Harding and Lamarche (2016), Ito, Ida and Tanaka (2018), Fabra et al. (2021), and Blonz et al. (2026) suggests that there are benefits from combining time-varying prices (time-of-use or real-time pricing) and informational technology similar to the smart thermostat we study. Future work should add to our understanding of the extent to which the returns to smart thermostats are enhanced when matched with smart economics - e.g., dynamic pricing plans that get energy prices right. A related avenue of inquiry would explore the impact that such technologies have on the price elasticity of energy demand (some preliminary evidence from Herter (2007) suggests that they do impact how consumers respond to energy prices). If technology can enable people to better optimize their energy consumption, then price might become even more salient and therefore make people more marginal.

From the perspective of policymakers looking to promote energy conservation, our results provide little justification for subsidies directed towards smart thermostats. However, this does not mean that the technology has no private benefits. Perhaps the value of such technologies arises through increased comfort and convenience. A second avenue for future work is to explore these conjectures in greater detail. In this regard, we see promise in work designed to understand why smart thermostats are so popular among consumers given their costs and limited impact on energy consumption. Perhaps our focus on mitigation and energy savings has us thinking about the benefits of smart thermostats incorrectly. Rather than valuing smart thermostats for their expected savings, perhaps consumers value smart thermostats as a means to greater comfort in their home, especially in the face of a changing climate and the increased severity of extreme weather events. This avenue speaks to the energy efficiency gap literature as outlined by Allcott and Greenstone (2012) by shifting attention to a broader set of characteristics than mitigation and expected energy savings.

A final avenue for extension would be to better understand when and how different features of smart thermostats affect energy consumption. In particular, smart thermostats that automate the behaviors that engineers assume strike us as

especially promising. For example, smart thermostats that automatically adjust setpoints when the residents of a home are likely to be out as inferred by motion sensors or the smartphones of the residents. Recent research on the automation of other features of smart thermostats lends some credibility to this conjecture. For instance, Harding and Lamarche (2016) and Blonz et al. (2026) estimate the effects of a feature that automate temperature setting changes in response to time-of-use pricing. However, more direct tests of our conjecture should be conducted before basing policy on it.

REFERENCES

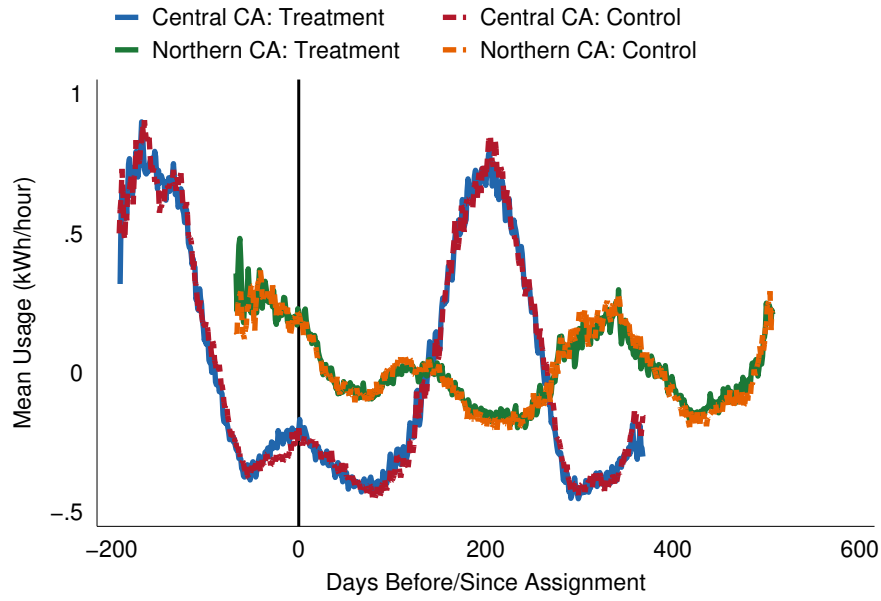
- Allcott, Hunt, and Michael Greenstone.** 2012. “Is There an Energy Efficiency Gap?” *Journal of Economic Perspectives*, 26(1): 3–28.
- Alpizar, Francisco, María Bernedo, and Paul J. Ferraro.** 2024. “Input Efficiency as a Solution to Externalities and Resource Scarcity: A Randomized Controlled Trial.” *Journal of the Association of Environmental and Resource Economists*, 11(1): 171 – 211.
- Al-Ubaydli, Omar, John A. List, and Dana Suskind.** 2020. “2017 Klein Lecture: The Science of Using Science: Toward an Understanding of the Threats to Scalability.” *International Economic Review*, 61(4): 1387–1409.
- Al-Ubaydli, Omar, John A. List, Danielle LoRe, and Dana Suskind.** 2017. “Scaling for Economists: Lessons from the Non-Adherence Problem in the Medical Literature.” *Journal of Economic Perspectives*, 31(4): 125–44.
- Blonz, Joshua, Karen Palmer, Casey J. Wichman, and Wietelmann Derek C.** 2026. “Smart Thermostats, Automation, and Time-Varying Prices.” *American Economic Journal: Applied Economics*, 17: 90–125.
- Borenstein, Severin.** 2012. “The redistributive impact of nonlinear electricity pricing.” *American Economic Journal: Economic Policy*, 4(3): 56–90.
- Brandon, Alec, Christopher M. Clapp, John A. List, Robert D. Metcalfe, and Michael K. Price.** 2014. “ZIP Code to California County Crosswalk [dataset].” Constructed by the authors. Included in the replication package at [data/public/raw/zipCountyMapping.csv](#).
- Brandon, Alec, Christopher M. Clapp, John A. List, Robert D. Metcalfe, and Michael K. Price.** 2026. “Data and Code for: “The Human Perils of Scaling Smart Thermostats: Evidence from Field Experiments.” Nashville, TN: American Economic Association; distributed by Inter-university Consortium for Political and Social Research, Ann Arbor, MI.” *TBD*.
- California Independent System Operator.** 2014. “Interval Locational Marginal Prices, Real-Time Market, All Aggregated Pricing Nodes

- (PRC_INTVL_LMP) [dataset].” *Folsom, CA: CAISO Open Access Same-time Information System (OASIS)*, Accessed April 27, 2014.
- California Independent System Operator.** 2021. “Alert, Warning and Emergencies Record [dataset].” *Folsom, CA: CAISO*, Accessed July 23, 2021.
- California Municipal Utilities Association.** 2017. “Savings Estimation Technical Reference Manual.” ERS.
- Cheng, Sierra, Erica J. McDonald, Matthew C. Cheung, Vanessa S. Arciero, Mahin Qureshi, Di Jiang, Doreen Ezeife, Mona Sabharwal, Alexandra Chambers, Dolly Han, Natasha Leighl, Kelley-Anne Sabarre, and Kelvin K.W. Chan.** 2017. “Do the American Society of Clinical Oncology Value Framework and the European Society of Medical Oncology Magnitude of Clinical Benefit Scale Measure the Same Construct of Clinical Benefit?” *Journal of Clinical Oncology*, 35(24): 2764–2771. PMID: 28574778.
- Christensen, Peter, Paul Francisco, Erica Myers, and Mateus Souza.** 2021. “Decomposing the Wedge between Projected and Realized Returns in Energy Efficiency Programs.” *The Review of Economics and Statistics*, 1–46.
- Clarke, Damian, and Kathya Tapia-Schythe.** 2021. “Implementing the Panel Event Study.” *The Stata Journal*, 21(4): 853–884.
- Cornelissen, Thomas, Christian Dustmann, Anna Raute, and Uta Schönberg.** 2016. “From LATE to MTE: Alternative Methods for the Evaluation of Policy Interventions.” *Labour Economics*, 41: 47 – 60. SOLE/EALE Conference Issue 2015.
- Davis, Lucas W., Alan Fuchs, and Paul Gertler.** 2014. “Cash for Coolers: Evaluating a Large-Scale Appliance Replacement Program in Mexico.” *American Economic Journal: Economic Policy*, 6(4): 207–38.
- Davis, Lucas W., Sebastian Martinez, and Bibiana Taboada.** 2020. “How effective is energy-efficient housing? Evidence from a field trial in Mexico.” *Journal of Development Economics*, 143: 102390.
- Department of Energy.** 2020. “Thermostats.” <https://www.energy.gov/energysaver/thermostats>, Accessed: July 21, 2020.
- Department of Treasury.** 2024. “Coordinating DOE Home Energy Rebates with Energy-Efficient Home Improvement Tax Credits: An Explainer.” <https://home.treasury.gov/news/featured-stories/coordinating-doe-home-energy-rebates-with-energy-efficient-home-improvement-tax-credits-an-explainer>, Accessed: September 14, 2026.
- Energy Information Administration.** 2018. “Energy Consumption and Expenditures Tables.” U.S. Department of Energy, 2015 Residential Energy Consumption Survey.

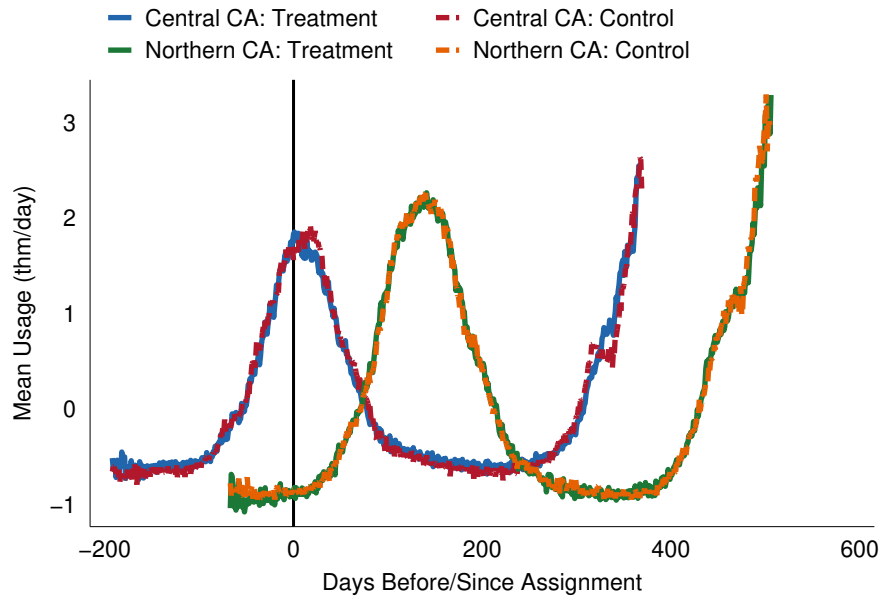
- Energy Information Administration.** 2025. “Energy Consumption and Expenditures Tables.” U.S. Department of Energy, 2020 Residential Energy Consumption Survey.
- Environmental Protection Agency.** 2019. “ENERGY STAR Summary of HVAC & Smart Thermostat Programs.” U.S. Environmental Protection Agency. Accessed: July 24, 2020.
- Environmental Protection Agency.** 2026. “Special Offers and Rebates from ENERGY STAR Partners.” U.S. Environmental Protection Agency. Accessed: March 30, 2026.
- Fabra, Natalia, David Rapson, Mar Reguant, and Jingyuan Wang.** 2021. “Estimating the Elasticity to Real-Time Pricing: Evidence from the Spanish Electricity Market.” *AEA Papers and Proceedings*, 111: 425–29.
- Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram.** 2018. “Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program.” *The Quarterly Journal of Economics*, 133(3): 1597–1644.
- Gottfredson, Denise C, Thomas D Cook, Frances EM Gardner, Deborah Gorman-Smith, George W Howe, Irwin N Sandler, and Kathryn M Zafft.** 2015. “Standards of Evidence for Efficacy, Effectiveness, and Scale-up Research in Prevention Science: Next Generation.” *Prevention science*, 16(7): 893–926.
- Harding, Matthew, and Carlos Lamarche.** 2016. “Empowering Consumers Through Data and Smart Technology: Experimental Evidence on the Consequences of Time-of-Use Electricity Pricing Policies.” *Journal of Policy Analysis and Management*, 35(4): 906–931.
- Harrison, Glenn W., and John A. List.** 2004. “Field Experiments.” *Journal of Economic Literature*, 42(4): 1009–1055.
- Herter, Karen.** 2007. “Residential implementation of critical-peak pricing of electricity.” *Energy Policy*, 35(4): 2121 – 2130.
- Houde, Sébastien, and Joseph E. Aldy.** 2017. “Consumers’ Response to State Energy Efficient Appliance Rebate Programs.” *American Economic Journal: Economic Policy*, 9(4): 227–55.
- Ito, Koichiro.** 2014. “Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing.” *American Economic Review*, 104(2): 537–563.
- Ito, Koichiro, Takanori Ida, and Makoto Tanaka.** 2018. “Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand.” *American Economic Journal: Economic Policy*, 10(1): 240–67.

- Jessoe, Katrina, and David Rapson.** 2014. “Knowledge Is (Less) Power: Experimental Evidence from Residential Energy Use.” *American Economic Review*, 104(4): 1417–38.
- Joskow, Paul L.** 2012. “Creating a Smarter U.S. Electricity Grid.” *Journal of Economic Perspectives*, 26(1): 29–48.
- Kahn, Matthew, and Frank Wolak.** 2013. “Using information to improve the effectiveness of nonlinear pricing: Evidence from a field experiment.” *Working Paper*.
- Kilbourne, Amy M, Mary S Neumann, Harold A Pincus, Mark S Bauer, and Ronald Stall.** 2007. “Implementing evidence-based interventions in health care: application of the replicating effective programs framework.” *Implementation Science*, 2(1): 1–10.
- Levinson, Arik.** 2016. “How Much Energy Do Building Energy Codes Save? Evidence from California Houses.” *American Economic Review*, 106(10): 2867–94.
- National Oceanic and Atmospheric Administration, National Climatic Data Center.** 2014. “Quality Controlled Local Climatological Data (QCLCD) [dataset].” *Asheville, NC: NOAA National Centers for Environmental Information*, Dataset identifier gov.noaa.ncdc:C00679. Hourly observations for stations WBAN 23155, 23232, 23237, 23254, 93193, 93241 and 93242, July 2012–December 2013. Accessed April 7, 2014.
- Nelson, Lorne W., and Ward MacArthur.** 1978. “Energy savings through thermostat setback.” *ASHRAE Journal*, 84(2): 49–54.
- Nest.** 2026. “Learn about the Nest Leaf.” Google. Accessed: February 16, 2026.
- Opower.** 2014. “Smart Thermostat Field Experiment Data, Northern and Central California [dataset].” Redwood Shores, CA: Oracle Corporation. Multiple electronic files, provided to the authors under a Data Use Agreement. Not publicly available; access by request to Oracle.
- Pacific Gas and Electric Company.** 2021. “SmartDay History [dataset].” *San Francisco, CA: PG&E*, Accessed July 24, 2021.
- Schmidheiny, Kurt, and Sebastian Siegloch.** 2023. “On Event Study Designs and Distributed-Lag Models: Identification, equivalence, and generalization.” *Applied Econometrics*, 38(5): 695 – 713.
- Shin, Jeong-Shik.** 1985. “Perception of price when price information is costly: evidence from residential electricity demand.” *Review of Economics and Statistics*, 591–598.

- Supplee, Lauren H, and Aleta L Meyer.** 2015. “The Intersection Between Prevention Science and Evidence-Based Policy: How the SPR Evidence Standards Support Human Services Prevention Programs.” *Prevention Science*, 16(7): 938–942.
- Supplee, Lauren H, and Allison Metz.** 2015. “Opportunities and Challenges in Evidence-Based Social Policy.” *Social Policy Report*, 28(4): 1–16.
- Weiss, Michael J., Howard S. Bloom, and Thomas Brock.** 2014. “A Conceptual Framework for Studying the Sources of Variation in Program Effects.” *Journal of Policy Analysis and Management*, 33(3): 778–808.
- Zivin, Joshua Graff, and Kevin Novan.** 2016. “Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs.” *Energy Journal*, 37(4): 1–23.



(a) Electricity



(b) Natural Gas

FIGURE 2. AVERAGE RESIDUAL ENERGY USE BY EXPERIMENTAL STATUS AND WAVE

Note: Subfigure (a) plots average hourly electricity use (kWh/hour) by day, and Subfigure (b) plots mean daily natural gas consumption (thm/day) by day. Means are calculated after projecting out household fixed effects. Only means based on 30 or more homes per day are included in the figures.

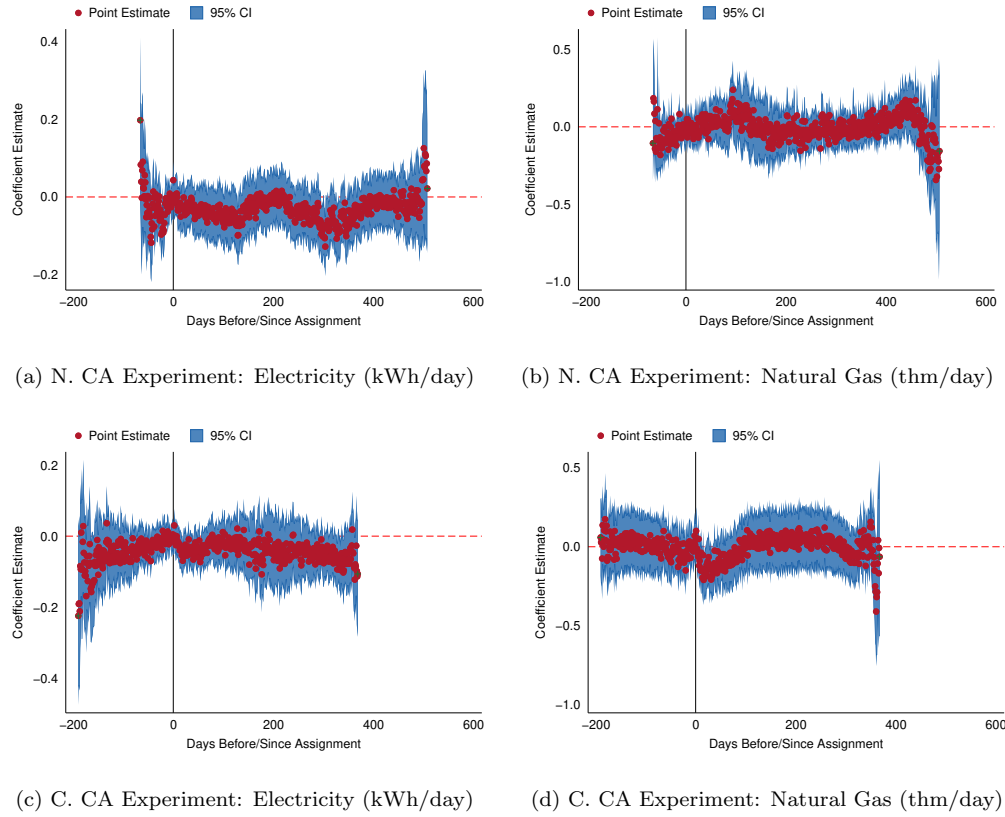


FIGURE 3. EVENT STUDY ESTIMATES OF ENERGY USE BY EXPERIMENT

Note: These figures plot the daily effect of assignment to the treatment group on electricity (kWh/day) and natural gas (thm/day) consumption. The omitted baseline period is one week before a household was randomly assigned to the treatment or control group. Red circles indicate the point estimate and blue shading indicates the 95% confidence interval constructed with standard errors clustered by household that are robust to heteroskedasticity.

TABLE 1—ESTIMATES OF THE AVERAGE EFFECT OF A SMART THERMOSTAT ON ADOPTER ENERGY CONSUMPTION (ATT)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Energy Consumption (kWh/hour or thm/day)						Mean
Panel A: Electricity (kWh/hour)							
N. CA Experiment	-0.055 (0.058)	-0.061 (0.058)	-0.016 (0.046)	-0.016 (0.046)	-0.016 (0.046)	-0.014 (0.046)	1.103 (1.196)
C. CA Experiment	0.009 (0.029)	0.005 (0.028)	0.002 (0.025)	0.002 (0.025)	0.002 (0.025)	0.003 (0.025)	1.191 (1.273)
Both Experiments	-0.031 (0.036)	-0.031 (0.035)	-0.003 (0.022)	-0.001 (0.022)	-0.001 (0.022)	-0.001 (0.022)	1.140 (1.230)
Panel B: Natural Gas (thm/day)							
N. CA Experiment	-0.009 (0.061)	0.009 (0.063)	0.085 (0.068)	0.075 (0.066)	0.075 (0.066)	0.067 (0.065)	1.422 (1.761)
C. CA Experiment	-0.003 (0.044)	0.004 (0.030)	-0.002 (0.026)	-0.001 (0.026)	-0.001 (0.026)	0.002 (0.025)	1.129 (1.332)
Both Experiments	0.062 (0.060)	0.061 (0.049)	0.024 (0.027)	0.021 (0.026)	0.021 (0.026)	0.022 (0.026)	1.298 (1.599)
Weather Controls		x	x	x	x	x	
HH Fixed Effects			x	x	x	x	
Month-of-Year Effects				x	x		
Day-of-Week Effects					x		
Hour or Day Effects						x	

Note: Standard errors in parentheses are clustered at the household level. Columns (1) through (6) report ATT estimates of the effect of a smart thermostat on energy use ($\hat{\gamma}^j$) based on separate DDIV regressions corresponding to the experimental sample indicated in the given row and the controls indicated at the bottom of the table. To put the magnitudes of these effects in context, Column (7) reports mean energy use in the control group in the corresponding sample. The samples used to produce the estimates in Panel A are based on *hourly* electricity meter readings in kWh, while the samples underlying the estimates in Panels B are based on *daily* natural gas meter readings in thm. Thus, the hour effects noted in Column (6) are included in the electricity model (Panel A) only and are day effects in the natural gas model (Panel B). Note that the estimates reported in Column (2) for samples that combine data from both experiments include an indicator equal to one for observations in the Northern California experiment. This indicator is perfectly co-linear with household fixed effects, so it is dropped from subsequent models. See Supplemental Materials Section I for full regression diagnostics and the results of alternative specifications. Based on the values of the rk *LM* and Wald *F* statistics reported in Supplemental Materials Section I, we reject the nulls of an under or weakly identified model across all specifications.

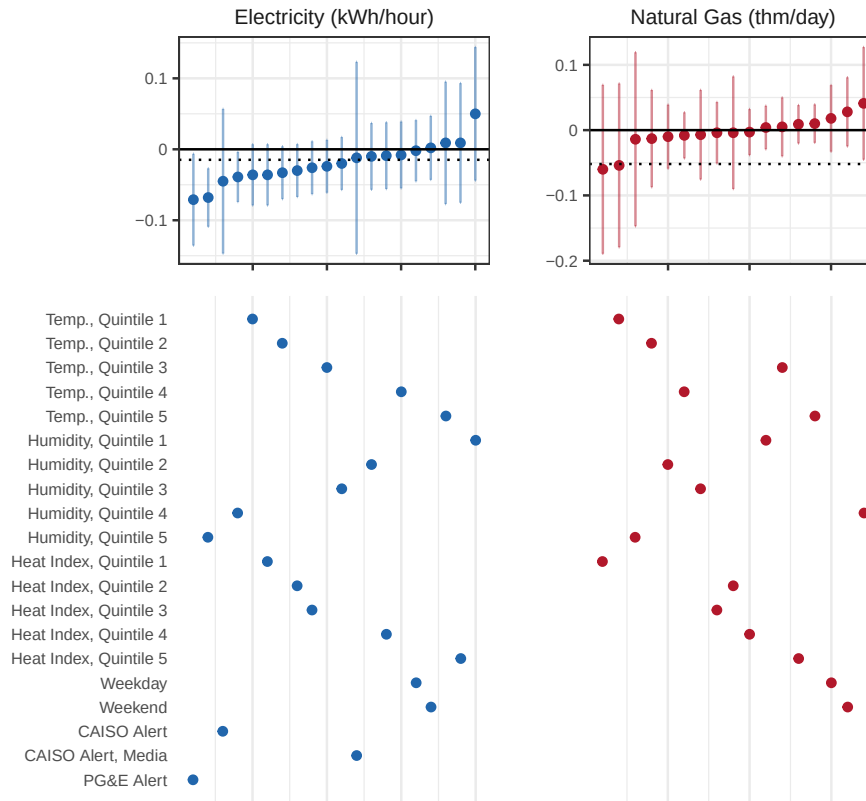


FIGURE 4. HETEROGENEITY IN ESTIMATES OF THE AVERAGE EFFECT OF A SMART THERMOSTAT ON ADOPTER ENERGY CONSUMPTION

Note: This figure plots ATT estimates of the effect of a smart thermostat on energy use and the associated 95% confidence intervals for specific subsamples of the combined data from both experiments. Estimates are obtained from specifications analogous to those reported in Columns (4), (5), or (6) of Table 1 and vary as the subsample warrants. The panel on the left plots the ATT on electricity consumption per hour, and the panel on the right plots the ATT on natural gas consumption per day. In each panel, estimates are sorted in ascending order. A legend under each figure denotes the subsample that obtains the estimate. The dotted lines on each panel correspond to the engineering prediction of a 1.3% and 4.0% reduction in consumption of electricity and natural gas for the full sample, respectively. Confidence intervals are constructed with standard errors clustered by household. Full regression diagnostics and additional heterogeneous treatment effect results can be found in Supplemental Materials Section J.

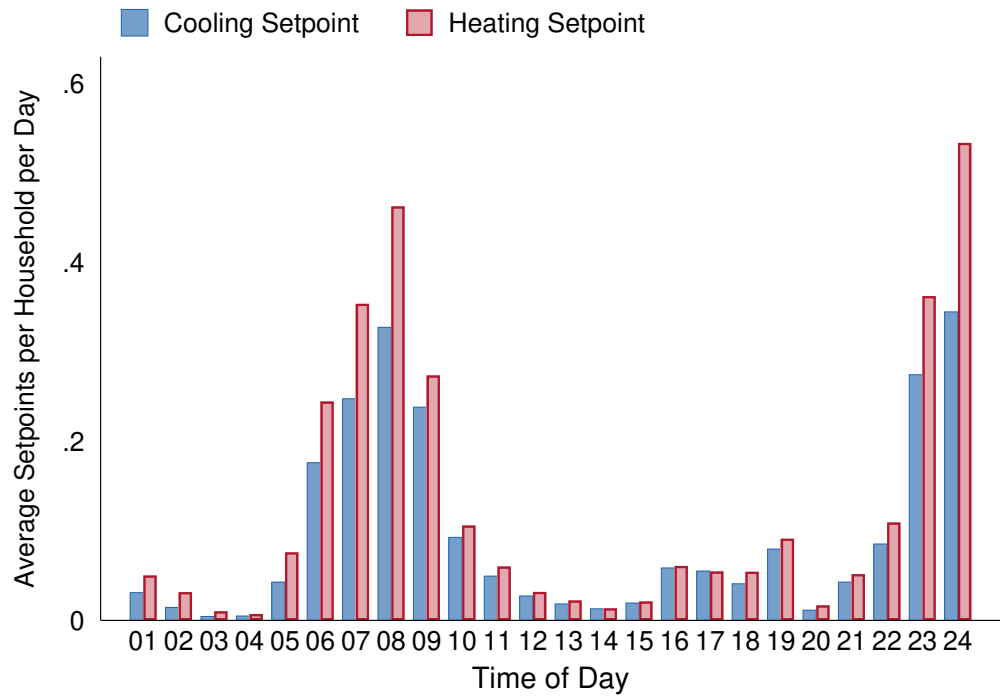


FIGURE 5. AVERAGE SCHEDULED SETPOINT CHANGES PER HOUSEHOLD PER DAY BY TIME OF DAY

Note: This figure plots the average number of setpoint changes in the given hour across households and over hours when the HVAC system is active. Thus, the plotted statistics are the proportion of observed household-days with a setpoint change in the given hour. Note that since households can schedule more than one setpoint change per day, the averages do not sum to one over the hours in the day. Blue bars denote changes to cooling setpoints, and red bars indicate changes to heating setpoints. As an example to illustrate how to interpret the bars in the figure, we observe households scheduling a cooling setpoint change during the 6:00 AM hour approximately 20% of the time in our data. See Supplemental Materials Section K for additional summary statistics.

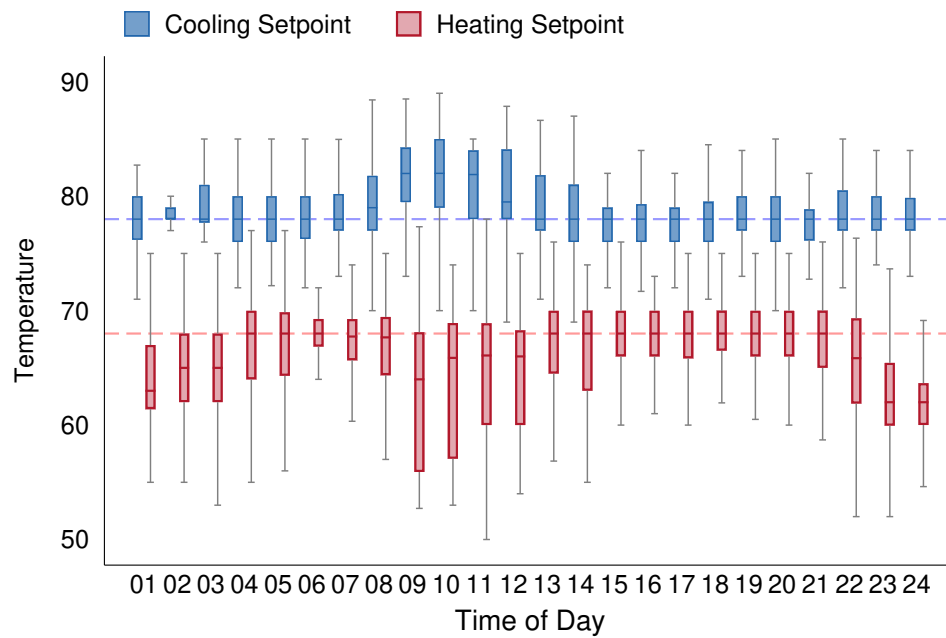


FIGURE 6. SCHEDULED SETPOINT CHANGE TEMPERATURE SETTINGS BY TIME OF DAY

Note: This figure presents a box and whisker plot of the temperature settings associated with scheduled setpoint changes by hour of day and whether the setpoint is for cooling or heating. The horizontal lines in the shaded boxes represent the median temperature setting, the ends of the boxes indicate the first and third quartiles, and the ends of the whiskers denote the upper and lower adjacent values. The dashed lines indicating the DOE's recommendation of 78 °F for cooling and 68 °F for heating. Across all hours of the day, the average (median) cooling setpoint change was 79 (78) °F. Both the average and median heating setpoint changes were 64 °F. See Supplemental Materials Section K for additional summary statistics.

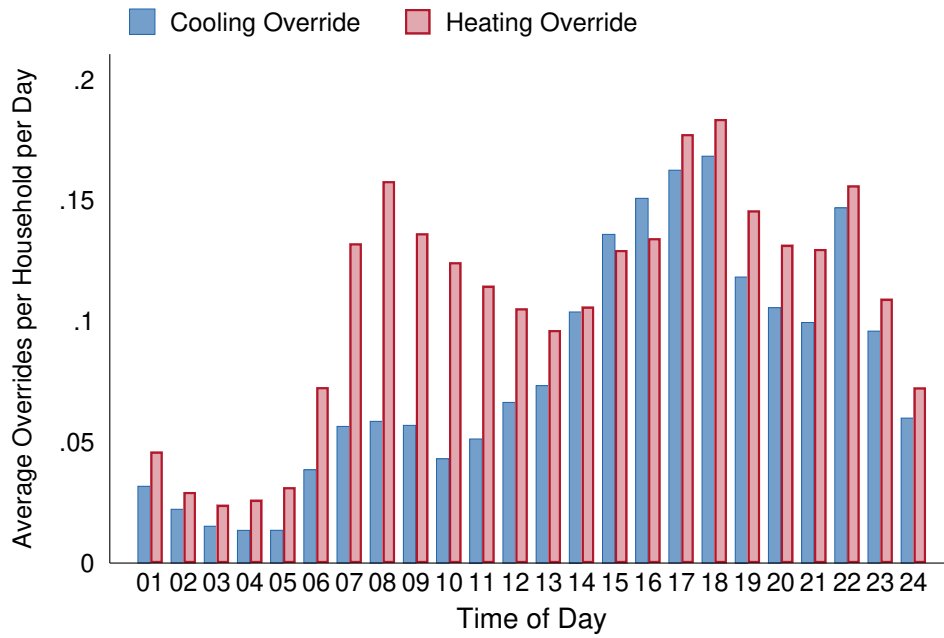


FIGURE 7. AVERAGE OVERRIDES PER HOUSEHOLD PER DAY BY TIME OF DAY

Note: This figure plots the average number of overrides in the given hour across households and over hours when the HVAC system is active. Averages for cooling are calculated over days in which the HVAC system cooled the home and for heating over days in which the HVAC system heated the home. Thus, the plotted statistics are the proportion of observed cooling/heating household-days with an override in the given hour. Note that since households can schedule more than one override per day, the averages do not sum to one over the hours in the day. Blue bars denote cooling overrides, and red bars indicate heating overrides. As an example to illustrate how to interpret the bars in the figure, we observe households scheduling a cooling override during the 6:00 AM hour approximately 4% of the time in our data.

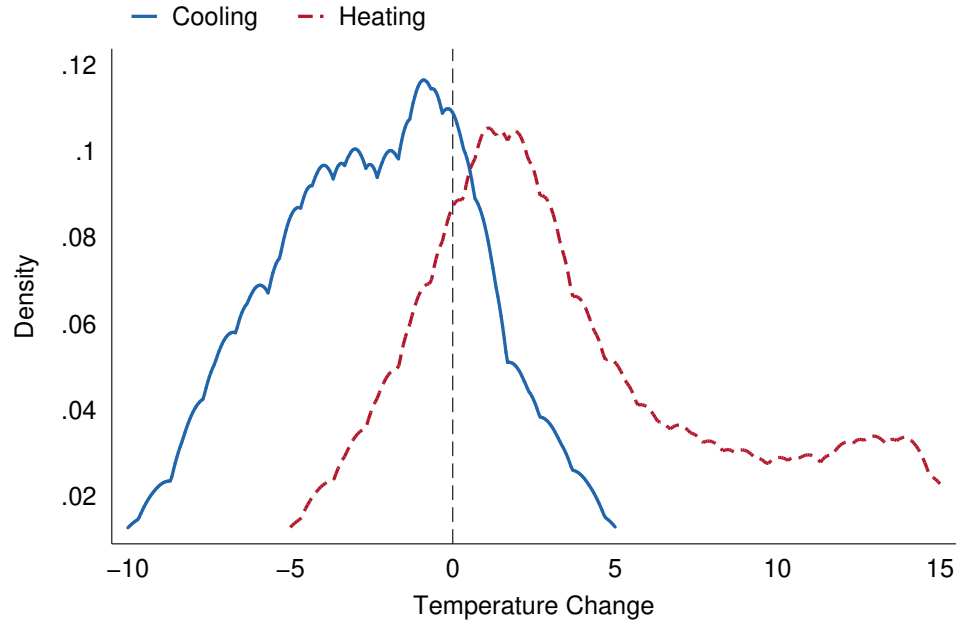


FIGURE 8. DENSITY OF DIFFERENCE BETWEEN OVERRIDE AND SCHEDULED SETPOINT TEMPERATURES BY HEATING/COOLING

Note: This figure presents the density of the difference between an override and a scheduled setpoint for heating and cooling. Densities are truncated at the 5th and 95th percentiles.

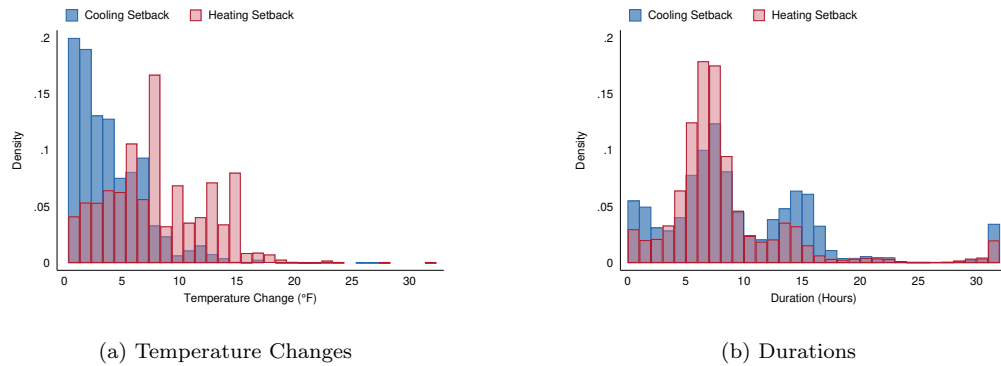


FIGURE 9. MARGINAL DENSITIES OF SETBACK TEMPERATURE CHANGES AND DURATIONS FOR COOLING AND HEATING SETBACKS

Note: The figure in Panel (a) plots the marginal density of setback temperature changes (in °F). Setbacks are defined such that positive temperature changes are energy conserving for both heating and cooling. The figure in Panel (b) plots the marginal density of setback durations (in hours). Durations are top-coded at 32 hours.

TABLE 2—ESTIMATES OF THE ASSOCIATION BETWEEN A SMART THERMOSTAT AND NATURAL GAS CONSUMPTION BY HOUSEHOLD TYPE

	(1)	(2)	(3)	(4)	(5)	(6)
	Baseline	10	25	50	75	90
	Natural Gas Consumption (thm/day)					
Panel A: Baseline Specification						
All Types: ITT ($\hat{\gamma}^{thm}$)	0.046 (0.044)					
Panel B: Scheduled Setpoint Type Classification						
High-Type: ITT ($\hat{\gamma}_H^{thm}$)		-0.060 (0.050)	-0.107 (0.049)	-0.120 (0.052)	-0.187 (0.068)	-0.316 (0.084)
Low-Type: ITT ($\hat{\gamma}_L^{thm}$)		0.088 (0.101)	0.139 (0.084)	0.031 (0.066)	0.004 (0.053)	-0.015 (0.049)
Panel C: Override Type Classification						
High-Type: ITT ($\hat{\gamma}_H^{thm}$)		0.258 (0.154)	-0.039 (0.092)	-0.021 (0.072)	-0.061 (0.055)	-0.061 (0.050)
Low-Type: ITT ($\hat{\gamma}_L^{thm}$)		-0.066 (0.048)	-0.045 (0.050)	-0.062 (0.050)	0.001 (0.063)	0.094 (0.087)
Panel D: Setback Temperature Change Type Classification						
High-Type: ITT ($\hat{\gamma}_H^{thm}$)		-0.036 (0.051)	-0.014 (0.058)	0.010 (0.070)	-0.057 (0.081)	0.026 (0.151)
Low-Type: ITT ($\hat{\gamma}_L^{thm}$)		-0.074 (0.090)	-0.082 (0.062)	-0.068 (0.054)	-0.040 (0.052)	-0.048 (0.048)
Panel E: Setback Duration Type Classification						
High-Type: ITT ($\hat{\gamma}_H^{thm}$)		-0.022 (0.049)	-0.059 (0.054)	-0.095 (0.077)	-0.201 (0.093)	-0.211 (0.087)
Low-Type: ITT ($\hat{\gamma}_L^{thm}$)		-0.163 (0.105)	-0.024 (0.067)	-0.024 (0.052)	-0.027 (0.049)	-0.037 (0.048)
Panel F: Setback Temperature Change $\times$ Duration Type Classification						
High-Type: ITT ($\hat{\gamma}_H^{thm}$)		-0.016 (0.049)	-0.016 (0.051)	-0.007 (0.062)	-0.022 (0.080)	-0.092 (0.105)
Low-Type: ITT ($\hat{\gamma}_L^{thm}$)		-0.183 (0.110)	-0.097 (0.077)	-0.071 (0.058)	-0.048 (0.052)	-0.039 (0.049)

Note: Standard errors in parentheses are clustered at the household level. All estimates are based on a sample comprised of the Northern California experiment. All specifications include weather controls, household fixed effects, and day effects. The sample underlying the estimates in all panels is based on *daily* natural gas meter readings in thm. The coefficient estimate for the unknown type ($\hat{\gamma}_?^{thm}$) is 0.177 across all specifications, and it is statistically significant at the 1% level. Across all specifications, $N = 805$, $N \times T = 398,243$, and $R^2 = 0.650$. See Supplemental Materials Section K for more on the types and the results obtained with alternative specifications.