CALIFORNIA
ENERGY COMMISSION

Energy Research and Development Division

FINAL PROJECT REPORT

Reinventing Residential Demand Response

Appendix E: Causal Effect of Short-Term Monetary Incentives and Automation on Residential Electricity Consumption

Gavin Newsom, Governor

March 2021 | CEC-500-2021-019-APE

PREPARED BY:

Primary Authors:
James M. Gillan
Datong P. Zhou
Maximilian Balandat

University of California, Berkeley
123 Main Street
Berkeley, CA 94720

Contract Number: EPC-15-083

PREPARED FOR:
California Energy Commission

David Hungerford
Project Manager

Virginia Lew
Office Manager
ENERGY EFFICIENCY RESEARCH OFFICE

Laurie ten Hope
Deputy Director
ENERGY RESEARCH AND DEVELOPMENT DIVISION

Drew Bohan
Executive Director

ABSTRACT

This report summarizes the design, implementation, and results of a randomized controlled trial conducted by OhmConnect, Inc. and evaluated by the University of California, Berkeley to investigate the causal effect of monetary and non-monetary incentives on the reduction of electricity consumption. For this purpose, a total of ~13,000 residential households serviced by the three main electric investor-owned utilities (Pacific Gas and Electric, Southern California Edison, and San Diego Gas & Electric) across California were subjected to hour-ahead interventions over a period of 14 months.

Households were found to reduce their consumption by 12-14% and appeared to be insensitive to the incentive level. Heating and cooling load were identified to be the main drivers for reductions, indicating that rebates offered for automated household devices have a large potential for residential Demand Response. Further, it was found that targeting households with high and low incentive levels based on historical behavior can lower the cost of the program significantly. Lastly, it was found that the social value of Demand Response events studied in this experiment was small.

Future work can explore the potential of conveying relative value instead of monetary value to customers, the sensitivity of users in regard to targeting events, and further drivers or obstacles for automating households.

Keywords: Demand Response, residential electricity pricing, automation, targeting, machine learning, field experiment

Please use the following citation for this report:

Gillan, James M., Datong P. Zhou, Maximilian Balandt. (University of California, Berkeley), 2021. *The Causal Effect of Short-Term Monetary Incentives and Automation on Residential Electricity Consumption.* California Energy Commission. Publication Number: CEC-500-2021-XXX-APE.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

The purpose of this report is to provide a comprehensive analysis of the residential behavioral demand response study conducted by OhmConnect, Inc. and evaluated by the University of California, Berkeley. Specifically, this report addresses how demand response (DR) events impact residential electricity consumption, how varying incentive levels affects responses to DR events, and how automation (i.e., using energy-efficient smart devices to respond to DR) affects responses to DR events.

This report attempts to answer the following key empirical questions: How can a valid experiment be designed that both suits the practicality of the problem as well as fulfills rigorous academic standards (Chapters 1 and 2)? How many participants respond to monetary incentives for DR, and what is the magnitude of this effect (Chapters 3)? Can large or small reducers be identified in a systematic fashion in order to optimize the cost of the program (Chapter 4)? Do people appeal to their environmental consciousness to reduce electricity consumption despite not being offered monetary incentives (Chapter 5)? Does academic machinery provide value in evaluating the econometric setup presented in Chapters 1 and 2 (Chapter 6)? Finally, can we avail ourselves of a survey to learn more granular user preferences than observed through pure monetary incentives (Chapter 7), and what is the economic value of these preferences (Chapter 8)?

To answer this set of questions, this report shall use the following structure:

Chapter 1 summarizes the design of the entire experiment. The experiment was conducted from November 15, 2016 to June 1, 2018. The experiment consisted of three distinct phases, which define the experimental regime a customer experienced based on the number of days that had passed since they had enrolled. Phase 1 consisted of the first 90 days for all users. Phase 2 consisted of days 91-180 for users that were not assigned to the control group. Phase 3 consisted of days 181-270 for users that experience Phase 2. Chapter 1 will provide informative illustrations about the segmentation of users into different groups within the three phases of the experiments and is the foundation for subsequent chapters which describe the three phases, their design, underlying econometric theories, and their results in detail.

Chapter 2 provides summary statistics of the experiment. Firstly, we will introduce recruitment statistics. Recruitment began on 11/14/2016 with the launch of the pilot and concluded on 8/14/17. The pilot period recruitment ran from 11/14/2016 to 12/31/16 and the study period recruitment from 1/1/2017 to 8/14/2017. Secondly, we will provide background information on the users. Specifically, we investigate how participants are distributed by region, utilities, and availability of historical smart meter data. Thirdly, the distribution of users into various experimental groups is illustrated with tables, which we refer to as the initial assignment of users. Lastly, we provide comprehensive statistics of the dispatch statistics of messages across users within the different experimental groups.

Chapter 3 reports results of Phase 1 of the experiment, which is concerned with the following two primary questions: 1) How do participants respond to varying monetary incentives during DR events and 2) how does adopting automation affect those responses? We describe the research design in further detail and report the empirical results. It also explores how these effects vary by time of day and month of year. The main dimension of randomization in Phase 1 was the assignment to receive DR messages as compared with the 90-day delayed Control

Group which received no messages. The other dimension of randomization during Phase 1 was assignment to an automation encouragement. In order to measure the causal effect of adopting an automation technology, the encouraged households were offered a rebate for the full purchase price of a new smart home device up to $240 in value.

Moreover, we provide balance statistics for the Phase 1 assignment. Given constraints to the enrollment pipeline, we were unable to stratify assignment to ensure balance a priori. This is a challenge to the design pursued in this section. Further, since the evaluation team did not perform the randomization, this check is vital to ensuring experimental validity. We analyze the average electricity consumption by hour-of-day in the 90 days prior to enrollment and test for balance in observables characteristics between treatment groups to provide evidence on the validity of assignment.

Furthermore, we explain how we use a difference-in-differences estimation strategy to evaluate the effect of DR events. This involves controlling for between-household differences by including household by hour-of-day fixed effects and data from the 90 days prior to enrollment.

Chapter 3 concludes by estimating the average effect of being enrolled in the program during an event hour and the effect of being called. The effect heterogeneity along the dimensions of hour of day, temperature, and automation is being investigated as well.

The main results of Chapter 3 are that households reduce electricity on the order of 12- 14% per Demand Response Event. This amount is largely independent from the actual incentive level, suggesting that users are insensitive to prices or that electricity consumption follows a binary decision model. The main axes along which we observe heterogeneity are ambient air temperature and automation status.

Chapter 4 summarizes the design and outcomes of Phase 2 of the experiment, which attempts to answer the question of whether or not the incentives sent during Phase 1 could be modified and targeted to improve program efficiency. The targeting strategy developed by UC Berkeley reduces costs by identifying the largest responders and sending them lower incentives, but does not reflect the optimal cost-reducing dispatch of individuals due to constraints in the experimental structure. We achieved this by estimating household-level responses using a ML model to predict each individual's counterfactual consumption during an event and then averaged the difference between these counterfactuals and observed consumption to find the so-called "Individual Treatment Effect (ITE)". Next, in a similar fashion to the balance checks from Chapter 3, we compare users in the non-targeted and targeted groups by their mean consumption in pre-treatment periods (i.e. before enrollment in Phase 1 of the experiment) as well as ambient air temperature.

Before launching Phase 2, we used extensive simulations on actual observations from Phase 1 and Phase 2 for a small subset of users who had already experienced both phases and simulated a continuation of the experiment with hypothetical targeting strategies. We then chose the one that had the largest kWh/point reduction in aggregate since it proxies well for the cost of the response to the DR provider. Finally, we estimated the causal effect of implementing our targeting strategy using the experimental assignment to targeted versus non-targeted groups for households that experienced Phase 2 when we implemented the preferred targeting strategy on a sample of 2,725 households.

Chapter 5 analyzes Phase 3 of the experiment, which is concerned with the potential of using messages with moral suasion in the form of green/environmental messaging rather than financial incentives, which we explore in Chapters 3 and 4 of the experiment. This section describes the research design in further detail and reports the empirical results. It also explores how these effects vary by automation status. Phase 3 occurred 180 days after enrollment for the Standard and Encouraged users and lasted for 90 days, after which the household was transitioned out of the experiment. Using a similar Differences- in-Differences approach as in Phase 1, we suggest that moral suasion yields smaller reductions and that there is something unique to the monetary incentives, which is consistent with intuition. It is found that environmental messaging does not yield a satisfactory amount of electricity reduction, but we remark that further analysis of the targeting strategy investigated in Chapter 4 might help shed light on this interesting aspect of the experiment.

Chapter 6 is dedicated to Machine Learning models that estimate treatment effects for Phases 1-3 using only experimental variation. Specifically, we re-estimate the average treatment effect identified in the previous sections using a non-experimental estimator, which does not require a control group. Instead, this estimator is capable of estimating individual treatment effects, which enable the design of an adaptive targeting scheme (see Chapter 4) to increase the per-dollar-reductions of users. Since this approach is not inherently causal, we run extensive simulations to benchmark the results predicted by this Machine Learning approach on the more classical, econometric approach to find that there exists a promising similarity of the results predicted by both techniques.

Chapter 7 reports results from a survey conducted after users completed the experiment. The questions ask about ownership of automation, load sources in the home, strategies to reduce, and motivations for joining. We find survey respondents owned few automatable devices and that our rebate offer increased the devices as expected. We find the majority of participants rank financial gains as their primary motivation for joining.

Chapter 8 provides an economic valuation of the experiment and hypothetical programs with similar structures. We look at the effect of different incentive levels sent during Phase 1 and quantify the kWh reduced by points paid for the sample and for groups of automated and non-automated users. We also examine what the energy return is to the automation technologies deployed by the rebates. We evaluate the wholesale market value from the energy markets using data from the California Independent System Operator and include the value of reducing the externalities from generation.

Our findings show the savings for households range from \$2.66-\$6.13 for the 90 days of Phase 1. For hypothetical programs that call 100 events per year and set incentives consistent with the wholesale price, average household savings range from \$1.82-\$4.80 annually for all households. Savings are larger for automated households, ranging from \$3.95-\$10.40 and adopting automation increases savings by \$9.68-\$25.48 in the hypothetical programs we consider.

Our estimates on the short-run avoided costs from energy procurement and externalities range from \$0.17-\$3.96 per household per year for a hypothetical 100 event program. The wide range comes from assumptions about whether events are called during the highest price periods. We find that on average, the DR events for the experiment were called during periods when wholesale prices were about $0.04/kWh compared. These estimates do not include the capacity value of the resource which may increase the value of the programs we consider.

Lastly, we calculate the private payback period and the discounted social value of automation adoption. Payback periods for the 2 smart plugs range from 3.6-12.3 years depending on the wholesale market valuation. Thermostat payback periods are longer due to higher up-front costs and range from 11.2 years to no finite payback period. The discounted social value is also sensitive to assumptions about wholesale market valuation, but range from $24 to an upper bound of $147 for a 5-year device life span (plugs) and $47 to $289 for a 12-year life span (thermostats). These numbers use experimental estimates on the effect of adopting automation, but also make assumptions about future energy prices and externalities that may affect the results.

Chapter 9 concludes and outlines areas for future research.

CHAPTER 1:
Summary of Experiment

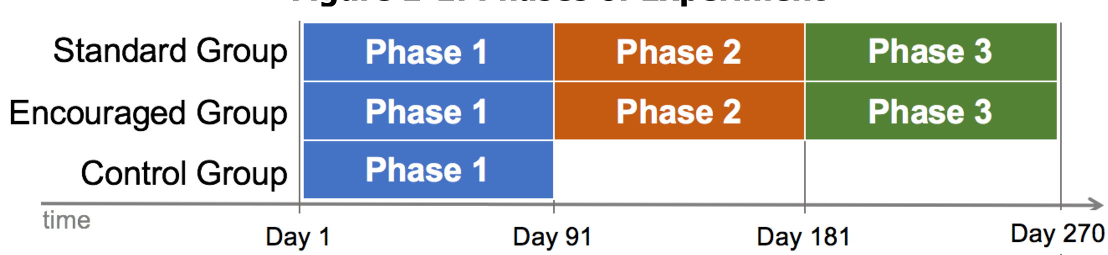
OhmConnect, Inc. (OC) is a third-party Demand Response (DR) provider in California. They implemented and experiment designed in collaboration with the team from the Energy Institute at Haas and Claire Tomlin's lab in the Electrical Engineering and Computer Science department at University of California Berkeley. OC offers a DR product to residential consumers where they can get paid for for reducing their electricity consumption during DR events they call #OhmHours. The events can be originated by OC for internal reasons or from a scheduling coordinator dispatching OC if it is awarded a bid in the Proxy Demand Resource market or from the Demand Response Auction Mechanism designed to procure DR capacity.

OC's product has two features that make studying it unique from previous residential DR studies. First, it calls events more frequently and with shorter notice than the day-ahead studies of the past that focused on critical peak pricing during a limited number of summer events. Second, OC offers a unique automation technology that shuts off appliances that households have chosen to connect to OC's platform during DR events. While direct load control is not a new concept in DR more broadly, there has been little work studying an automation technology with these features within the residential setting. Further, while OC calls #OhmHours with day-ahead notice outside of the experiment, we restricted the experimental DR events to occur with hour-ahead notice.

The experiment was conducted from November 15, 2016 to June 1, 2018 with the primary period of analysis falling between January 1, 2017 and December 31, 2017. The period of November 15, 2016 to December 31, 2016 consisted of a pilot period and the period from January 1, 2018 to June 1, 2018 consisted of an unplanned continuation of two of the experimental phases.

There were three phases which defined the experimental regime a customer experienced based on the number of days that had passed since they enrolled.

Figure E-1: Phases of Experiment



Source: UC Berkeley

Phase 1 consisted of the first 90 days for all users. Phase 2 consisted of days 91-180 for users that were not assigned to the control group. Phase 3 consisted of days 181-270 for users that experience Phase 2.

Phase 1 had several dimensions for randomization designed to understand the effect of varying incentive levels and automation on DR responses. First, users were assigned to three groups that varied along two treatment dimensions as shown in Figure E-2. The dimensions were:

- Pricing Events: Users received an average of 25 Demand Response (DR) events over 90 days communicated via email and SMS (see Chapter 3 for more details). Each incentive level for a particular DR event was selected at random with equal 20% probability from the set of all possible reward levels {5, 25, 50, 100, 300} points/kWh.
- Automation Rebate: Users were offered a rebate of up to $240 for purchasing a smart home automation device, which is paid out to the user upon successful connection of the device to their utility account.

Figure E-2: Phase 1 Experimental Design

Experimental Treatment	Control	Standard Enrolled	Enrolled + Encouraged
Pricing Events + Automation Option	×	✓	✓
$240 Rebate for Automation Adoption	×	×	✓
Pr(Assignment)	20%	40%	40%

Source: UC Berkeley

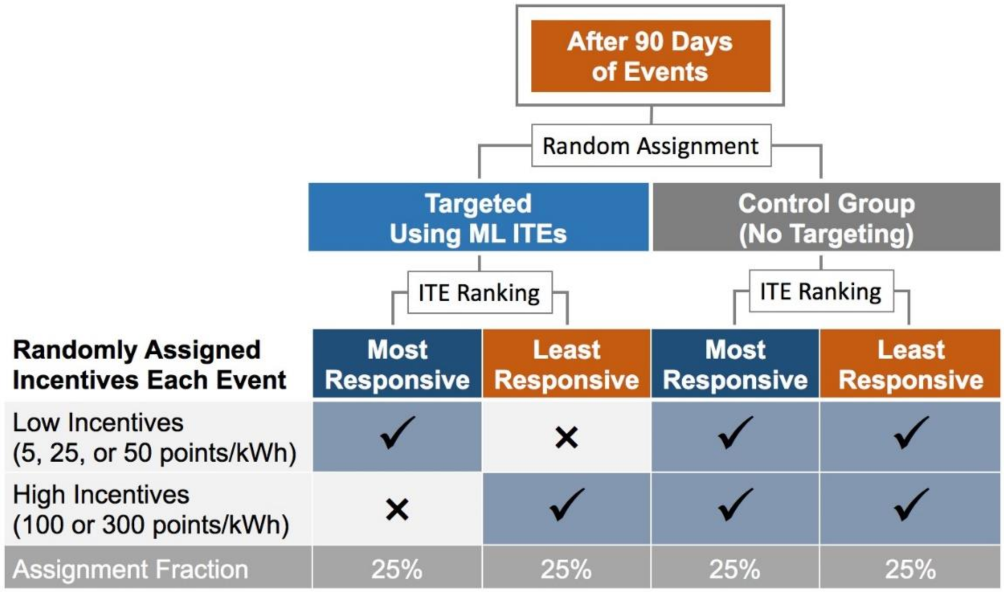
Check marks indicate households in that column received the treatment for that row and exes denote they did not.

Users were assigned to three treatment groups upon enrolling with OhmConnect (OC). With 40% probability users were assigned to "Standard Enrolled" the standard OC experience where they received Pricing Events and had the option to connect their Smart Home devices to OC's automation service. With 40% probability users were also assigned to receive the Automation Rebate "Enrolled + Encouraged". With 20% probability were assigned to a recruit and delay "Control" group that did not receive either treatment arm, but instead received an email telling them they would not receive any events for about 90 days and offering them a $10 reward for remaining enrolled over the period.

The design allows us to address three key empirical questions: 1) What is the effect of DR events on electricity consumption for those enrolled versus those who were delayed? 2) What is the effect of varying incentive levels? and 3) What is the effect of adopting automation on DR responses?

After 90 days, users in the Standard Enrolled and Enrolled + Encouraged groups were pooled and randomly assigned to two groups to understand how targeting users based on estimated responses could improve the efficiency of dispatch. Figure 3 shows the design where users are assigned after 90 days to a "Targeted" and a control "Non-Targeted" group. Targeted users were ranked by a machine learning individual treatment effect (ITE) estimator as most or least responsive and then sent either low or high incentives accordingly (see Chapter 4 for more details). Non-targeted users continued to receive all five incentive levels the same as they had been receiving during Phase 1. Figure E-3 shows the tree-diagram for Phase 2's experimental design.

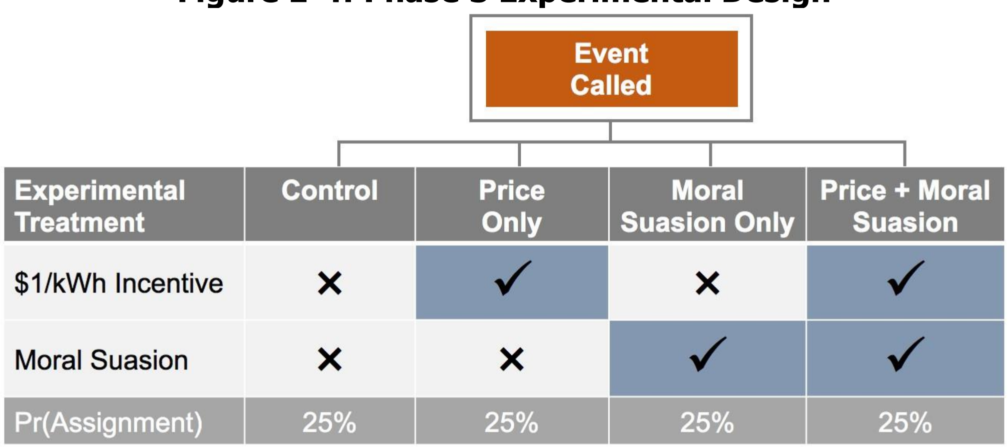
Figure E-3: Phase 2 Experimental Design



Source: UC Berkeley

Phase 3 was the final phase of the experiment and was meant to understand if moral suasion and environmental priming had a differential effect from financial incentives. After Phase 1 enrolled users had concluded Phase 2 and control users had completed Phase 1, they were pooled into an experience for 90 days where each event was randomized between four treatments with equal 25% probability as shown in the assignment diagram pictured in Figure E-4.

Figure E-4: Phase 3 Experimental Design



Source: UC Berkeley

Lastly, all users that have reached 90 days of age in Phase 2 are then rolled into Phase 3 of the experiment, which is concerned with the effect of moral suasion on the reduction in electricity consumption during a DR event (see Chapter 5). Interventions in this group occur on an event-by-event level, that is, for a particular event, each user has the same likelihood of experiencing one of the following four treatments:

- Control: Users did not receive a DR event.
- Price Only: Users received a DR event with a 100 point/kWh reward level and the same language as in Phase 1 and 2.
- Moral Suasion Only: Users received an event with no financial reward, but included the language such as "Environmental #OhmHour today from 6PM-7PM! Saving energy now could keep a dirty power plant turned off!"
- Price + Moral Suasion: Users received an event that had environmental priming language and a 100 point/kWh financial incentive.

Upon completing Phase 3, users were given the opportunity to complete a survey (Chapter 7) to reflect on their experiences and preferences formed during the experiment. These users are offered a monetary incentive for successfully completing the survey.

CHAPTER 2: Summary Statistics

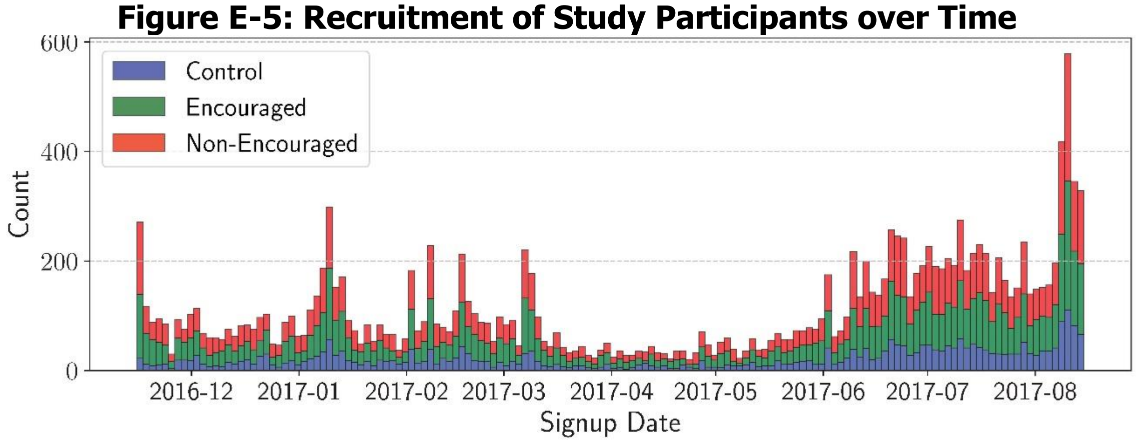
In this section we describe summary statistics on recruitment and the data used in the analysis to provide a foundation for interpreting the results.

2.1: Recruitment Statistics

Recruitment began on 11/14/2016 with the launch of the pilot and concluded on 8/14/17. The pilot period recruitment ran from 11/14/2016 to 12/31/16 and the study period recruitment from 1/1/2017 to 8/14/2017. Due to a technical implementation problem, the recruitment period was cut short two weeks and ended prior to the originally planned 9/1/2017. While this represents an unfortunate loss in data, it did not seriously affect the statistical power of the study.

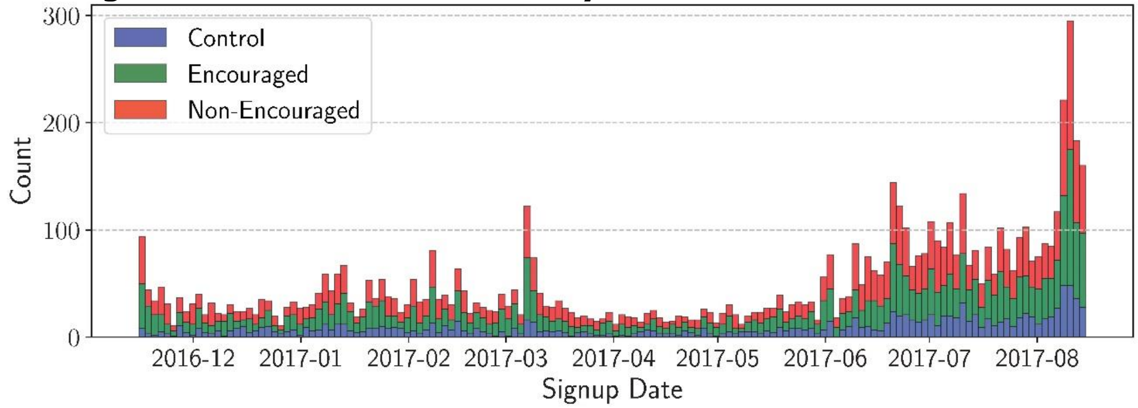
Figure E-5 illustrates the number of study participants that were recruited for the RCT broken out by time of enrollment. This is done separately for users assigned to the three different experimental groups of Phase 1 (Control, Encouraged, Non-Encouraged). Recruitment began on November 15, 2016 and ended on August 15, 2017. We observe lower enrollment figures from April 2017 – June 2016 with a noticeable peak towards the end of the recruitment period. As can be seen from the figure, the height of the red and green bars for a particular vertical slice appear to have approximately the same height, indicating that encouraged and non-encouraged users are balanced in size. In contrast, the blue bar is about half as large as the green or red bar, which is consistent with the 40/40/20 assignment of users into encouraged, non-encouraged, and control groups we elaborated on in Chapter 1 of this report.

Figure E-5: Recruitment of Study Participants over Time



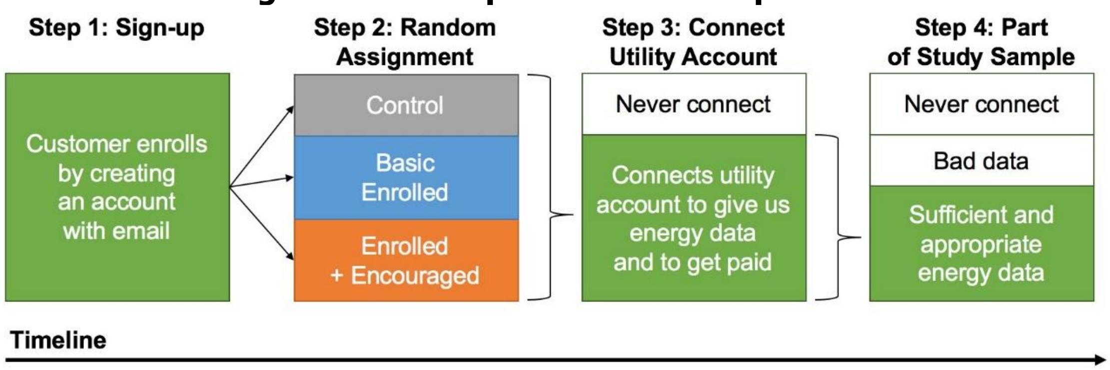
In a similar fashion, Figure E-6 plots the number of users that were recruited into the study and successfully connected their electric utility accounts. About half of all recruited users connected their utility accounts. We were unable to use the recruited users who did not connect their utility accounts because we have no energy data for them. We observe that the shape of the boxplot looks similar to the one in Figure E-5, suggesting that users across the three different experimental groups were equally likely to connect their electric utility accounts. The average number of recruits per day was 58 with a standard deviation of 48, a minimum of 5, and a maximum of 295.

Figure E-6: Distribution of Utility Account Connects over Time



The study sample consisted of users who connected their utility accounts and survived a data-cleaning process. Figure 2.3 describes this process. Step 1 shows users were deemed recruited by creating an account with their email. After recruitment, users were randomly assigned to their Phase 1 experience assignment– designated Step 2. Note again, users were not notified of any assignment other than the Control group delay messaging. Step 3, users completed the enrollment process and connected their utility accounts. Step 4, the research team cleaned the data, removing users with insufficient pre-enrollment energy data or erroneous meter data.

Figure E-7: Sample construction process



2.2: Characteristics of Experiment Participants

Figure E-8 illustrates the distribution of the lengths of available historical smart meter data among all users that have successfully connected their utility accounts. Users from Southern California Edison (SCE) have the shortest availability and those serviced by San Diego Gas & Electric (SDG&E) have the longest. We observe peaks at 365 days and 730 days, which correspond to 1 or 2 years of data availability. The black dashed lines reflect the median availability of historical smart meter data, which is 374 days for PG&E, 273 days for SCE, and 403 days for SDG&E.

Figure E-8: Availability of Smart Meter Data Across Experimental Users

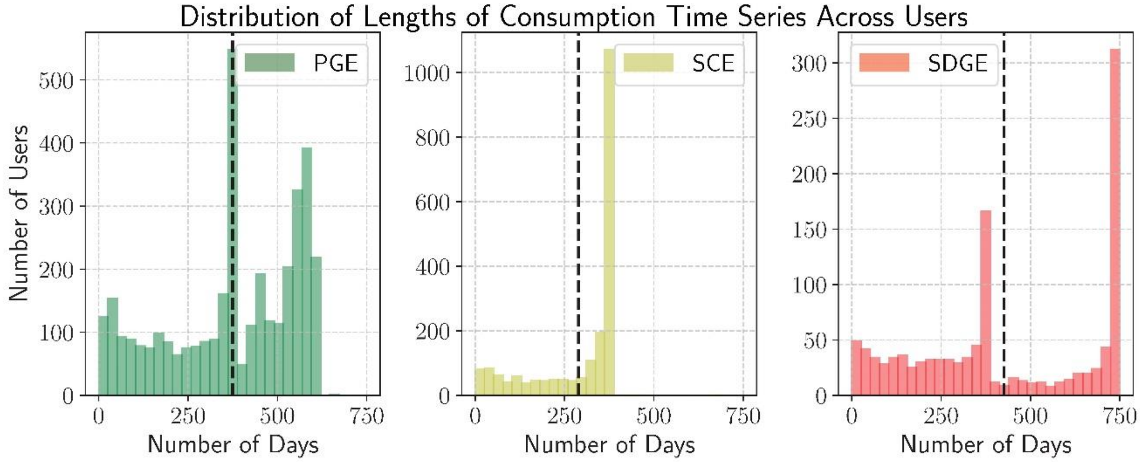


Figure E-9 provides a scatter plot of the geographic distribution of control, encouraged, and non-encouraged users broken out by electric utility. As expected, most users are concentrated in the urban areas of the San Francisco Bay Area, San Diego, and Los Angeles. It is visually striking that there appear to exist no structural differences in the distribution of users across either treatment group or electric utility, which is an intuition to be confirmed in the balance checks provided in Chapters 3 and 4.

Figure E-9: Geographic Distribution of Enrolled Users

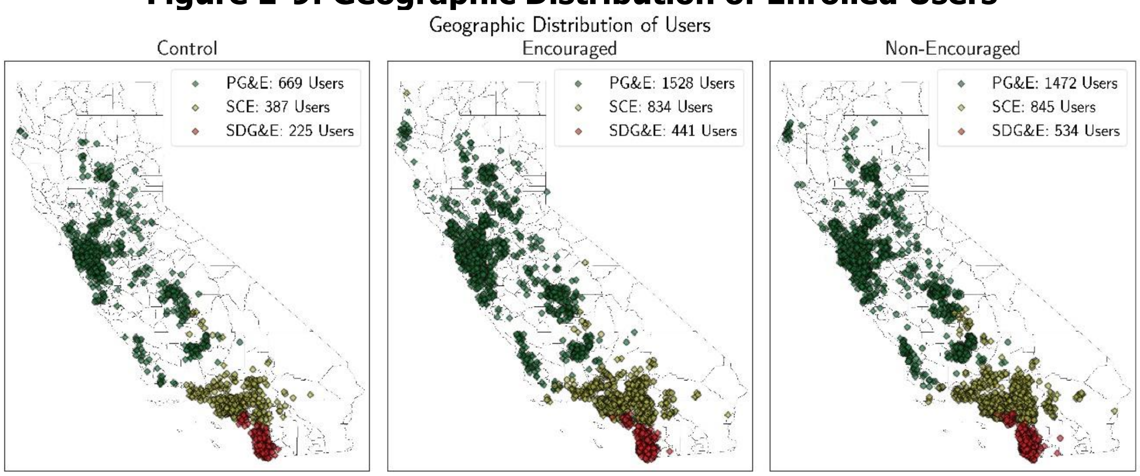
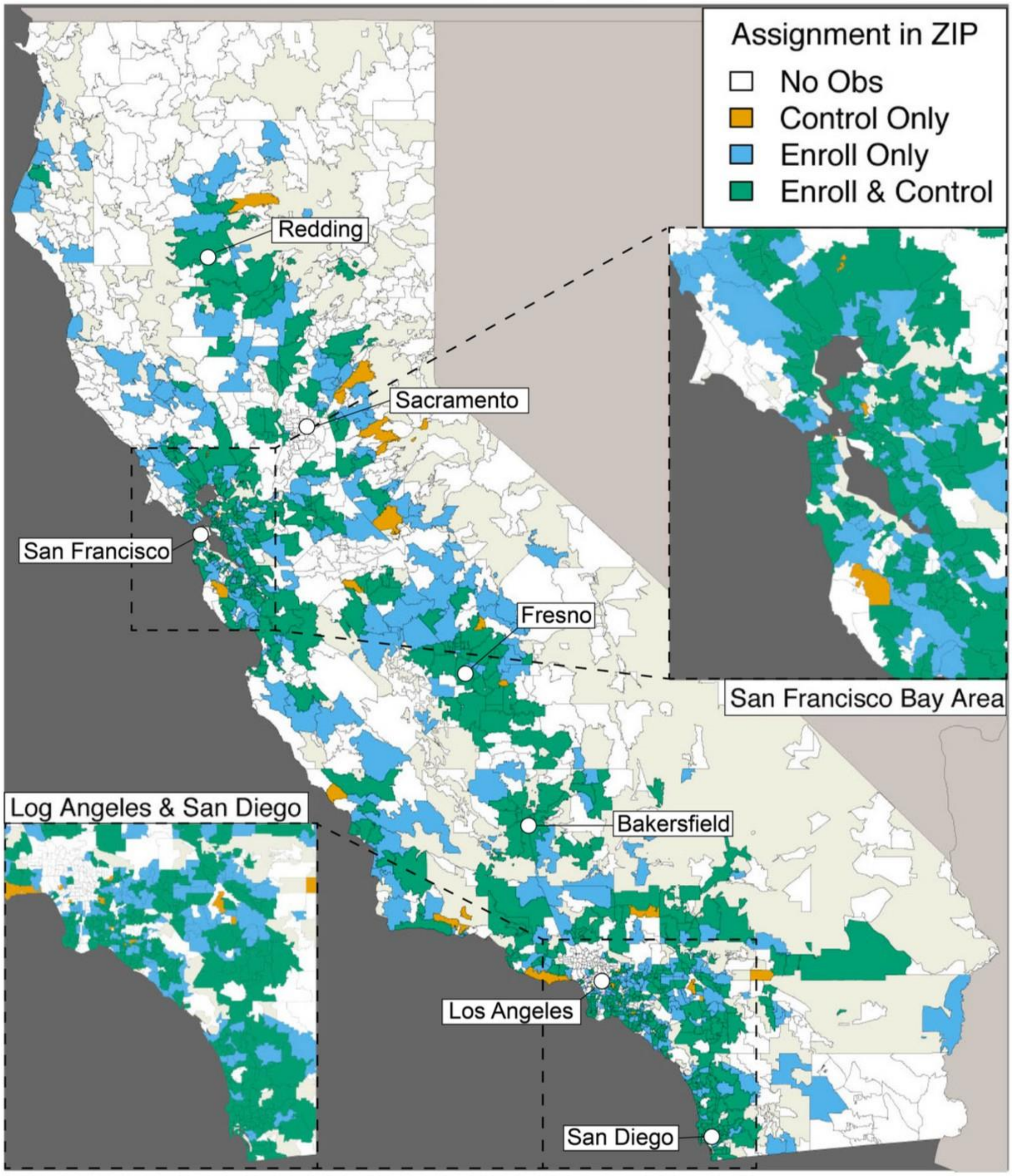


Figure E-10 shows the map of initial assignment by ZIP code. Each ZIP is colored orange if there are only Control users, blue if there are only enrolled (E or NE) users, and green if there are both enrolled (E or NE) and Control users. The map shows the assignment looks to have no significant spatial correlation in the assignment.

Figure E-10: Map of Assignment by ZIP Code



2.3 Initial Assignment Statistics

In this subsection, we provide data on the distribution of users across the experimental groups.

Table E-1 shows the total assignment numbers during the study period.[1] The assignment ratios are very close to the intended design of 40/40/20 for E/NE/C. The number of users with data is shown in the third column to be 6,227. The fifth column shows the numbers of users in the WP sample is 5,531, which drops users with erroneous meter data or those that appear to have on-site generation (net metering).

[1] This constitutes dropping 751 users who no populated join date ("NA") and 2,157 who joined during the pilot period.

Table E-1: Treatment Assignment Counts and Ratios for Study Period (1/1/17 – 8/14/17)

Treatment Group	Recruited	Fraction of Total	Enrolled with Data	Fraction with Data	Enrolled in WP Study	Fraction in WP Study
Control (C)	2,759	20.0%	1,166	18.7%	1,035	18.7%
Non-encr. (NE)	5,533	40.2%	2,507	40.3%	2,271	41.1%
Encouraged (E)	5,490	39.8%	2,554	41.0%	2,225	40.2%
Total	13,782	100%	6,227	100%	5,531	100%

Table E-2 shows there is a small amount of differential attrition, about 4%, in the Control group versus the treatment groups between the point of recruitment (making an account with OC) and connecting their utility data. Table 2.3.2 also shows there is no attrition between the enrollment and being included in the study sample. There is no differential attrition between the NE and E groups. Selection concerns cannot be ruled out due to the attrition, but the tables shown below indicate there is no observable difference between treatment groups.

Table E-2: Sample attrition for study period (1/1/17 – 8/14/17)

Event Incentive Level	Pr(Connected if Recruited)	Pr(In WP Study if Connected)
Control (C)	0.417*** (0.009)	0.899*** (0.009)
Non-encouraged (NE)	0.043*** (0.012)	-0.005 (0.011)
Encouraged (E)	0.034*** (0.012)	0.002 (0.011)
p-value for H_0: NE = E	0.39	0.44
Total	13,782	6,227

2.4 Dispatch Statistics

In this subsection, we illustrate the distribution of Demand Response events along various

Table E-3 shows summary statistics corresponding to the number shown in Figure 2.4.1. The numbers are summarized over user-level observations of the number of messages sent per 90 days for users with at least 90 days in each phase. During Phase 1, users were sent an average of 25 messages per 90 days with a median of 28 and a standard deviation of 7.6. During Phase 2, users were sent an average of 22 messages per 90 days with a median of 26 and a standard deviation of 9.4. During Phase 3, users were sent an average and median of 15.7 messages per 90 days with a standard deviation of 7.5. The numbers also show there are around 20% of users who received zero messages for the 90 day periods.

Table E-3: Messages per 90 days by user by experimental phase

(1/1/17 – 11/12/17)

Experimental Phase	Mean	Std. Dev.	Minimum	25th percentile	Median	75th percentile	Maximum
Phase 1	25.1	7.6	0	25	28	29	31
Phase 2	21.8	9.4	0	20	26	29	30
Phase 3	15.7	7.5	0	11.6	15.7	23	28

Figure E-11 shows the distribution of the number of dispatched Phase 1 DR events across all treatment users. The mean number is 25.10. Only ~5% of all treatment users were contacted less than 10 times during the entirety of Phase 1.

Figure E-11: Distribution of DR Dispatches in Phase 1 Across Users

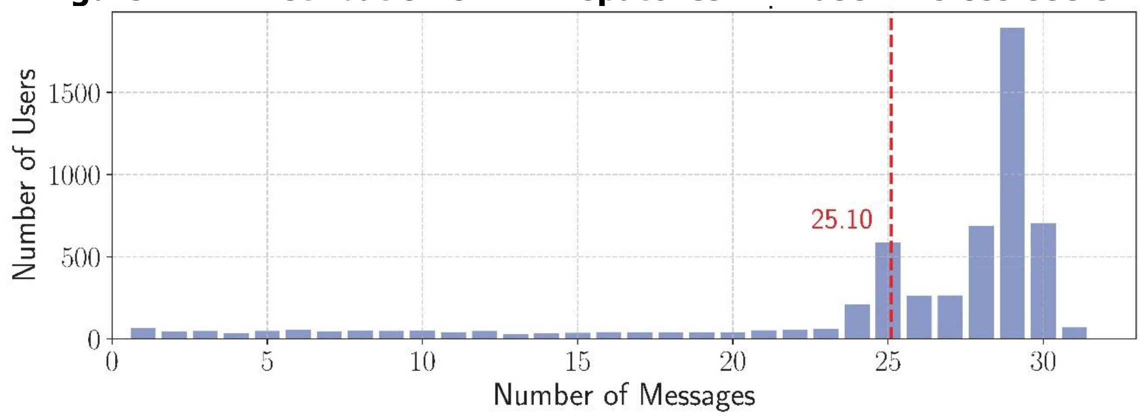
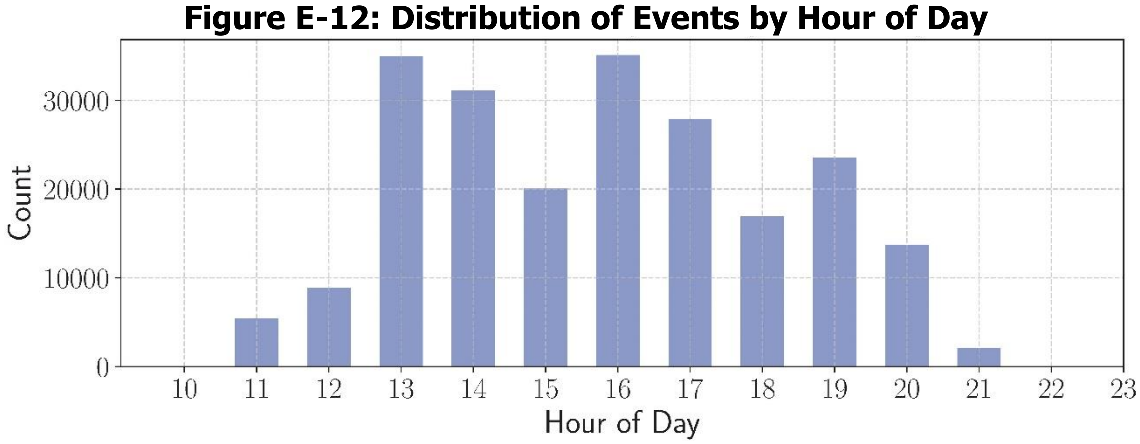


Figure E-12 shows the distribution of DR events in Phase 1 across all users by hour of the day. As can be seen from the figure, most events occurred in the late afternoon and early evening.

Figure E-12: Distribution of Events by Hour of Day



Similarly, Figure E-13 and Figure E-14 show the distribution across day of the week and month of the year.

Figure E-13: Distribution of Events by Day of Week

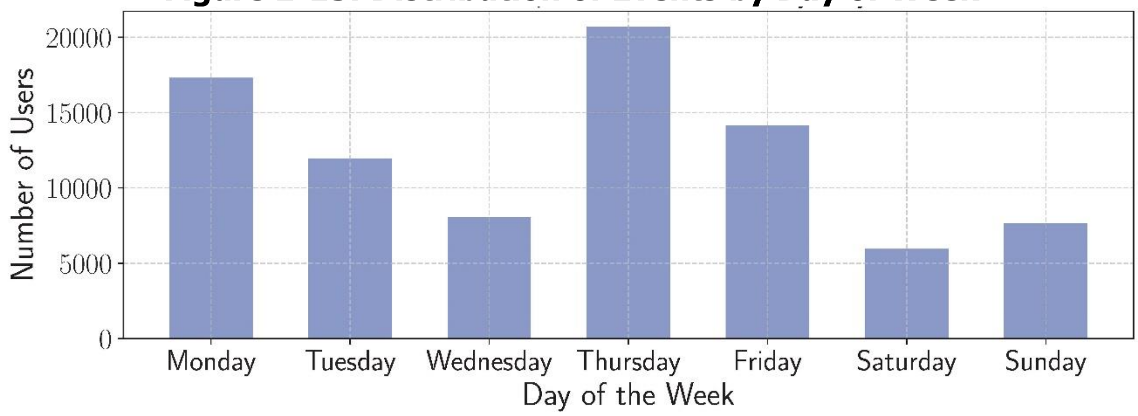


Figure E-14: Distribution of Events by Month of Year Users

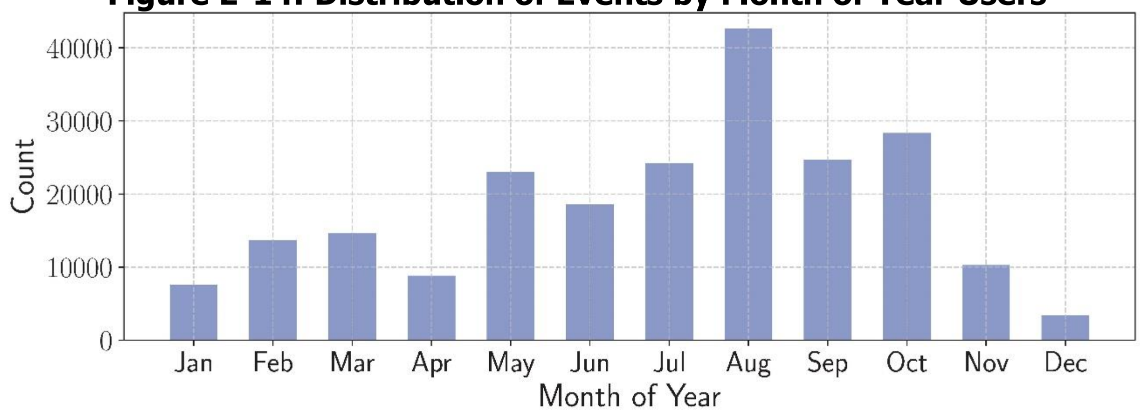


Figure E-15 shows the number of hours between consecutive Demand Response Events. We will later see (see Chapter 3) that a minimum gap of 18 hours is necessary for our estimation framework, as we wish to avoid “spillover” effects that might occur from one event to the next. Less than 3% of all events occurred within 18 hours. These events were dropped from the analysis.

Figure E-15: Distribution of Times between Consecutive Demand Response Events

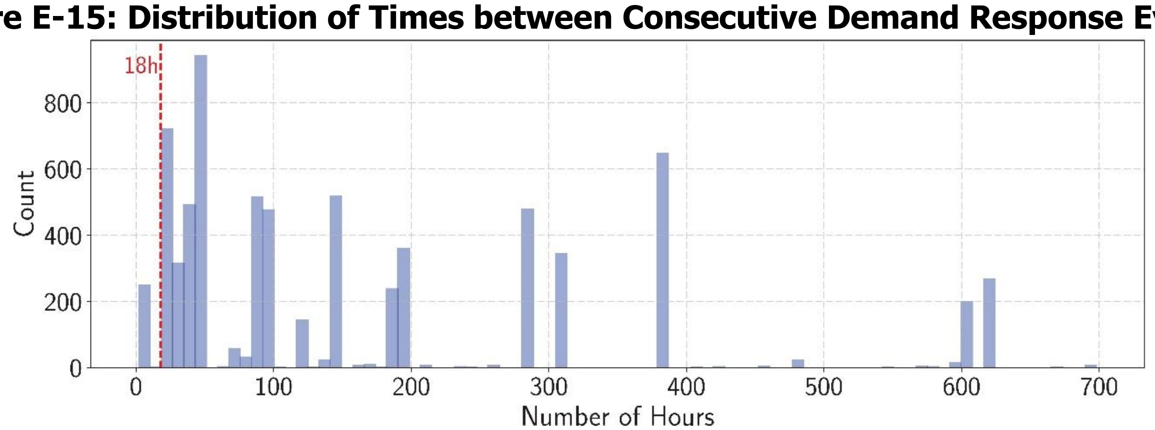


Table E-4 shows the incentive level assignment ratios for Phase 1. The second column reports the number of messages sent across all users in the study sample in the E and NE groups for each incentive level during Phase 1 of the experiment. These empirical assignment ratios are very close to the intended equal likelihood of receiving each of the incentive levels (20%). In total, there were 122,833 events called across 4,496 users across 94 unique events. Not every user was called during each event, only around 89% received a message when an event was called.

Table E-4: Incentive level assignment count and ratios for Phase 1 (1/1/17 – 11/12/17)

Event Incentive Level	Messages Sent	Fraction of Total
5 points per kWh	24,397	19.8%
25 points per kWh	24,525	20.0%
50 points per kWh	24,500	20.0%
100 points per kWh	24,688	20.1%
300 points per kWh	24,723	20.1%
Total	122,833	100%

Table E-5 shows the incentive level assignment ratios for Phase 2. The second column reports the number of messages sent across all users initially assigned to treatment in the study sample and assigned to the HT targeted group. The third column reports the number of messages sent to the LT targeted group. The fourth column reports the messages sent to the non-targeted (NT) group. The assignment probabilities were 50/50 and 33/33/33 for the HT and LT groups and 20 percent for each incentive level in the NT group. The empirical assignment ratios replicate this closely indicating the treatment was implemented according to the design at the time this report was written. In total, there were 71,464 events called during Phase 2 with the numbers being evenly split between the targeted and the NT groups.

Table E-5: Incentive level assignment count and ratios for Phase 2 (2/13/17 – 11/12/17)

Event Incentive Level	Messages Sent to HT	Messages Sent to LT	Messages Sent to Non-Targeted
5 points per kWh	-	5,976 (33.0%)	6,991 (19.6%)
25 points per kWh	-	5,921 (32.7%)	7,093 (19.9%)
50 points per kWh	-	6,215 (34.3%)	7,189 (20.2%)
100 points per kWh	8,837 (49.9%)	-	7,277 (20.4%)
300 points per kWh	8,856 (50.1%)	-	7,109 (19.9%)
Total	17,693 (100%)	18,112 (100%)	
	35,805		35,659 (100%)

Table E-6 shows the message level assignment ratios for Phase 3. The second column shows the number of messages sent in each group and the third column shows the fraction of total messages. The intended assignment ratio is 25 percent each group, which is closely replicated with slightly more being assigned to the event-level control group. In total 38,634 messages had been sent as of the time this report was written. These spanned 53 unique events and

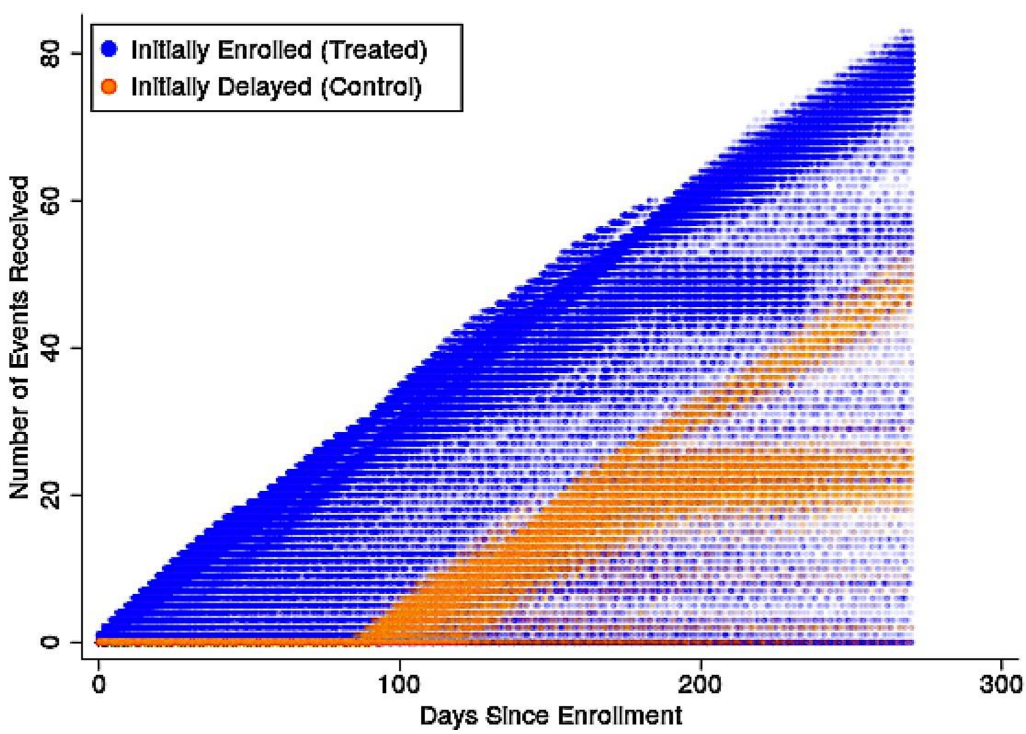
1,535 users. The user number is lower than the Phase 1 and Phase 2 numbers because Phase 3 occurred after Phase 1 and 2 and was not complete at the time this report was written. Phase 3 is due to be completed on 12/31/17.

Table E-6: Incentive level assignment count and ratios for Phase 3 (5/18/17 – 11/12/17)

Event Message Level	Messages Sent	Fraction of Total
Control Group	10,315	26.7%
Moral Suasion Only (M)	9,514	24.6%
Price Only (P)	9,354	24.2%
Both Moral Suasion & Price (B)	9,451	24.5%
Total	38,634	100%

Figure E-16 summarizes the distribution of messages sent per user over time by showing the number of messages received for each user as a function of the number of days since enrollment. The figure groups E and NE users in blue as "Initially Enrolled" and the Control group as "Initially Delayed" in orange. It is clear that the C group did not receive any messages for the first 90 days as is consistent with Phase 1. After 90 days C group users begin receiving messages, indicated by the increasing shading. Each point represents a single user and points are plotted with transparent shading so that darker parts of the plot indicate more users. The majority of users receive a consistent amount of 25 events per 90 days enrolled. The flattening of the lines indicates some users stopped receiving messages after about 180 days in the treated and 90 days in the control group.

Figure E-16: Message Count per User versus Days Since Enrollment



CHAPTER 3:
Phase 1 – Monetary Incentives and Automation

Phase 1 investigated two primary questions: 1) How do participants respond to varying monetary incentives during DR events and 2) how does adopting automation affect those responses. This section describes the research design in further detail and reports the empirical results. It also explores how these effects vary by time of day and month of year.

3.1 Research Design and User Experience

The main dimension of randomization in Phase 1 was the assignment to receive DR messages as compared with the 90-day delayed Control Group which received no messages. The control group received an email with the following message: "Due to overwhelming demand for our service, there will be a delay before we can send you #OhmHours. We estimate this delay will last approximately 3 months. In return for your patience, we'll issue you an extra $10 bonus when your account delay is over."

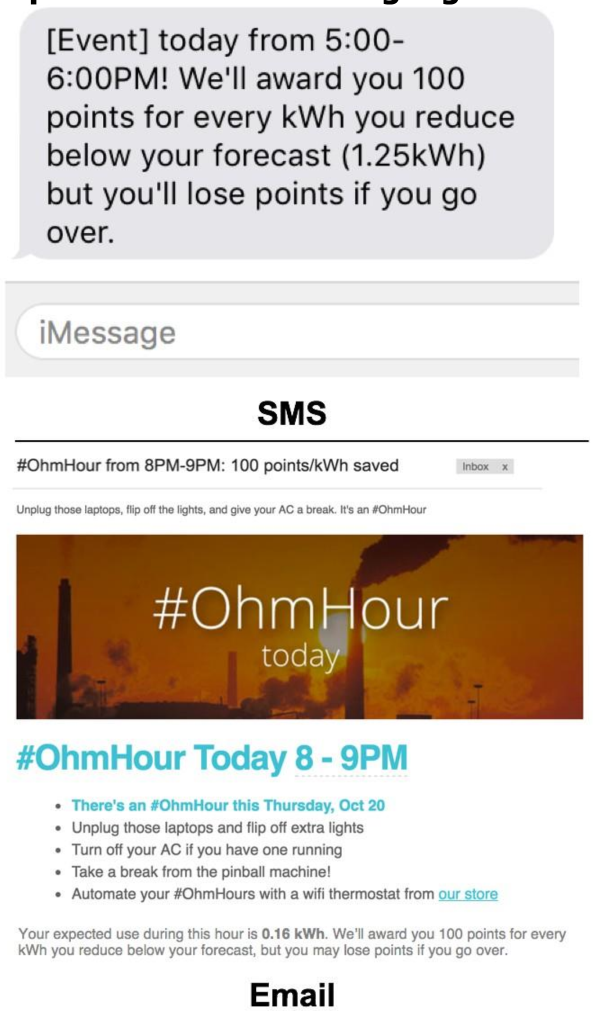
Figure E-17 shows sample event language for the Phase 1 DR events as experienced by Standard (also interchangeably referred to as Non-Encouraged) and Encouraged respondents. This language was consistent across all Phase 1 messages. The only difference was that during Phase 1 DR events, the incentive level was randomized between levels of 5, 25, 50, 100 and 300 points per kWh. The example figures show language for an event of 100 points per kWh. The point reward for each event was calculated as follows:

$$Reward = Incentive * (Forecast - Consumption)$$

The term Incentive was randomized among the 5 incentive levels with equal probability.

The forecast was calculated consistent with the CAISO 10-in-10 methodology. The methodology could generally be replicated by the evaluation team, but did show some non-systematic differences. These should not affect the estimation of the causal effects reported below, but is worth noting.

Figure E-17: Sample SMS and Email language for Phase 1 DR Events



The other dimension of randomization during Phase 1 was assignment to an automation encouragement. In order to measure the causal effect of adopting an automation technology, the Encouraged households were offered a rebate for the full purchase price of a new smart home device up to $240 in value. Upon creating an account, these households were shown a pop-up notification on the web-portal in addition to being sent an email notifying them they had been selected to receive a rebate for purchasing a new smart device. The household was offered a choice between 3 smart thermostats ranging in retail prices from $198 to $240 or one package of two smart plugs with a retail price of $80. The households were told that they would have the purchase price equivalent of points added to their balance when they connected the device as to ensure rebates encouraged automation

3.2 Balance Check

This section reports balance statistics for the Phase 1 assignment. Given constraints to the enrollment pipeline, we were unable to stratify assignment to ensure balance a priori.

This is a challenge to the design pursued here. Further, since the evaluation team did not perform the randomization, this check is vital to ensuring experimental validity.

Figure E-18 reports the average electricity consumption by hour-of-day in the 90 days prior to enrollment. The Standard (Non-Encouraged) and Encouraged are shown to be overlapping quite nicely, but the Control group clearly consumes more in a systematic fashion. Averaging across these delivers a statistically insignificant difference, but the figure justifies the use of the difference-in-differences estimation strategy described below to ensure cross-sectional differences are not interpreted causally. Figure E-19 performs the same exercise for average temperature and shows a more overlapping assignment. Figure E-19 shows the distributions of the length of historical data for control and treatment (Standard and Encouraged) groups are also generally overlapping with control households showing on average longer pre-period series.

Figure E-18: Electricity consumption by hour-of-day by Phase 1 Assignment

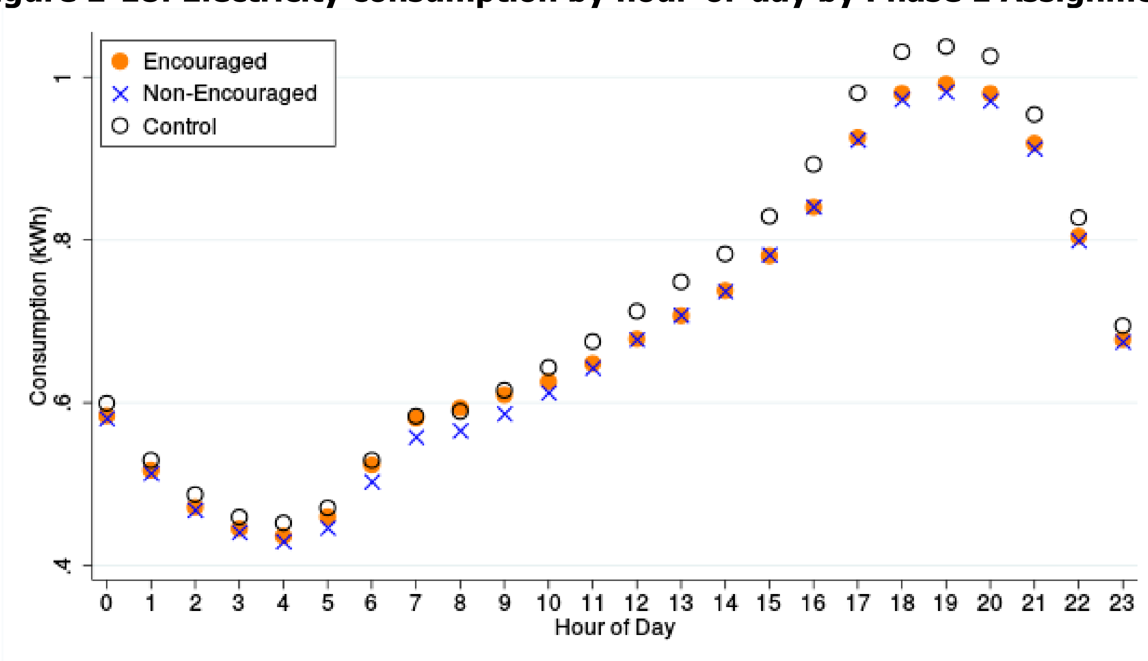


Figure E-19: Average temperature by hour-of-day across by Phase 1 treatment

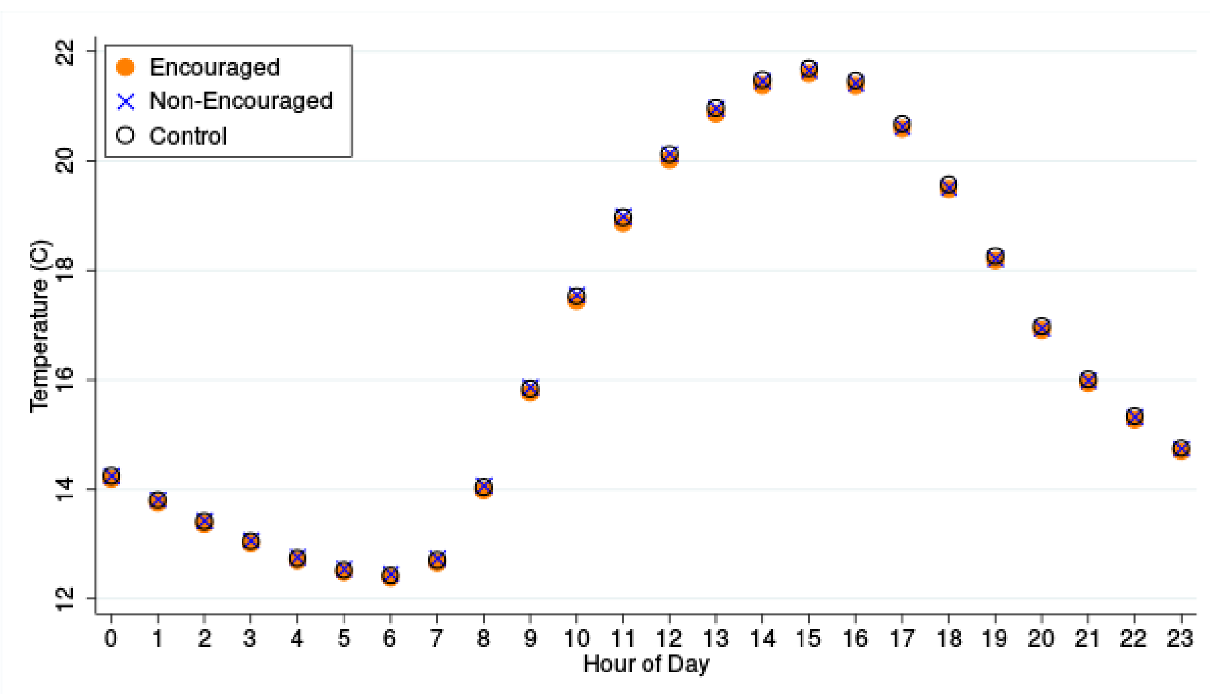


Table E-7 reports the full balance check. We test for balance in observables characteristics between treatment groups to provide evidence on the validity of assignment. Columns (1), (2), and (3) report means and standard deviations in parentheses for the Control, Standard, and Encouraged groups calculated using pre-enrollment data. Columns (4)-(6) report p-values on the t-test on the difference in means permuted between each of the treatment groups. Standard errors for the balance tests are assumed to be independent between households. Rows 1-6 report statistics for the data used in the empirical analysis and row 7 and above for the census variables, where available. Only one comparison is significantly different at the 10 percent significance level: the maximum consumption between the Encouraged and the Control group. This provides strong evidence that the assignment was random.

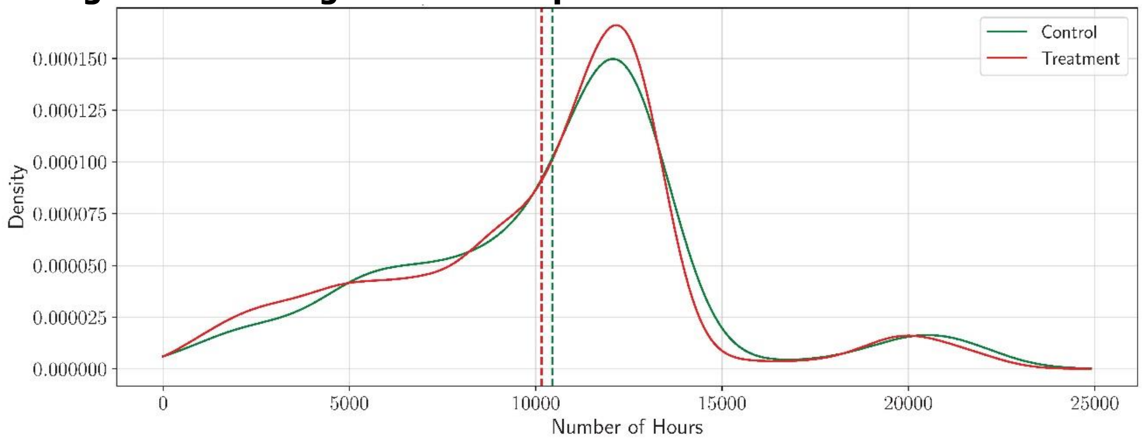
Table E-7: Statistical Checks of Phase 1 Randomization

	Control Group	Treatment Groups Standard	Encouraged	p-value on t-test with H_0: $\mu_S = \mu_C$	$\mu_E = \mu_C$	$\mu_E = \mu_S$
	(1)	(2)	(3)	(4)	(5)	(6)
Energy & Weather Data						
Daily Consumption (kWh)	16.5	16.0	16.0	0.23	0.21	0.94
	(11.2)	(10.8)	(10.7)			
Max Consumption (kWh)	5.0	4.9	4.8	0.14	0.09*	0.71
	(2.7)	(2.6)	(2.7)			
Hourly Outdoor Temp. (°C)	16.7	16.7	16.6	0.94	0.38	0.23
	(2.6)	(2.7)	(2.8)			
Mean Daily CDHs	33.8	34.0	33.4	0.90	0.69	0.51
	(30.6)	(31.8)	(31.7)			
Mean Daily HDHs	113.8	113.7	115.5	0.98	0.27	0.15
	(41.5)	(41.8)	(43.2)			
Pre-Enrollment Obs.	8755.6	8720.5	8686.5	0.83	0.68	0.79
	(4342.7)	(4362.4)	(4399.8)			
Demographic Data						
% HH Income < $25K	18.8	19.2	19.4	0.59	0.40	0.69
	(14.2)	(13.7)	(13.8)			
% Population Age 21-39	30.1	29.6	29.8	0.36	0.56	0.68
	(11.8)	(12.1)	(12.1)			
% Family HHs	66.3	66.1	66.7	0.85	0.66	0.42
	(19.1)	(19.0)	(18.5)			
% Population w/ Bachelors	22.8	22.5	21.8	0.68	0.14	0.18
	(13.0)	(13.0)	(12.4)			
Median Year Built	1973.5	1973.5	1973.0	0.92	0.54	0.50
	(17.2)	(16.9)	(16.9)			
% Renters	48.7	48.3	48.4	0.79	0.81	0.98
	(26.7)	(26.0)	(26.3)			
% HHs w/ 3+ bedrooms	53.0	52.4	52.8	0.72	0.88	0.78
	(30.3)	(29.7)	(29.1)			
% Detached Units	54.2	54.2	54.4	0.99	0.91	0.89
	(33.9)	(33.0)	(32.2)			
% Electric Heating	28.2	27.9	27.2	0.80	0.31	0.34
	(17.3)	(17.8)	(17.0)			
Median Monthly Rent ($)	1344.1	1361.8	1344.3	0.54	0.99	0.44
	(518.3)	(520.7)	(519.5)			
Households/Observations	1,043	2,246	2,202			

Standard deviations in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Figure E-20: Length of Consumption Histories Before Enrollment



3.3 Phase 1 Empirical Strategy

We use a difference-in-differences estimation strategy to evaluate the effect of DR events. This involves controlling for between-household differences by including household by hour-of-day fixed effects and data from the 90 days prior to enrollment. We also include hour-of-sample fixed effects to take out seasonal shifts in consumption across all treatment and control groups. We also include temperature controls to improve the precision of our estimates. The specific estimating equation is:

$$Y_{@AB} = \delta_{\mathrm{DEDFA}} EventHour_{@A} + \delta_{\mathrm{FJFDEDFA}} NoEvent_{@A} + \alpha_{@B} + \gamma_{A} + \beta_{0} CDH_{@A} + \beta_{Q} HDH_{@A} + \eta_{@AB}$$

Where the indices are *i* = household, *t* = hour-of-sample, *h* = hour-of-day (redundant with t). The outcome variable is either kWh consumed during that hour or the log of it to provide estimates that are approximate to percent changes. The *EventHour* variable indicates an event hour occurred (after enrollment) for that household and the *NoEvent* variable indicates the household was enrolled, but no event occurred during that hour.

$\alpha_{@B}$ are household by hour-of-day fixed effects and γ_{A} are hour-of-sample fixed effects. We use linear parametric controls in terms of cooling and heating degree hours, defined as positive deviations above and below 65 degrees Fahrenheit, respectively. There was a slight non-compliance issue in the assignment of hours because around 10% household opted-out of receiving events during certain hours of the day. To address this selection issue we instrument EventHour with Enrollment interacted with EventHour, but it does not substantively affect the estimates.

To estimate the effect of automation, we leverage the encouragement design. This involves estimating an instrumental variables (IV) approach as follows. We estimate the same difference-in-differences style approach among the Standard and Enrolled households, excluding the Control Group:

$$Y_{@AB} = \delta_{\mathrm{STAJDEDFA}} AutoxEventHour_{@A} + \delta_{\mathrm{STAJFJFDEDFA}} AutoxNoEvent_{@A}$$

$$+\alpha_{@B} + \gamma_{A} + \beta_{0} CDH_{@A} + \beta_{Q} HDH_{@A} + \eta_{@AB}$$

AutoxEventHour indicated the household was automated and AutoxNoEvent indicates hours after enrollment for automated households when no event was called. The outcome variable, controls, and fixed effects are defined the same as above. The key difference in the IV approach is the Auto variables are instrumented using the Encouragement assignment and 2SLS.

3.4 Phase 1 Experimental Results

3.4.1 Average Effect of DR Events

Table E-8 reports the results for estimating the average effect of being enrolled in the program during an event hour and the effect of being called. The first two columns report results in terms of kWh and the third and fourth in terms of log which approximately represent percent changes. The results show statistically significant effects at the 99 percent level and estimates on the order of 0.12 kWh or 13 percent pooling across all incentive levels. The standard errors are clustered two-ways by household and hour-of- sample. Because only around 10 percent of household opted-out the OLS and the IV estimates are similar with the OLS being closer to 0 due to the averaging across households that were not called during that event hour.

Table E-8: Effect of DR Events

	kWh	kWh	log(kWh)	log(kWh)
Enrolled x DR Hour vs. Control (OLS)	-0.107*** (0.012)		-0.117*** (0.009)	
Called x DR Hour vs. Control (IV)		-0.120*** (0.014)		-0.132*** (0.010)
Enrolled x No DR vs. Control	-0.017 (0.006)	-0.017 (0.006)	-0.017 (0.006)	-0.017 (0.006)
Households	5,491	5,491	5,491	5,491
N (observations)	22,926,631	22,926,631	22,926,631	22,926,631

We also show the effect of events as an event-study that shows how consumption differed in the hours leading up to and following the event. Figure E-21 and Figure E-22 show the event-study for kWh and log(kWh), respectively, with zero indicated the DR event hour. The dashed line delineates the time before the household was notified and the two solid lines the event hour. Vertical bars indicate 95 percent confidence intervals and the estimates are normalized to period -2.

Figure E-21: Effect of DR Events in kWh

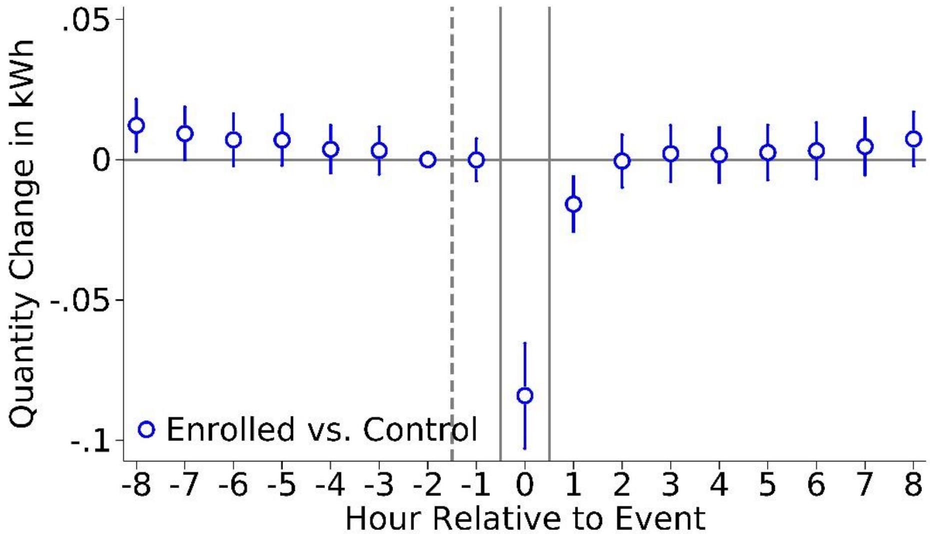
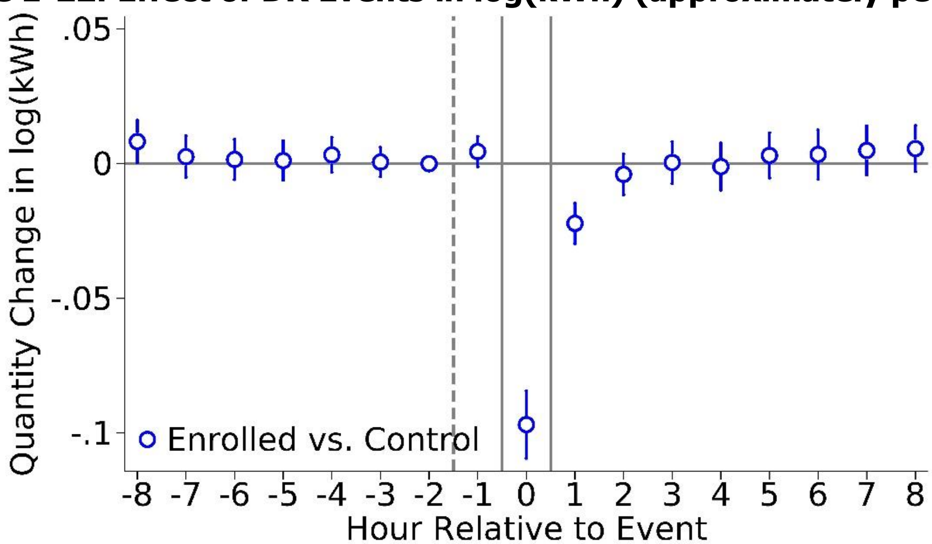


Figure E-22: Effect of DR Events in log(kWh) (approximately percent)



Automation

Before moving to the effect of adopting automation we also report differences in automated and non-automated household relative to control. These provide non-causal estimates on the difference between automated and non-automated DR events and serve to show how much of the average effects reported in Table E-8 are due to automation. Figure E-23 and Figure E-24 show the same event-study figures as above and indicate automated users reduced about 25 kWh and 25 percent of their consumption as compared with the non-automated who reduced around 0.07 kWh or 7 percent. This comparison cannot be interpreted causally because it compares households who chose to sign up for the automation service to those who did not.

Figure E-23: Automated vs. Non-automated DR Events in kWh

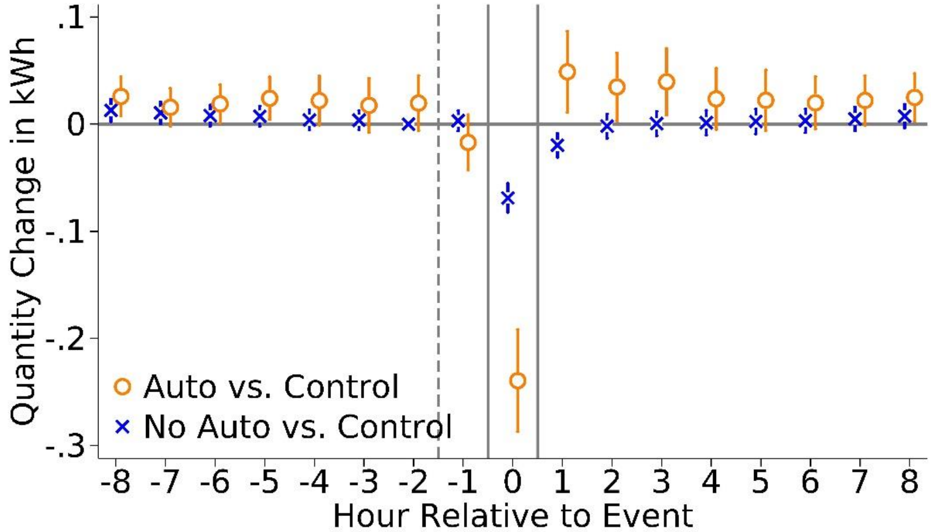
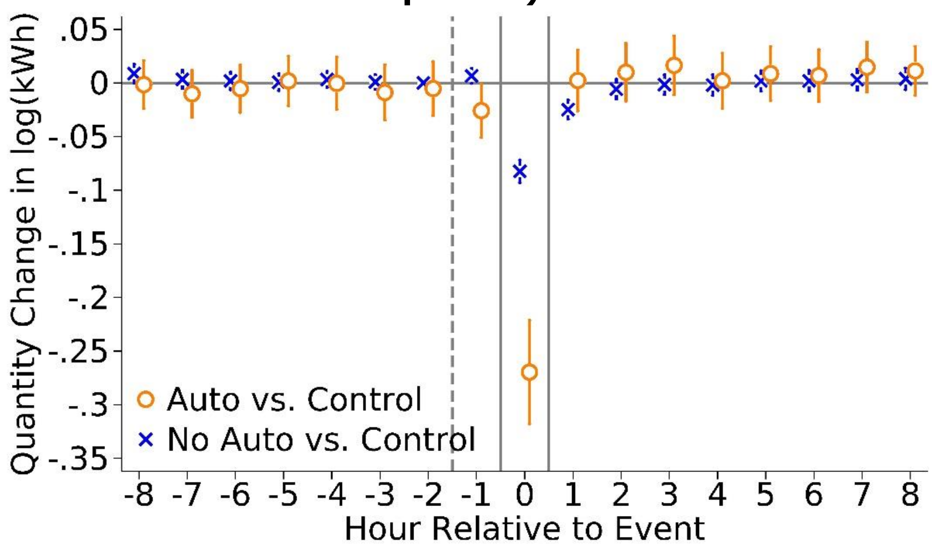


Figure E-24: Automated vs. Non-automated DR Events in log(kWh) (approximately percent)



The encouragement provides random variation that can be used to estimate the causal effect of automation. First we report the effect of the encouragement on the automation take-up. Figure E-25 plots the number of households and the number of devices for the Standard group (no rebate offer) and the Encourage group (rebate offered). Given the groups were roughly of equal size, the results show a sharp increase in the number of households with automation driven by thermostats and smart plug adoption.

Figure E-25: Effect of Rebate Encouragement on Automation Take-up

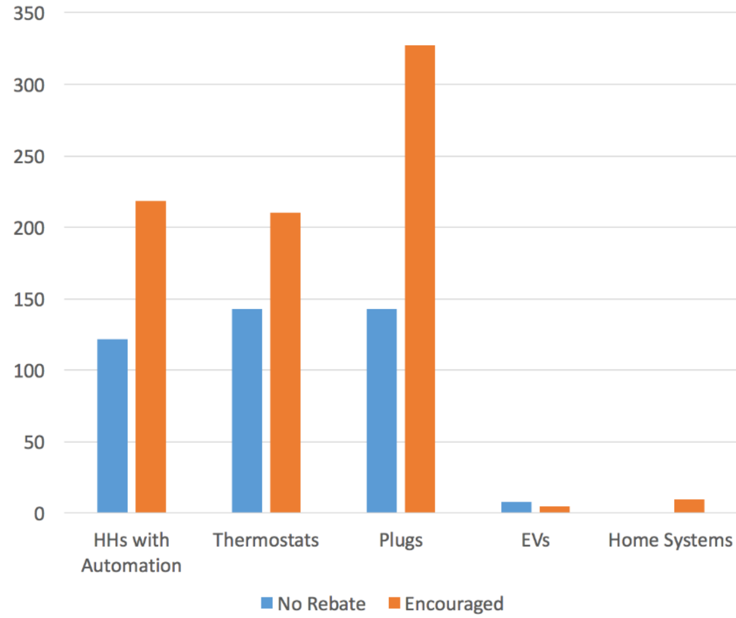


Table E-9: Summary of Encouragement Results

	Standard (S) (1)	Encouraged (E) (2)	Difference (E−S) (3)
Panel A: Automation Type and Counts			
Households	2,246	2,202	-
Households with at least one connected device	122	218	96
Total connected devices	295	554	259
Total connected thermostats (subsidized)	143	210	67
Total connected plugs (subsidized)	143	327	184
Total connected home systems (not subsidized)	0	10	10
Total connected electric vehicles (not subsidized)	8	5	-3
Panel B: Automation Take-up			
Households with any automation	0.053	0.097	0.045*** (0.008)
Take-up by consumption level:			
1^{st} quartile (0-0.34kWh)	0.022	0.103	0.080*** (0.015)
2^{nd} quartile (0.34-0.55kWh)	0.064	0.103	0.039** (0.016)
3^{rd} quartile (0.55-0.85kWh)	0.066	0.098	0.031* (0.016)
4^{th} quartile (0.85-6kWh)	0.060	0.091	0.031* (0.016)

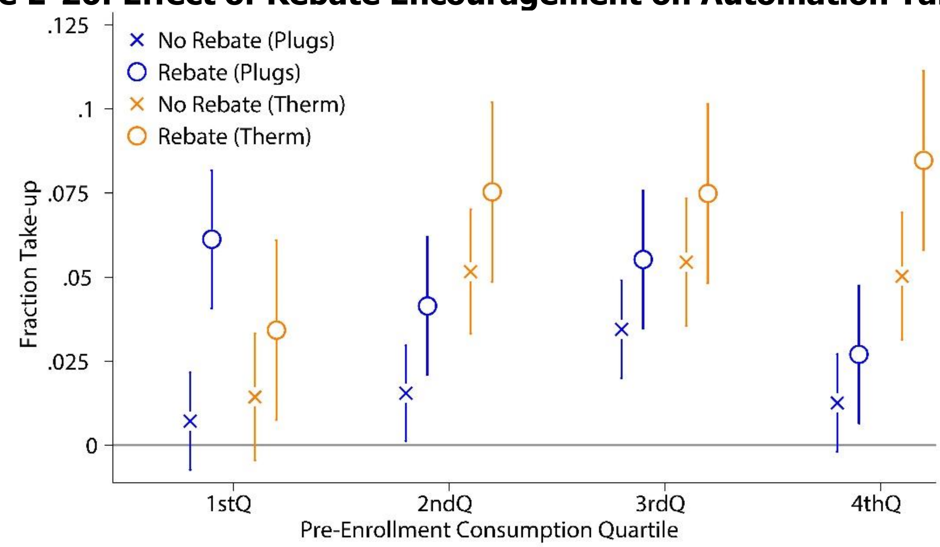
Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table E-9 summarizes the effect of the encouragement on automation take-up, reporting the number of devices in each of the Standard and Encouraged groups in Panel A. Panel B reports estimates of the instrument's "first stage", showing that the fraction of automation take-up increased a statistically significant amount of 4.5 percentage points (83 percent) over the baseline take-up of 5.3 percentage points.

Panel B also breaks out the take-up by the average consumption during the 90-day pre-period. This shows that the rebate was disproportionately taken-up by households in the bottom quartiles of consumption (0-0.34kWh). This has an important implication for interpretation. The households who are induced to take up automation by the rebate are observably different from those who sign up without the rebate. This means the causal adoption estimates we provide below may not be representative of causal estimates for the households who sign up without a rebate.

Figure E-26 plots the take-up by consumption quartile by technology and shows that the significant increase in the bottom quartile is driven by the adoption of plugs whereas the increase in the top quartile is driven by thermostats. These differences are intuitive if thermostats are typically used in larger households and plugs may be more appropriate for households without central heating/cooling.

Figure E-26: Effect of Rebate Encouragement on Automation Take-up



Using the adoption variation induced by the encouragement, we report the estimates of the causal effect of adopting automation on responses in Table E-10. The Results show the OLS non-causal estimates for comparison in columns 1 and 3. The IV estimates of the causal effect are reported in columns 2 and 4 and show that adopting automation yields an additional 0.567 kWh reduction or around 0.668 log points. Logs are a worse approximation of percent the further from 0 so the second estimate is approximately a 49 percent reduction. The results are statistically significant at the 99 percent level. Comparing to the OLS estimates shows that the marginal treatment effect is larger for those nudged into adopting by the rebate. One interpretation of this is that household who have already adopted have less flexible demand because they've already taken adaptation measures. Another interpretation would be individuals who have chosen to adopt without rebate are less inclined to take additional action. Important to note is that the adoption effect is in addition to what otherwise would have been done to respond, which may not be zero. Thus, the total DR event response for these households could be larger than the estimates provided.

Table E-10: Effect of Automation

	kWh	kWh	log(kWh)	log(kWh)
Auto x DR Hour vs. Non-Auto x DR Hour (non-causal OLS)	-0.190*** (0.028)		-0.222*** (0.029)	
Auto + DR Hour vs. Non-Auto x DR Hour (causal IV)		-0.567*** (0.165)		-0.668*** (0.160)
Auto x No DR Hour vs. Non-Auto x No DR Hour	0.021 (0.013)	-0.139 (0.106)	-0.005 (0.015)	-0.112 (0.127)
Households	4,448	4,448	4,448	4,448
N (observations)	18,559,930	18,559,930	18,559,930	18,559,930

Figure E-27 and Figure E-28 plot the same event-study style figure for the automation adoption and shows a similar pattern to those above for kWh consumed and log(kWh), respectively. A precise response during event periods that quickly reverts to past consumption levels.

Figure E-27: Effect of Adopting Automation on DR Events in kWh

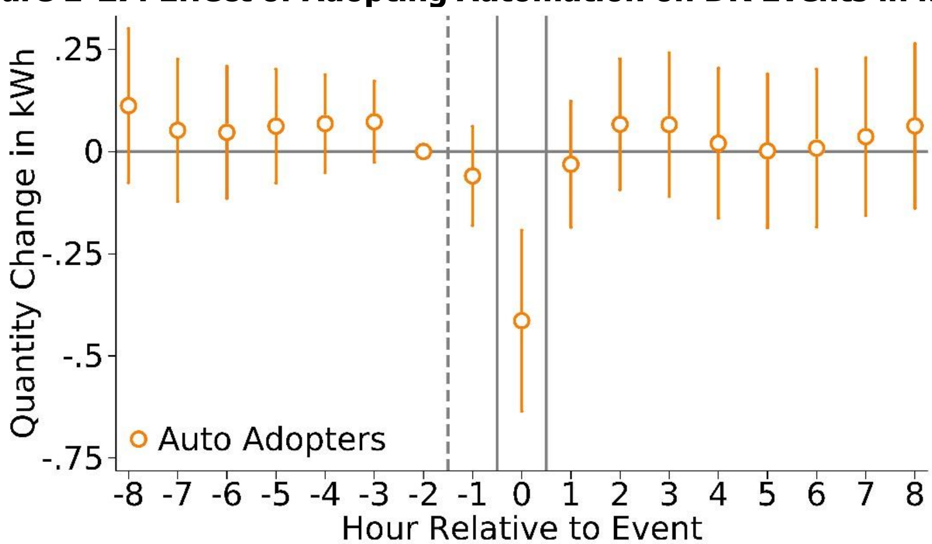
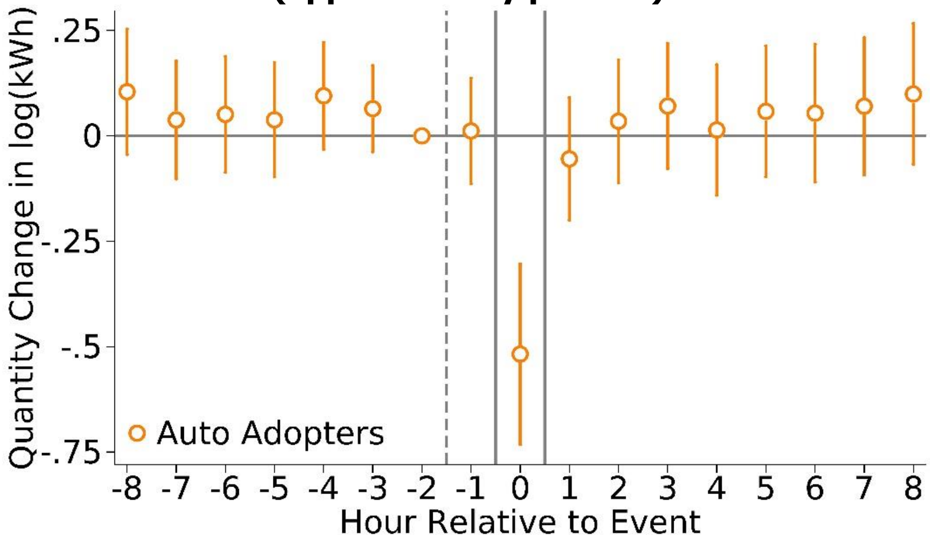


Figure E-28: Effect of Adopting Automation on DR Events in log(kWh) (approximately percent)



Effect of Difference Incentive Levels

We now report the effect of different incentive levels in the DR events. This is achieved by decomposing the average effects reported above by the incentive level. Because the incentive level was randomly assigned within each event and within individual over time, we can interpret the relative effect between incentive levels as the causal effect of changing the incentive level.

Figure E-29: Effect of Different Incentive Levels during DR Events in kWh

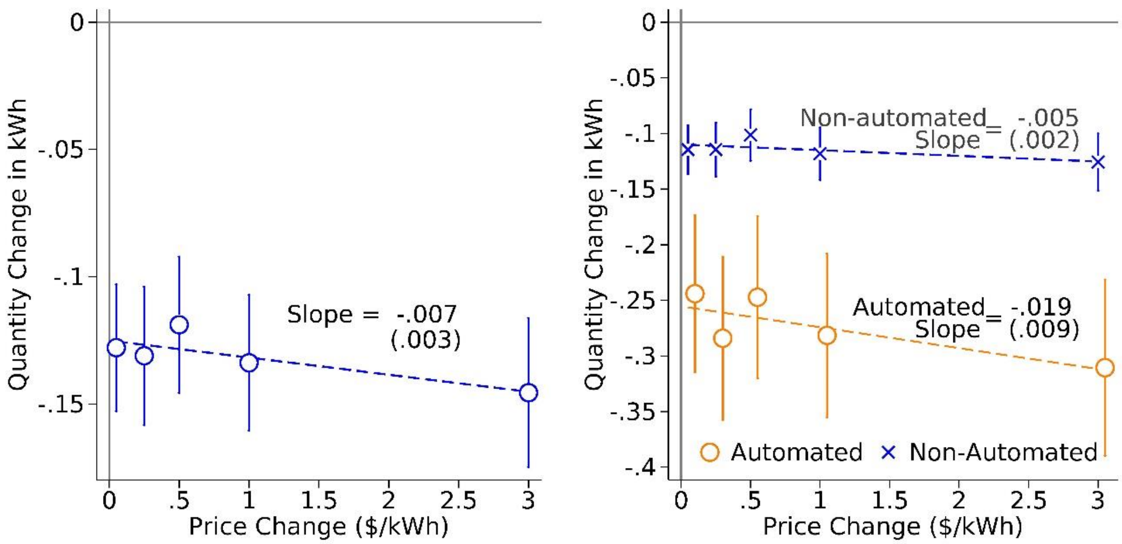


Figure E-30: Effect of Different Incentive Levels during DR Events in log(kWh) (approximately percent)

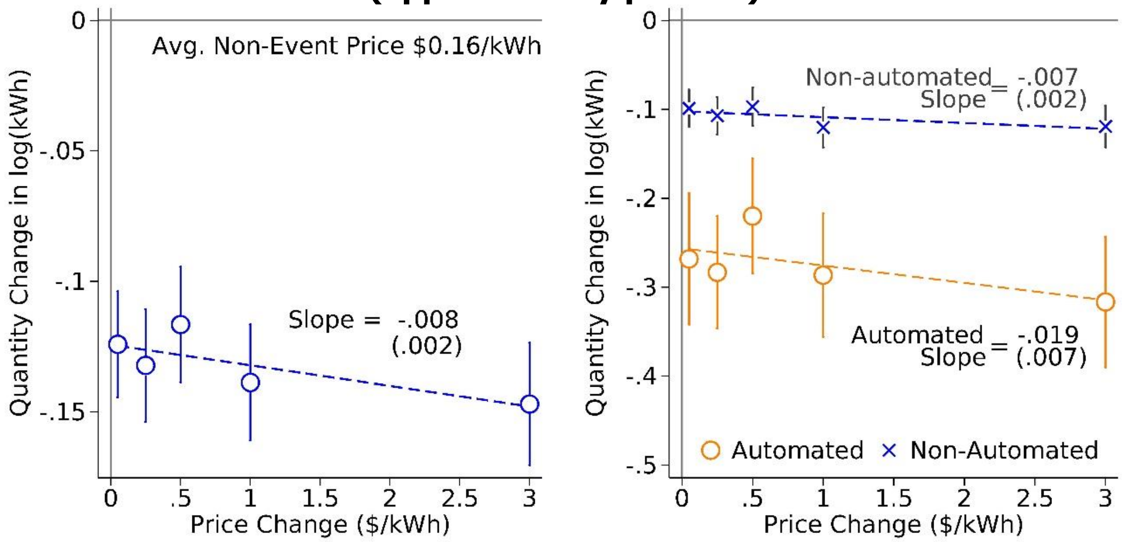


Figure E-29 and Figure E-30 plot the estimates of this effect by incentive level (written in $/kWh) for kWh consumed and log(kWh), respectively. Recall 100 points = $1 so the x- axis can be multiplied by 100 to get the points per kWh equivalent. The left panels in the figure show the causal estimates of each incentive level. Across both outcome measures, the results show a significant decrease in consumption of -0.125 kWh, around 12 percent for the 5 point per kWh ($0.05/kWh) incentive level and -0.145 kWh, around 14 percent for the 300 point per kWh ($3/kWh) incentive. We also report a linearly fitted slope through the incentive levels which is estimated to be negative and statistically significant from zero. However, the insensitivity of the response to the level is remarkable considering the change in incentive levels relative to the utility price. Table E-11 reports the estimates plotted in Figure E-29 and Figure E-30.

The right panels in Figure E-29 and Figure E-30 show the decomposition of the effect by users were automated to show that the automation was not driving the price insensitivity. In fact, automated users appear to be more price sensitive than their non-automated counterparts. The interpretation of the slopes in the right panel is not causal and the design was statistically underpowered to explore how price sensitivity changed due to automation adoption using the encouragement.

Table E-11: Effect of Varying Incentive Levels

	kWh	kWh	log(kWh)	log(kWh)
5 points per kWh ($0.05/kWh)	-0.128*** (0.013)		-0.124*** (0.010)	
25 points per kWh ($0.25/kWh)	-0.131*** (0.014)		-0.132*** (0.011)	
50 points per kWh ($0.50/kWh)	-0.119*** (0.014)		-0.117*** (0.011)	
100 points per kWh ($1.00/kWh)	-0.134*** (0.014)		-0.139*** (0.011)	
300 points per kWh ($3.00/kWh)	-0.146*** (0.015)		-0.147*** (0.012)	
Response Intercept		-0.125*** (0.010)		-0.124*** (0.009)
Response Slope per $0.01 in Incentive		-0.0067*** (0.0025)		-0.0079*** (0.0024)
Households	5,491	5,491	5,491	5,491
N (observations)	22,926,631	22,926,631	22,926,631	22,926,631

We explored whether the insensitivity was short-lived by estimating the slope and intercept parameters as shown by the dashed lines in the left panels of Figure E-29 and Figure E-30 for kWh consumed and log(kWh), respectively. Figure E-31 and Figure E-32 show these estimates broken out by 30-day period after enrollment. The left estimate in each panel is the pooled estimate as shown in Table E-11. The estimates do not change much over the 90 days period, indicating that familiarity with the program over a 90 day period does not substantively change the insensitivity result. Figure E-33 and Figure E-34 show the results broken out (non-causally) by automation to again show that the results are not driven by automated users.

Figure E-31: Demand Intercept and Slope Over Time in kWh

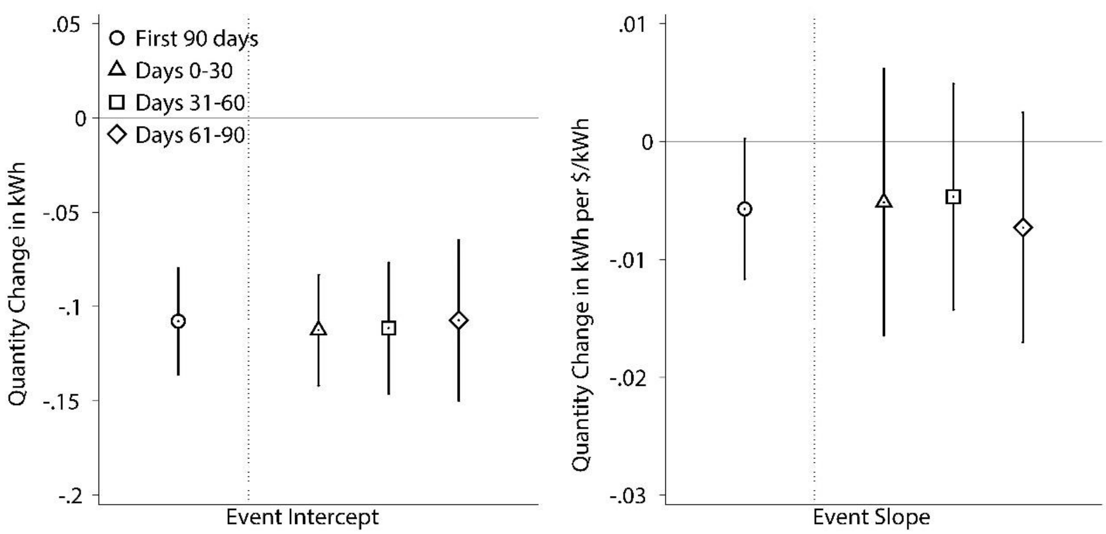


Figure E-32: Demand Intercept and Slope Over Time in log(kWh) (approximately percent)

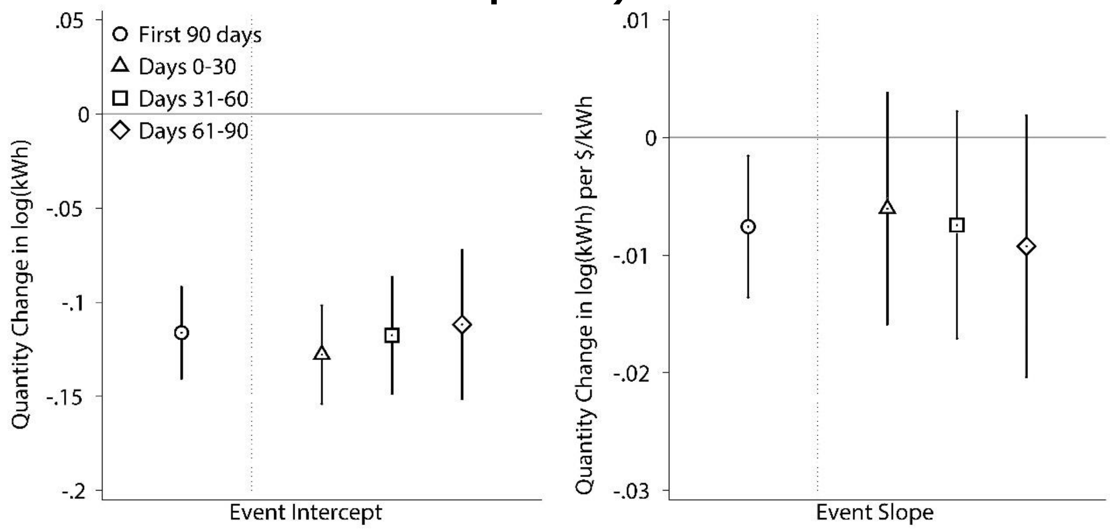


Figure E-33: Demand Intercept and Slope Over Time by Automated and Non-Automated in kWh

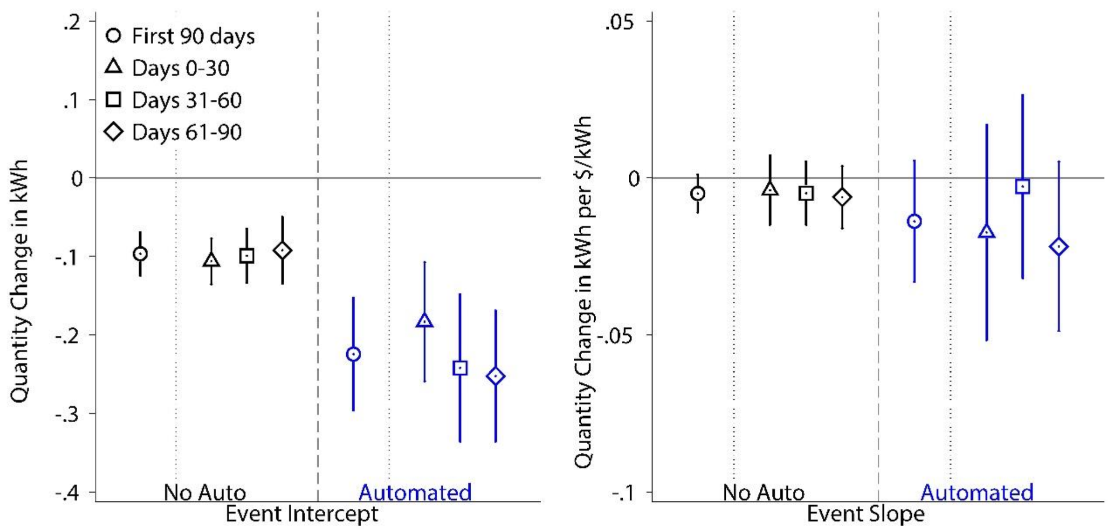
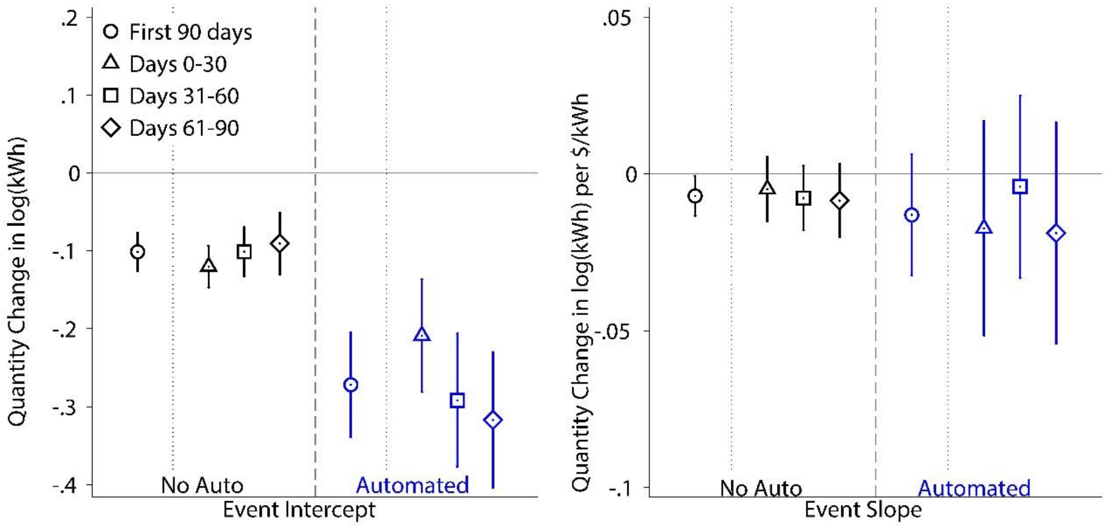


Figure E-34: Demand Intercept and Slope Over Time by Automated and Non-Automated in log(kWh) (approximately percent)

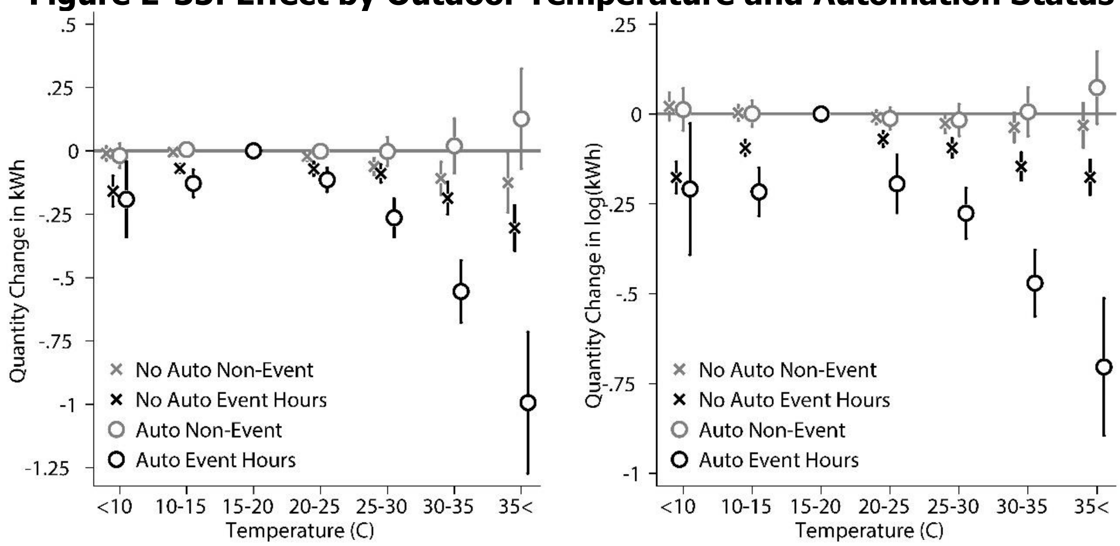


3.4.4 Effect Heterogeneity by Temperature, Hour-of-Day and Season

In this section we report heterogeneity in the average effect of a DR event by outdoor temperature, hour-of-day, and month-of-year.

For temperature, we interact the treatment indicator for an event hour or a non-event hour after enrollment with a set of temperature bin dummies that span 5 degrees Celsius between 10 and 35 degrees. This corresponds to temperatures ranging from 50 to 95 degrees Fahrenheit. The differences can be interpreted as the difference in effect relative to the omitted bin 15-20 degrees Celsius (59-68 degrees Fahrenheit). Figure E-35 plots the estimates broken out by automated and non automated households.

Figure E-35: Effect by Outdoor Temperature and Automation Status



The left and right panels show results in terms of kWh consumed and log(kWh), respectively. The rightmost estimates in each figure can be interpreted as follows. Automated households reduced consumption by 1 kWh or 0.75 log points (53 percent) more during DR events when the temperature was hotter than 35C relative to the baseline. Non-automated households also reduced significantly more on hot days, showing responses on the order of 0.25kWh and 0.2

log points (18 percent). We plot non-event hours as well and show that there is little evidence of spillover behavior that correlates with temperature. We might expect this if thermostat presets were being changed.We perform similar interactions for hour-of-day and month-of-year, both of which confirm that times of the day and times of the year with hotter temperatures tend to have larger effects. Figure E-36 and Figure E-37 show results for hour-of-day and Figure E-38 and Figure E-39 show results for month-of-year. For each figure the left panel reports results in kWh consumed and the right panel for log(kWh). Figure E-37 and Figure E-39 show automated and non-automated responses decomposed non-causally. Since the hour of the event was not randomized, comparing between hours should be performed with caution. Further, events may have been called at different times of the day for different times of the year. The results show a pattern that is consistent with the interpretation that cooling load drives larger responses.

Figure E-36: Effect of DR Event by Hour-of-Day

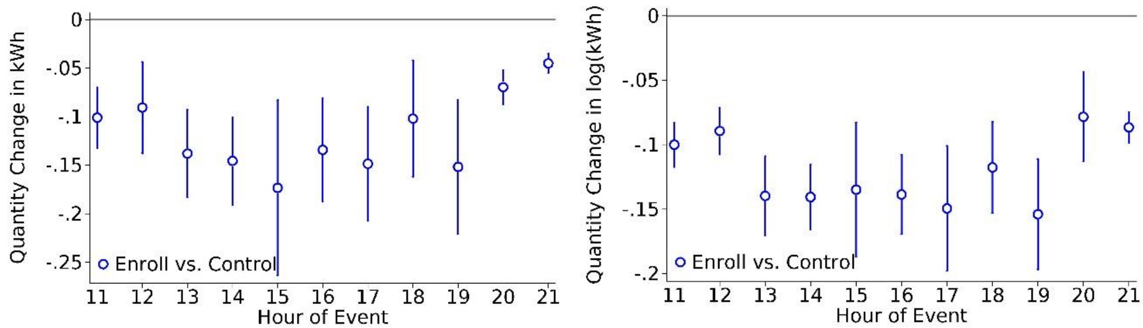


Figure E-37: Effect of DR Event by Hour-of-Day and Automation

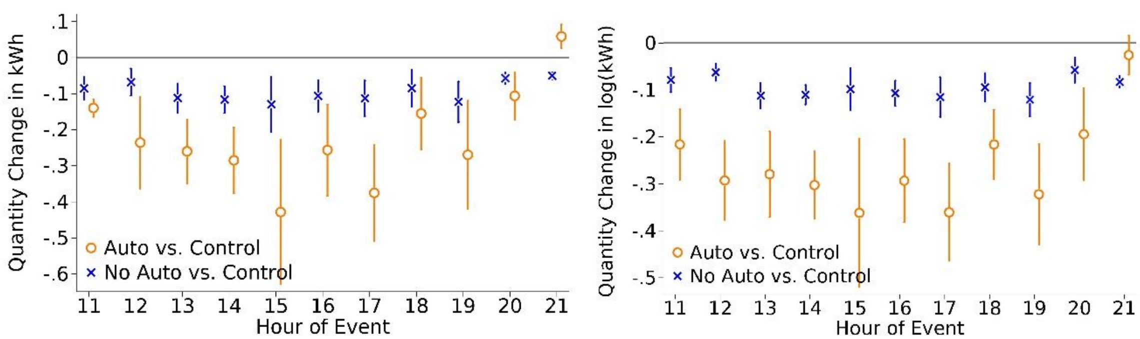


Figure E-38: Effect of DR Event by Month-of-Year

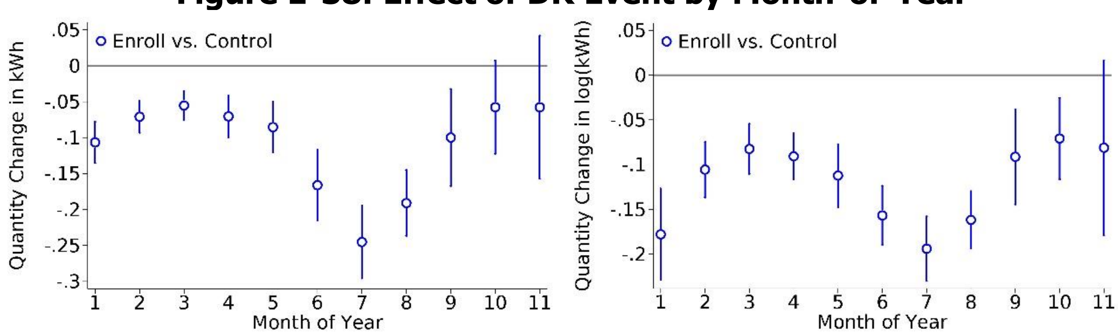
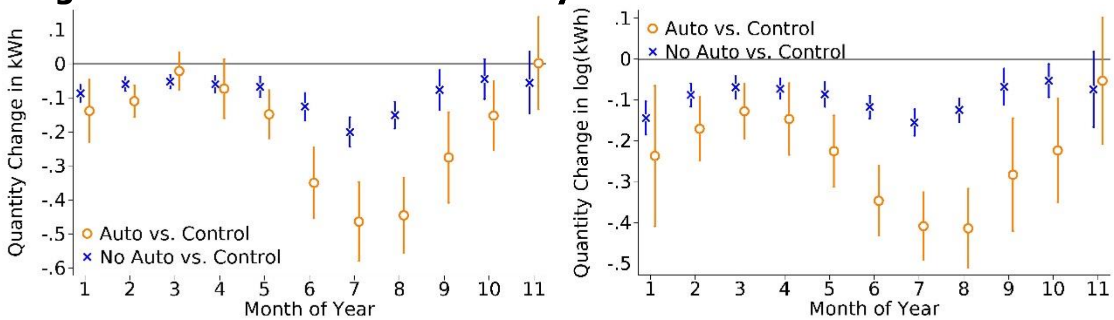


Figure E-39: Effect of DR Event by Month-of-Year and Automation



CHAPTER 4: Phase 2 – Targeting Household Incentives

Phase 2 investigated the question: Can the incentives sent during Phase 1 be modified and targeted to improve program cost efficiency? The targeting strategy studied was developed by UC Berkeley to reduce costs by identifying the largest responders and sending them lower incentives, but does not reflect the optimal cost-reducing dispatch of individuals due to constraints in the experimental structure. This section describes the research design in further detail and reports the empirical results. It also explores how these effects vary by month of year.

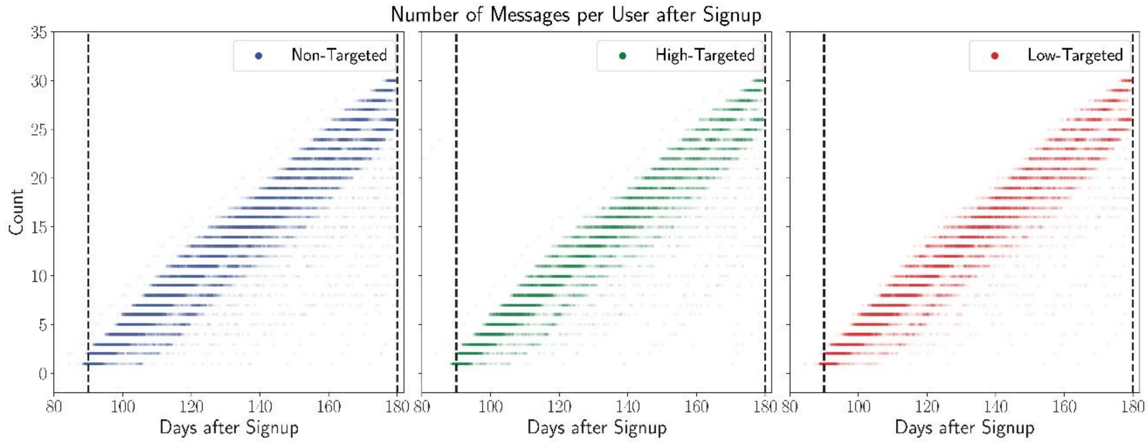
4.1 Research Design and User Experience

Phase 2 explored whether the incentives could be targeted based on Phase 1 responses in order to reduce the cost to the DR provider of calling events. We achieved this by estimating household-level responses using a ML model to predict each individual's counterfactual consumption during an event and then averaged the difference between these counterfactuals and observed consumption to come up with what we termed the Individual Treatment Effect (ITE). Each week as users reached 90 days we would estimate each eligible household's ITE and then rank them within the cohort to be transitioned. Phase 2 lasted for a duration of 90 days for households assigned to the Standard or Encouraged groups initially. On day 181 households were transitioned out of Phase 2.

To causally estimate the effect of the targeting strategy we employed, we randomly assigned individuals to a control group that continued to receive event messaging in the same fashion as Phase 1: receiving 5, 25, 50, 100, and 300 point per kWh incentive levels with equal probability. The targeted group received a limited set of incentive levels based on their ITE ranking. Given the insensitivity documented in the previous section, we pursued a strategy that sent the most responsive users lower incentive levels (5, 25, and 50 points per kWh) and the least responsive users the highest incentive levels (100 and 300 points per kWh). Comparing the entire targeted group to the control provides estimates of how effective this shuffling of incentives was at reducing payouts without reducing the kWh response.

Figure E-40 shows the number of messages for each group by days since sign-up. This shows the messages continued in a pattern similar for all three groups. The only differences were the incentive levels received.

Figure E-40: Number of Phase 2 DR Events per Individual Over Time



4.2 Balance Check

As in Phase 1 of the experiment, we check that the experimental groups are balanced, such that any observational differences in the groups – in expectation - can be attributed to the intervention of interest.

In a similar fashion to the balance checks from Chapter 3, we compare users in the non-targeted and targeted groups by their mean consumption in pre-treatment periods (i.e. before enrollment in Phase 1 of the experiment) as well as ambient air temperature. Figure E-41 and Figure E-42 show these metrics for both experimental groups broken out by the hour of the day. Both figures confirm that targeted and non-targeted users are indeed balanced in terms of electricity consumption and ambient air temperature.

Figure E-41: Consumption for Targeted vs. Non-Targeted Users by Hour of the Day

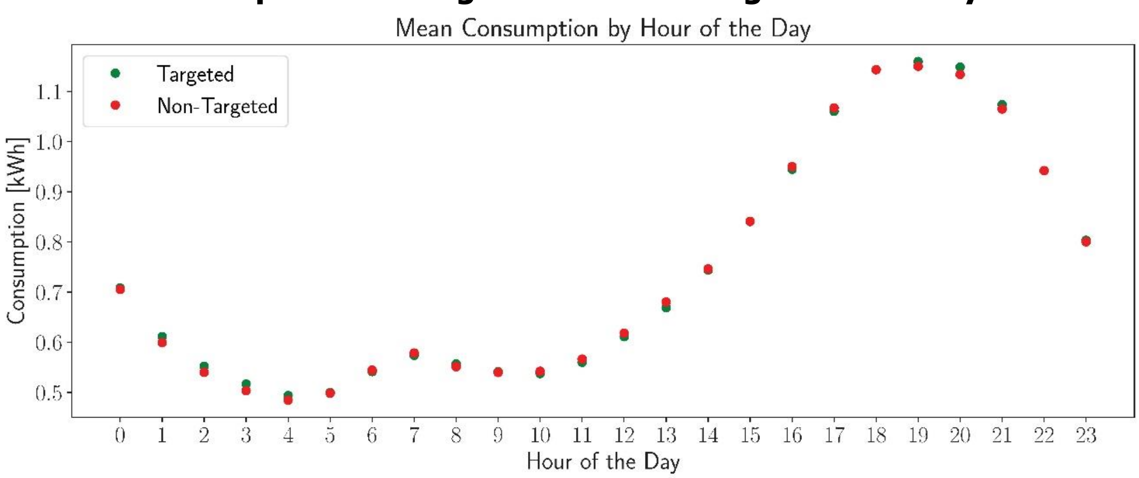
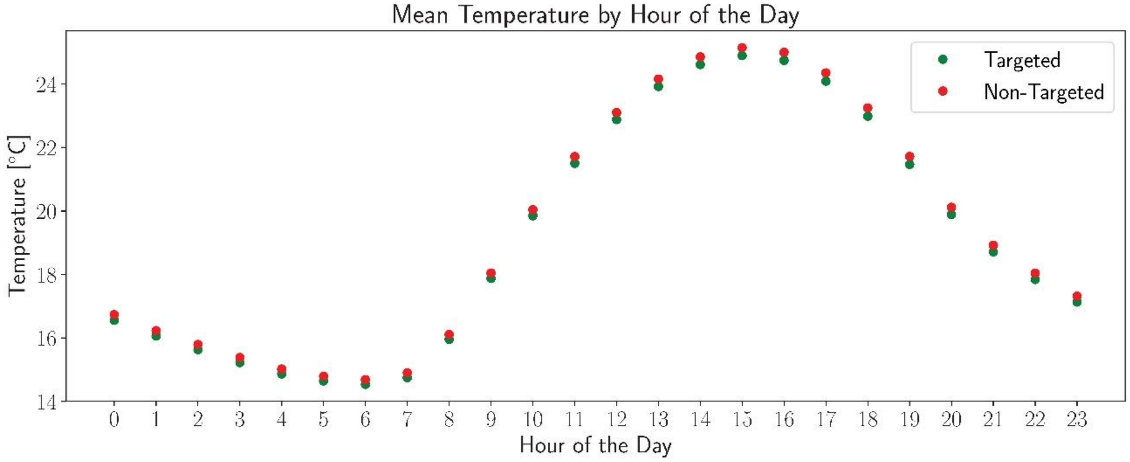


Figure E-42: Temperature for Targeted vs. Non-Targeted users by Hour of the Day



We also test for observable differences for the Phase 1 period. We use a sample of 2,725 households who experienced any Phase 2 before the completion of the experimental period.

4.3 Simulation to Determine Targeting Metric

At the outset of the experiment, we intended to estimate each individual's elasticity since we had within-household incentive variation. However, after the results began to come in that individuals were not very responsive to the incentive level, our ability to measure (and target) based on these slopes was drawn into question. This is due to the fact that differentiating users on slopes that were very close to zero would result in targeting based on a metric that was a poor differentiator. As such, we took on an intermediate step to determine if elasticity estimation was feasible and if not, if there were other strategies to target incentives.

We turned to simulation to tackle this problem and updated our targeting strategy on 6/27/2017. That is, we utilized actual observations from Phase 1 and Phase 2 for a small subset of users who had already experienced both phases and simulated a continuation of the experiment with hypothetical targeting strategies. We then chose the one that had the largest kWh/point reduction in aggregate since it proxies well for the cost of the response to the DR provider. We selected from the following targeting strategy candidates.

1. Individual Treatment Effect (ITE) – the pooled response per individual for any incentive level
2. ITE normalized by average reward level received – the response normalized by the incentive level offered
3. Intercept of estimated individual demand curve – the intercept of the individual demand estimation
4. Slope of estimated individual demand curve – the slope of the individual demand estimation (the original targeting criteria)
5. Random assignment – this provided a benchmark to validate the simulation

Each of the candidates was computed in kWh and %-values as the outcome variable, giving a total of 10 candidates.

The process of the simulation was as follows:

1. Use the sample of non-targeted users prior to 6/27/17.

2. Randomize 50% of the sample to simulated targeted and non-targeted groups
3. Sort targeted users on a given targeting criteria and assign households with the most negative estimates to the low incentive group and those with the least negative estimates to the high incentive group.
4. We then simulated Phase 2, drawing incentive levels with equal probability for each group. Simulated non-targeted were given all 5 incentive levels with 50% probability. Simulated low-targeted were given 5, 25, and 50 point per kWh incentive levels with approximately 33% probability. Simulated high-targeted were given incentive levels 100 and 300 points per kWh with 50% probability.
5. Using actual #OhmHours and variation from Phase we calculated the following:
    - The difference in average consumption in targeted vs. non-targeted groups
    - The difference in measured reductions (forecast less actual consumption)
    - The difference in payouts
6. Repeat this process 1000 times and take the mean across simulations

Figure E-43 plots the distribution of payouts from the Demand Response Provider (DRP) to targeted and non-targeted simulated users. The top panel verifies that random assignment of targeted users to simulated-high-targeted and simulated-low-targeted groups achieves an average payout that matches the payout of the non-targeted group, thereby verifying the validity of random assignment. Panels 2-5 include this random benchmark, which serves as a comparison for the payout to the targeted users under varying targeting criteria. We observe that the criteria ITE, ITE normalized, as well as intercept all reduce the payout significantly, however, ITE performs best. According to the bottom panel, the slope criterion underperforms.

Figure E-43: Differences in Payouts from Simulated Experiment

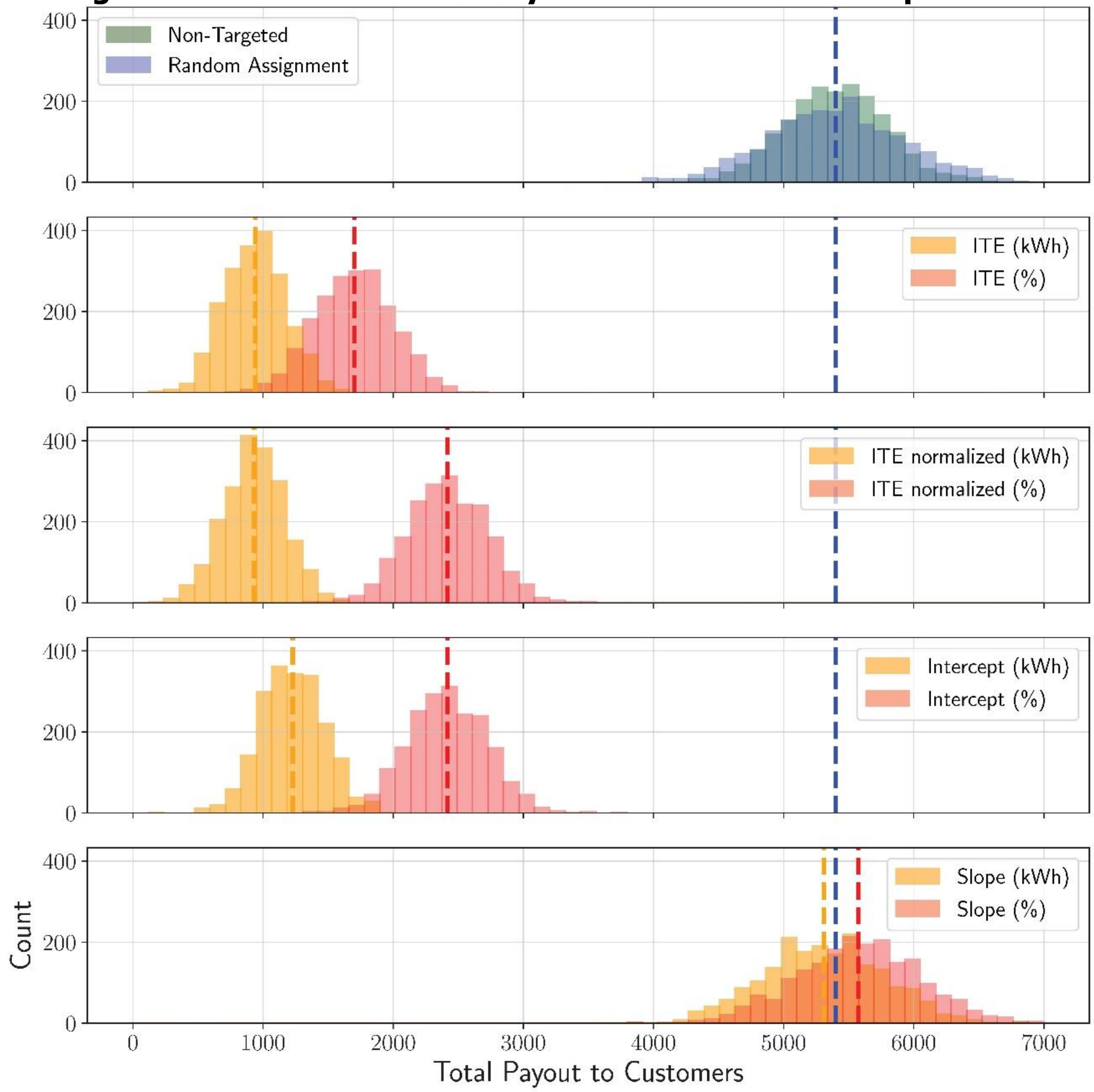
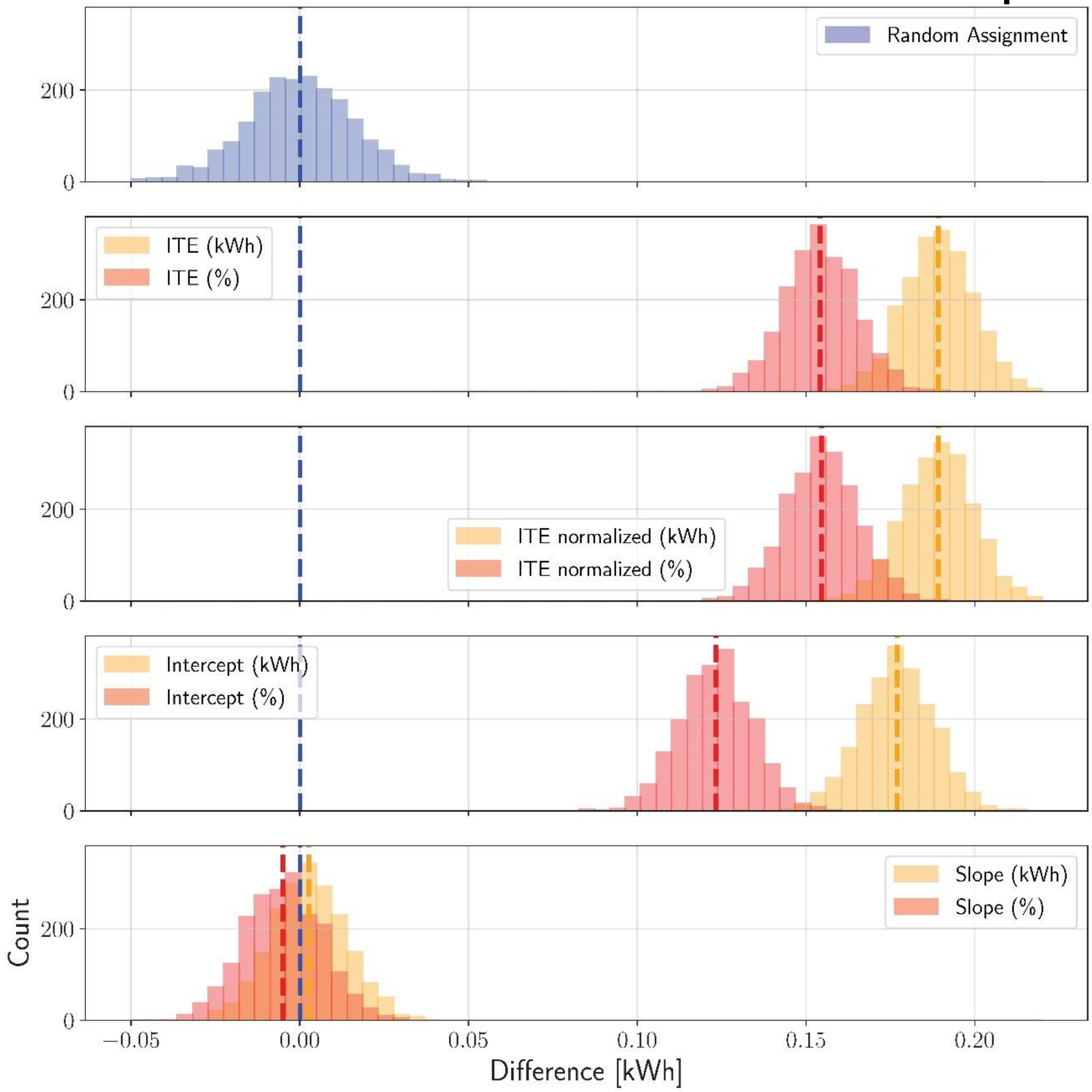


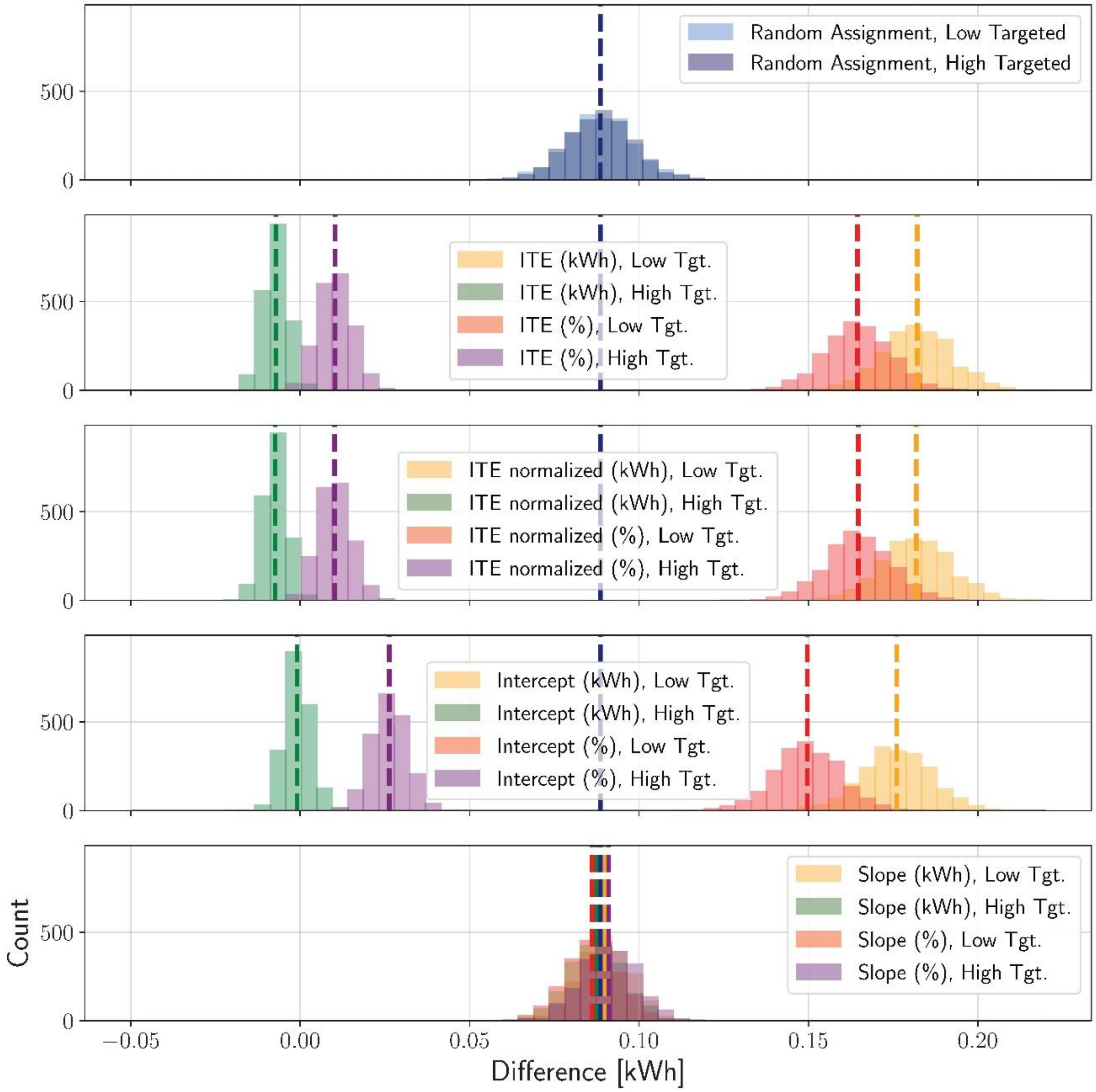
Figure E-44 plots the distribution of differences in means between estimated Phase 2 reductions of simulated low and high targeted users. Random assignment (top panel) shows a distribution centered around zero. Panels 2-4 graph distributions centered around a positive value, which is consistent with the objective of assigning high (low) rewards to low (large) Phase 1 reducers. ITE (kWh) achieves the largest mean difference whereas the slope criterion appears to be ineffective.

Figure E-44: Differences in Measured Reductions from Simulated Experiment



Lastly, Figure E-45 supports the results in Figure E-44, as it breaks down the reductions of simulated low and high targeted users. The differences in means between the reductions estimated for simulated-low and simulated-high-targeted users (for a particular targeting criterion) is exactly the mean given in Figure E-44.

Figure E-45: Differences in Consumption from Simulated Experiment



The simulations showed how poor the slope of each individual demand performed in differentiating users. Instead, the ITE in terms of kWh candidate provided the greatest reduction in payouts according to the simulation. Thus, after 7/1/17 we continued the experiment with this targeting strategy.

4.4 ITE Estimator and Targeting Methodology

The targeting methodology used required an estimate of each household's event responsiveness. The estimate we use is our ITE estimate. In its simplest form, each household's ITE estimate was given by the following formula:

where *i* = household and *t* = the hour-of-sample. The formula basically averages over the difference between the observed consumption in *kWh* and the *Counterfactual*. We produce the counterfactual using a random forest estimator with 5 lagged hours of consumption

and outdoor temperature along with a host of fixed effects. Chapter 6 goes into more detail on the machine learning algorithm. We then rank individuals based on the $I\Upsilon E_i$ which is what we estimate the response per event is for household *i*.

4.5 Phase 2 Experimental Results

We estimate the causal effect of implementing our targeting strategy using the experimental assignment to targeted versus non-targeted groups for households that experienced Phase 2 after 6/27/2017 when we implemented the preferred targeting strategy to 3/1/2018 – a sample of 2,725 households. We only use data from event hours during Phase 2, giving a total sample of 73,165.

Table E-12 shows the results of the targeting by comparing the averages between the two groups. We examine the reductions from baseline and the number of points paid out per event as outcomes. Each column represents a separate regression. Rows 1-3 report the difference in the groups which is the effect of targeting and rows 4-6 report the means in the non-targeted control group to facilitate interpretation. Rows 1 and 4 report the effect for both types of households and Rows 2-3 and 5-6 break out the effect by the Most and Least Responsive types.

Columns 1 and 2 of Table E-12 show the effect of targeting on reductions from baseline. The results show reductions were on average 0.070 kWh smaller, but on the order of the effects found in Phase 1 and that the targeting strategy made these reductions smaller by 0.013 kWh. This represents a 19 percent reduction, although the difference is not statistically significant at 1 and 5 percent levels. We also see the most responsive types reduce around 0.102 kWh versus 0.035 kWh which also suggests that the ITE estimation replicates the reductions as measured by the baseline.

Columns 3 and 4 of Table E-12 show the effect of targeting on points paid per event. The results show on average all participants were paid 6.3 points per event in the control group and that the targeting strategy decreased this payout by 3.1 points, a reduction of 49 percent. Households designated as most responsive households were paid 8.7 points per event in the control group and targeting (offering only 5, 25, and 5 point per kWh incentives) reduced this payment by 6.3 points and lowered the payout by 72 percent.

Conversely, the least responsive households are paid 3.8 points per event and targeting (offering only 100 or 300 point per kWh incentives) did not significantly change payouts.

Table E-12: Effect of Targeting on kWh Reduced and Points Paid per Event

	Reduction from BL (kWh)	Reductions from BL (kWh)	Points per Event (1 pt = $0.01)	Points per Event (1 pt = $0.01)
Difference from Targeting Strategy	-0.013* (0.008)		-3.06*** (0.82)	
Difference for Most Responsive		-0.012 (0.013)		-6.30*** (1.07)
Difference for Least Responsive		-0.014*** (0.007)		0.41 (1.23)
Mean for Non-	0.070***		6.31***	

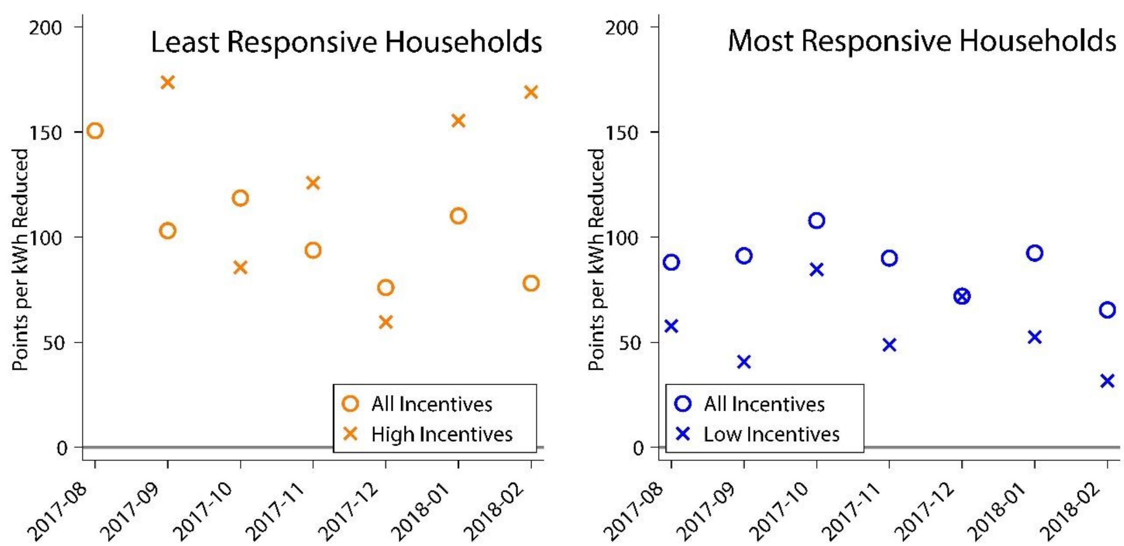
Targeted Group	(0.011)		(0.65)	
Mean for Most Responsive		0.102*** (0.010)		8.67*** (1.05)
Mean for Least Responsive		0.035*** (0.005)		3.76*** (0.72)
Households	2,725	2,725	2,725	2,725
N (observations)	73,165	73,165	73,165	73,165

The results in Table E-12 suggest the cost of payouts can be dramatically reduced by sending lower incentives to those designated as most responsive with little effect on the reduction. Dividing the points paid by the kWh reduced gives a rough metric on the average cost of the events. These suggest the 85 points per kWh cost of the most responsive types can be reduced to 26 points per kWh. Further, the least responsive types have a cost of 107 points per kWh. While we do not verify that these payments are cashed out, they can be roughly translated to dollars as 1 point for $0.01.

Figure E-46 shows the cost in points per kWh reduced from baseline broken out by month-of year. The left panel shows the least responsive households and the right panel the most responsive with All Incentives indicating the Control Group and High/Low Incentives indicating the Targeted Group. We omit one data point from the plot for scale: the Least Responsive targeted (high incentive) group estimate for August 2017 because it indicated a negative cost of 148.7 points per kWh.

The costs for both groups are comparable for the non-targeted group and generally around the 96 points per kWh average incentive level. The figure shows consistent reductions in the cost for summer months (August and September) and the colder winter months (October-February). The costs decrease for sending most responsive households lower incentives is fairly consistent around 50 points per kWh. The costs do not systematically rise if least responsive households are sent higher incentives, but do on average.

Figure E-46: Effect of Targeting on Cost in Points per kWh by Month-of-Year



CHAPTER 5: Phase 3 – Incentives vs. Moral Suasion

Phase 3 explored how households respond to different event messaging. Specifically, we address the question: are responses to financial incentives different from responses to messages with moral suasion in the form of green/environmental messaging? This section describes the research design in further detail and reports the empirical results. It also explores how these effects vary by automation status.

5.1 Research Design and User Experience

Phase 3 occurred 180 days after enrollment for the Standard and Encouraged users and lasted for 90 days, after which the household was transitioned out of the experiment. The randomization for Phase 3 was by event so that each time a DR event occurred, households had a 25% probability of being in a control group that did not receive an event that hour or equal 25% probability of receiving one of three messages. One of the messages was a 100 point per kWh message identical to the Phase 1 messaging – "Points Only". The other two included messaging that suggested there were potential environmental benefits to reducing electricity consumption. Specifically, there was a "Suasion Only" message that included the environmental messaging and no potential for a monetary reward in points and a "Price and Suasion" that included the messaging and a 100 point per kWh incentive level.

5.2 Empirical Strategy

In order to estimate the effect of each message we use a similar difference-in-differences strategy to Phase 1. We use pre-enrollment data to control for between household time-invariant differences and hour-of-sample fixed effects to control for differences in aggregate consumption. We also include the same parametric temperature controls as Phase 1. The estimating equation is given by:

$$
\begin{aligned}
Y_{ith} = \delta_{moral} MoralOnly_{it} + \delta_{points} PointsOnly_{it} + \delta_{both} Both_{it} \\
+\alpha_{ih} + \gamma_t + \beta_C CDH_{it} + \beta_H HDH_{it} + \eta_{ith}
\end{aligned}
$$

where *MoralOnly*, *PointsOnly*, and *Both* are indicator variables that the household received that event during that period. We do not include non-event hours since there is no obvious control group for comparison.

5.3 Phase 3 Experimental Results

This section reports estimates of the causal average effect for each message. Figure E-47 and Figure E-48 summarize the results visually in kWh consumed and log(kWh), respectively. The leftmost estimates show the effect of suasion and no monetary incentives, the middle estimate is the effect of points only consistent with Phase 1 and the rightmost estimates are for the combination of points plus suasion. The black circles are estimated from the equation described in 5.2 and the blue X's and orange triangles break that effect out by automated and non-automated households. All estimates can be interpreted as causal for the sample cut described, but comparisons between automated and non-automated cannot be causally interpreted.

Figure E-47: Effect of Moral Suasion Versus Points in kWh

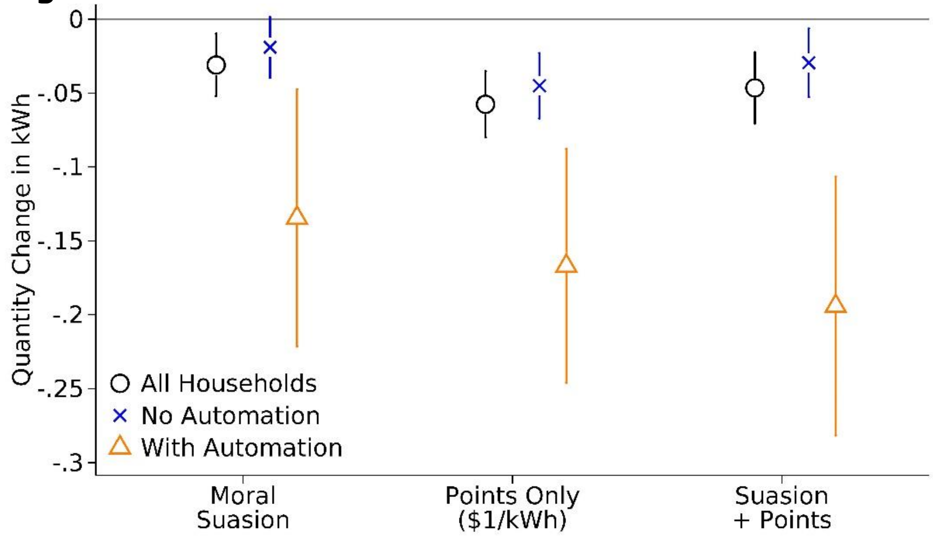
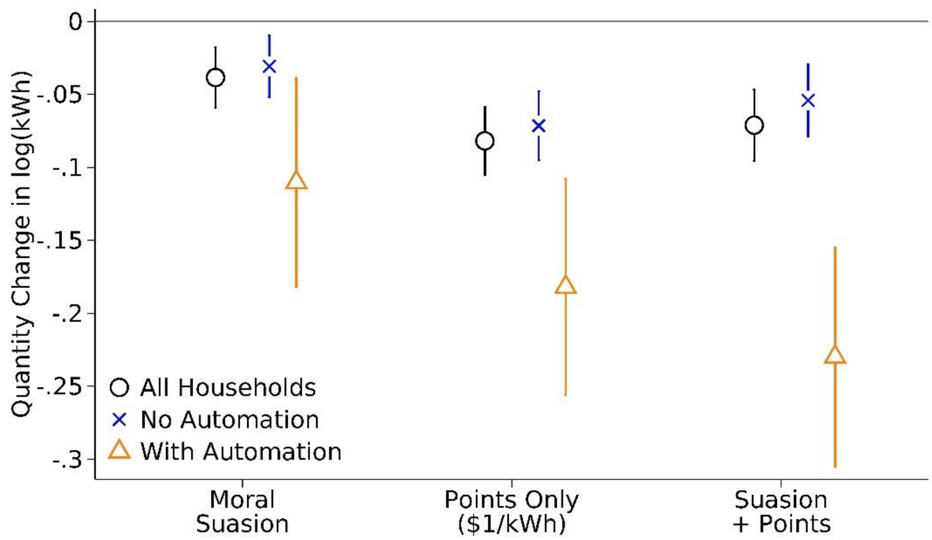


Figure E-48: Effect of Moral Suasion Versus Points in log(kWh) (approximately percent)



The results show that moral suasion has a smaller impact on consumption than the messages with the monetary incentives and this is driven in large part by the non- automated households that are making active decisions. Table E-13 reports the estimates for Figure E-47 and Figure E-48 along with the p-values on statistical tests that the coefficients are the same.

Table E-13: Effect of Moral Suasion vs. Monetary Incentives

<table>
<tr><th></th><th>kWh</th><th colspan="2">kWh</th><th>log(kWh)</th><th colspan="2">log(kWh)</th></tr>
<tr><td>Moral Suasion Only</td><td>-0.031***
(0.011)</td><td colspan="2"></td><td>-0.038***
(0.011)</td><td colspan="2"></td></tr>
<tr><td>Moral Suasion Only x No Automation</td><td></td><td colspan="2">-0.132***
(0.011)</td><td></td><td colspan="2">-0.030***
(0.037)</td></tr>
<tr><td>Moral Suasion Only x Automation</td><td></td><td colspan="2">-0.117***
(0.011)</td><td></td><td colspan="2">-0.110***
(0.011)</td></tr>
<tr><td>Points Only ($1/kWh)</td><td>-0.058***
(0.011)</td><td colspan="2"></td><td>-0.082***
(0.012)</td><td colspan="2"></td></tr>
<tr><td>Points Only ($1/kWh) x No Automation</td><td></td><td colspan="2">-0.147***
(0.012)</td><td></td><td colspan="2">-0.071***
(0.012)</td></tr>
<tr><td>Points Only ($1/kWh) x Automation</td><td></td><td colspan="2">-0.124***
(0.009)</td><td></td><td colspan="2">-0.182***
(0.038)</td></tr>
<tr><td>Suasion & Points</td><td>-0.046***
(0.012)</td><td colspan="2"></td><td>-0.071***
(0.012)</td><td colspan="2"></td></tr>
<tr><td>Suasion & Points x No Automation</td><td></td><td colspan="2">-0.147***
(0.012)</td><td></td><td colspan="2">-0.063***
(0.013)</td></tr>
<tr><td>Both Suasion & Points x Automation</td><td></td><td colspan="2">-0.124***
(0.009)</td><td></td><td colspan="2">-0.230***
(0.038)</td></tr>
<tr><td>p-values on null hypothesis:</td><td>Pooled</td><td>Auto-mated</td><td>No Auto</td><td>Pooled</td><td>Auto-mated</td><td>No Auto</td></tr>
<tr><td>Moral = Points</td><td>0.000</td><td>0.302</td><td>0.001</td><td>0.000</td><td>0.011</td><td>0.000</td></tr>
<tr><td>Moral = Both</td><td>0.043</td><td>0.054</td><td>0.201</td><td>0.000</td><td>0.000</td><td>0.010</td></tr>
<tr><td>Points = Both</td><td>0.104</td><td>0.343</td><td>0.055</td><td>0.158</td><td>0.073</td><td>0.038</td></tr>
<tr><td>Moral + Points = Both</td><td>0.001</td><td>0.029</td><td>0.008</td><td>0.000</td><td>0.152</td><td>0.000</td></tr>
<tr><td>Households</td><td>3,391</td><td colspan="2">3,391</td><td>3,391</td><td colspan="2">3,391</td></tr>
<tr><td>N (observations)</td><td>6,836,711</td><td colspan="2">6,836,711</td><td>6,836,711</td><td colspan="2">6,836,711</td></tr>
</table>

The first row Moral = Points tests that the MoralOnly coefficient is equal to the PointsOnly. In all models for the pooled estimates or the non-automated households, the Moral Suasion estimates are statistically significantly different at the 99 percent level. The results are generally consistent with the Moral and Combined as well. The test that Points is the same as Combined on the other hand is insignificant for the pooled estimates and less significant for the non-automated, although still borderline significant at the 90 and 95 percent significance levels.

Taken together, we suggest these estimates imply moral suasion yields smaller reductions and that there is something unique to the monetary incentives. Further, the fact that automated users respond to all messages could be evidence of a default effect in the automation technology.

CHAPTER 6:
Machine Learning Models

6.1 Comparison with Non-Experimental, Individual Treatment Effect Estimator

The methodology presented in the previous section(s) estimates the average treatment effect and follows the experimental gold standard to evaluate the causal effect of the Demand Response intervention on the temporary reduction in electricity consumption of residential households. We now re-estimate this average treatment effect using a non- experimental estimator, which does not require a control group. Instead, this estimator is capable of estimating *individual treatment effects*, which enable the design of an adaptive targeting scheme (see Chapter 4) to increase the per-dollar-reductions of users. Since this non-experimental estimator is not intrinsically causal, we benchmark the estimates to the experimental estimator to find that the estimates are close to each other. Importantly, this finding suggests that we are capable of estimating unbiased treatment effects on a more granular level compared to the standard, average treatment effects estimator.

6.2 Non-Experimental Treatment Effect Estimation

To estimate the effect of the DR intervention program, we adopt the potential outcomes framework introduced by Rubin (1974). Let $I = \{1, \dots, n\}$ denote the set of users. The indicator $D_{@A} \in \{0, 1\}$ encodes the fact whether or not user i received DR treatment at time t. Each user is equipped with a consumption time series $y_{@} = \{y_{@t}, \dots, y_{@u}\}$ and associated covariates $X_{@} = \{x_{@t}, \dots, x_{@u}\}$, where time is indexed by $t \in T = \{1, \dots, \tau\}$. Let y^{x} and y^{t} denote @A @A user i's electricity consumption at time t for $D_{@A} = 0$ and $D_{@A} = 1$, respectively. Let $C_{@}$ and $T_{@}$ denote the set of control and treatment times for user i.

The number of treatment hours is much smaller than the number of non-treatment hours.

Thus $0 < \frac{|T_i|}{|C_i|} \ll 1$. Further, let $D_{i,t}$ and $D_{i,c}$ denote user i's covariate-outcome pairs of treatment and control times, respectively. The one-sample estimate of the treatment effect on user i at time t, given the covariates $x_{it} \in \mathbb{R}^{n_x}$, is

$$\beta_{it}(x_{it}) := y_{it}^1(x_{it}) - y_{it}^0(x_{it}),$$

which varies across time, the covariate space, and the user population. Marginalizing this one-sample estimate over the set of treatment times T_i and the covariate space X_i yields the user-specific Individual Treatment Effect (ITE) β_i:

$$\beta_i := \mathbb{E}_{X_i}\mathbb{E}_{t\in T_i}[y_{it}^1 - y_{it}^0 \mid x_{it}] = \frac{1}{|T_i|}\sum_{t\in T_i}(y_{it}^1 - y_{it}^0).$$

The average treatment effect on the treated (ATT) follows from the definition of the ITE:

$$ATT = \mathbb{E}_{i\in I}[\beta_i] = \frac{1}{|I|}\sum_{i\in I}\frac{1}{|T_i|}\sum_{t\in T_i}(y_{it}^1 - y_{it}^0).$$

Since users were put into different experimental groups in a randomized fashion (see Chapter 3.2), the ATT and the average treatment effect (ATE) are identical.

The *fundamental problem of causal inference* refers to the fact that either the treatment or the control outcome can be observed, but never both (provided there are no missing observations). That is,

$$y_{it} = y^0_{it} + D_{it} \cdot (y^1_{it} - y^0_{it}).$$

Thus, the ITE is not identified, because one and only one of both potential outcomes is observed, namely $\{y^1_{it} \mid t \in T_i\}$ for the treatment times and $\{y^0_{it} \mid t \in C_i\}$ for the control times.

It therefore becomes necessary to estimate counterfactuals.

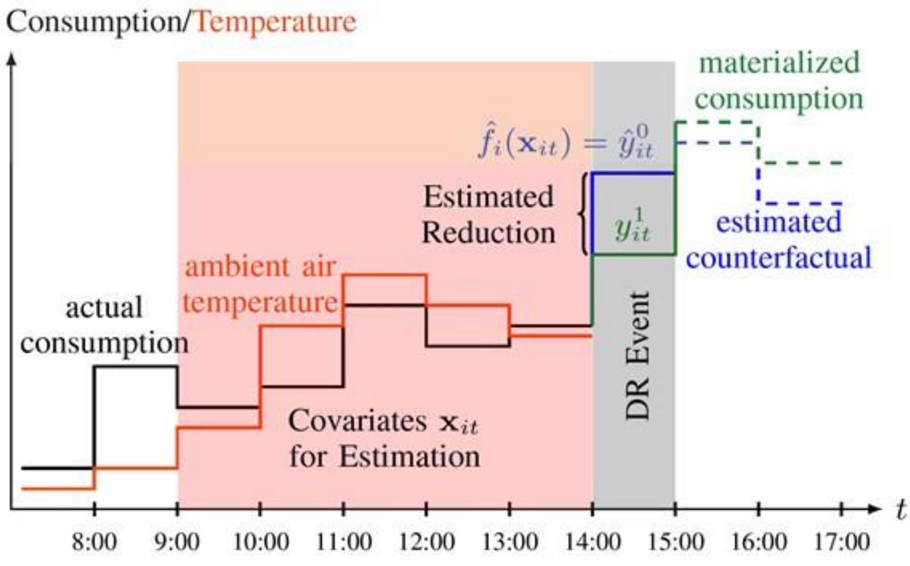
Consider the following model for the estimation of such counterfactuals:

$$y_{it} = f_i(x_{it}) + D_{it} \cdot \beta_{it}(x_{it}) + \varepsilon_{it},$$

where ε denotes noise uncorrelated with covariates and treatment assignment. $f_i(\cdot) : \mathbb{R}^{n_f} \mapsto \mathbb{R}$ is the conditional mean function and pertains to $D_{it} = 0$. To obtain an estimate for $f_i(\cdot)$, denoted with $\hat{f}_i(\cdot)$, control outcomes $\{y^0_{it} \mid t \in C_i\}$ are first regressed on $\{x_{it} \mid t \in C_i\}$, namely their observable covariates. In a second step, the counterfactual y^0_{it} for any $t \in T_i$ can be estimated by evaluating $\hat{f}_i(\cdot)$ on its associated covariate vector x_{it}. Finally, subtracting y^0_{it} from y^1_{it} isolates the one-sample estimate $\beta_{it}(x_{it})$, from which the user-specific ITE can be estimated.

Figure E-49 illustrates this process of estimating the reduction during a DR event by subtracting the actual consumption y^1_{it} from the predicted counterfactual $\hat{y}^0_{it} = \hat{f}_i(x_{it})$. Despite the fact that consumption can be predicted for horizons longer than a single hour, we restrict our estimators $f_i(\cdot)$ to a single hour prediction horizon as DR events are at most one hour long.

Figure E-49: Estimation of the Counterfactual y0/it using Treatment Covariates x_{it}.



To estimate the conditional mean function $f_i(\cdot)$, we use the following classical regression methods [Hastie et al. 2009], referred to as estimators:

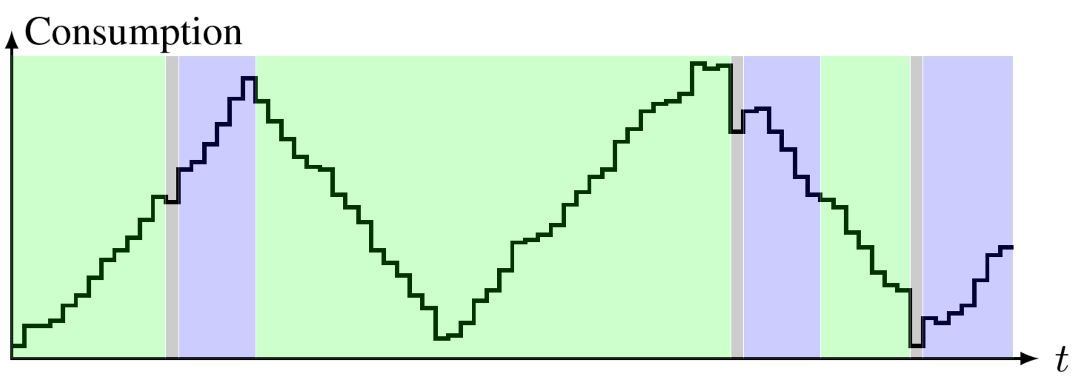
- (E1): Ordinary Least Squares Regression (OLS)

- (E2): L1 Regularized (LASSO) Linear Regression (L1)
- (E3): L2 Regularized (Ridge) Linear Regression (L2)
- (E4): k-Nearest Neighbors Regression (KNN)
- (E5): Decision Tree Regression (DT)
- (E6): Random Forest Regression (RF)

DT (E5) and RF (E6) follow the procedure of Classification and Regression Trees [Breiman et al. 1984]. We compare estimators (E1)-(E6) to the CAISO 10-in-10 Baseline (BL) [CAISO 2014], which, for any given hour on a weekday, is calculated as the mean of the hourly consumptions on the 10 most recent business days during the selected hour. For weekend days and holidays, the mean of the 4 most recent observations is calculated. This BL is further adjusted with a Load Point Adjustment, which corrects the BL by a factor proportional to the consumption three hours prior to a DR event [CAISO 2014].

Since users tend to exhibit a temporary increase in consumption in the hours following the DR intervention [Palensky and Dietrich 2011], we remove 8 hourly observations following each DR event in order to prevent estimators (E1)-(E6) from learning from such spillover effects. This process is illustrated in Figure E-50.

Figure E-50: Separation of Consumption Time Series into training set (green), DR Events (grey), and Spillover Periods (blue)



Hence the training data used to estimate the conditional mean function $f_{@}(\cdot)$ consists of all observations leading up to a DR event, excluding those that are within 8 hours of any DR event. To estimate user i's counterfactual outcome $\tilde{y}_{@}^{x}$ during a DR event, we use the following covariates:

- 5 hourly consumption values preceding time t
- Air temperature at time t and 4 preceding measurements
- Hour of the day, an indicator variable for business days, and month of the year as categorical variables

Thus, the covariate vector writes

$$x_{@A} = [y^{x}_{@A\,\hat{\imath}t} \cdots y^{x}_{@A\,\hat{\imath}\hat{\imath}}\ T_{@A} \cdots T_{@A\hat{\imath}\hat{\imath}}\ C(HoD_{@A}) : C(isBday_{@A})\ C(MoY)],$$

where $T_{@A}$ denotes temperature, $HoD_{@A}$ hour of day, $isBday_{@A}$ an indicator for business days, and $MoY_{@A}$ the month of year for user i at time t. "C" denotes categorical variables and ":" their interaction.

6.3 Simulation of Non-Experimental Estimators on Semi-Synthetic Data

To reduce the propagation of model bias into the estimation of treatment effects, we empirically de-bias estimators by subtracting the empirical bias, which is the difference in means between the observed control outcomes and their predictions, from all estimated counterfactuals:

$$\hat{y}^0_{it} \leftarrow \hat{y}^0_{it} - \frac{1}{|C_i|}\sum_{k\in C_i}(\hat{y}^0_{ik} - y^0_{ik}) \ \forall t \in T_i.$$

Although this operation leads to an increase of variance of counterfactual estimates, the reasoning behind this operation is that an unbiased estimator provides a fair economic settlement for DR reductions. If the estimator were biased in favor of the consumer, then the user, in expectation, would receive an additional payment proportional to the bias each time a DR event is called despite not having actually reduced his consumption by the amount of bias. Likewise, an estimator biased in favor of the utility results in the opposite effect.

Figure E-51 shows the distribution of one-sample prediction errors on the placebo treatment set for a selected subset of estimators. RF outperforms all other estimators as it is centered most sharply around zero, thus having the smallest sample standard deviation. The performance of L1, L2, and OLS is similar, indicating that the training data is of suffcient length such that overfitting is not a concern. The performance of KNN lies between L1/L2/OLS and BL. The CAISO BL performs worst. The sample bias and standard deviation of estimated residuals are provided in Table E-14. As the estimators have been calibrated with the de-biasing operation, the one-sample estimation errors (whose mean is the bias) for all estimators varies insignificantly around zero.

Figure E-51: Distribution of One-Sample Absolute Prediction Errors on Placebo Events

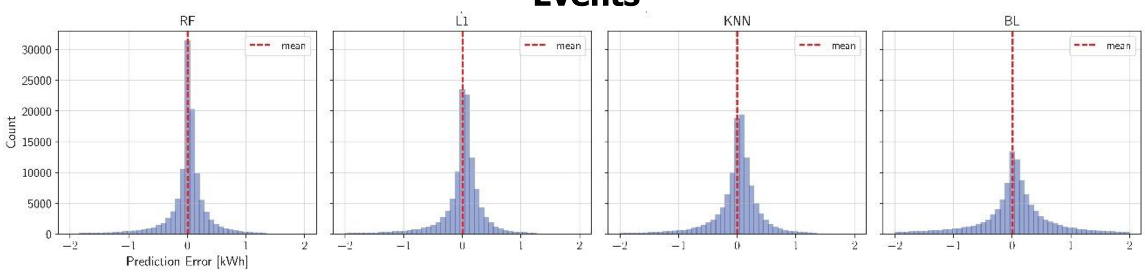


Table E-14 also provides the median of the set of Mean Absolute Percentage Errors (MAPE) across all users, for all estimators. The MAPE for a given user i is defined as follows:

$$MAPE = \frac{1}{|V_i|}\sum_{t\in V_i}\frac{|\hat{f}_i(x_{it}) - y^0_{it}|}{|y^0_{it}|}\cdot 100\%,$$

Table E-14: Sample Standard Deviations of ATT Estimates

Estimation Method	Bias	Standard Deviation	Median MAPE [%]
RF	0.00280	**0.34460**	**30.779**
OLS	0.00157	0.35981	35.088

L1	0.00184	0.35969	34.945
L2	0.00153	0.35977	35.079
DT	-8.26e-05	0.40386	35.461
KNN	-0.00129	0.41011	41.341
BL	0.00684	0.49550	50.496

where $V_{@} \subset C_{@}$ is a subset of the set of training times used for validation of the estimators during the training step. Using standard k-fold cross validation on the training data set $D_{@,\mathrm{Af}}$ (i.e. we chose k = 10), $V_{@}$ can be interpreted as the set of time indices in the holdout set of any given fold. Figure E-52 compares the MAPEs weighted across users, where users with more observations are given a larger weight.

Figure E-52: Distribution of weighted MAPEs across Training Data for Different Estimators

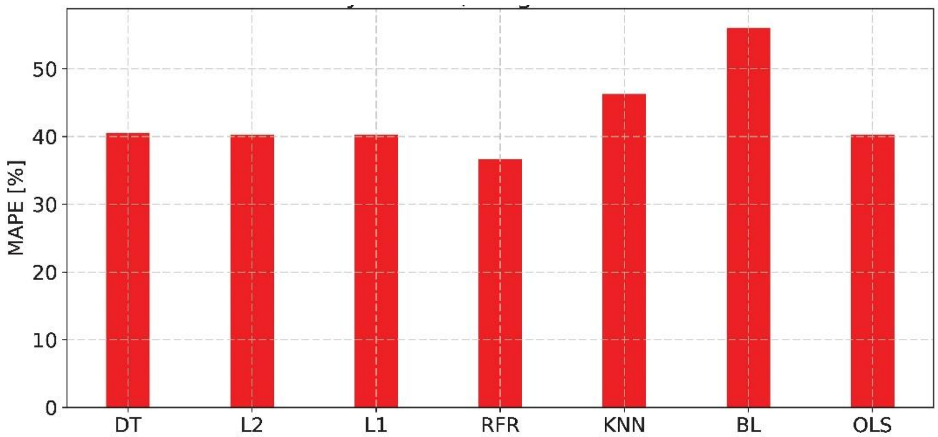
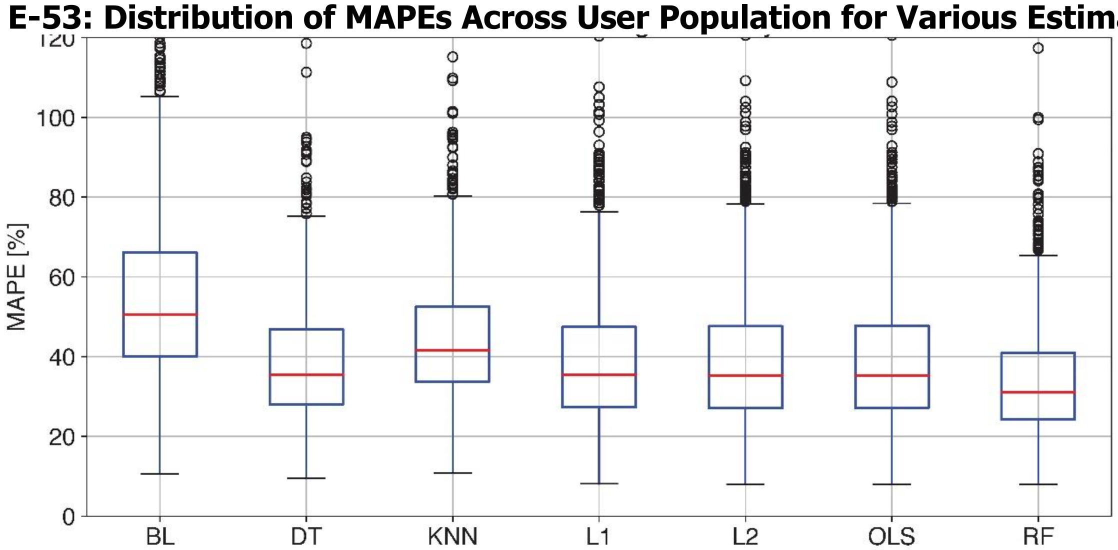


Figure E-53 shows the distribution of MAPEs across the user population for each of the estimators, which again illustrates the inferiority of the CAISO BL compared to RF, which is the estimator with the lowest MAPE.

Figure E-53: Distribution of MAPEs Across User Population for Various Estimators



We now generate synthetic treatments to evaluate how closely various estimators can replicate a known, synthetic treatment effect. The synthetic treatment set $D_{i,\text{cûF}}$ is used as a set of ground truth counterfactuals $\{y^{x} \mid t \in D_{i,\text{cûF}}\}$ for which treatment outcomes

$\{y^{\mathbf{t}} \mid t \in D_{i,\text{cûF}}\}$ are synthetically generated. We assume a constant ITE, namely $-1 \leq \beta_{\mathbf{t}} =$

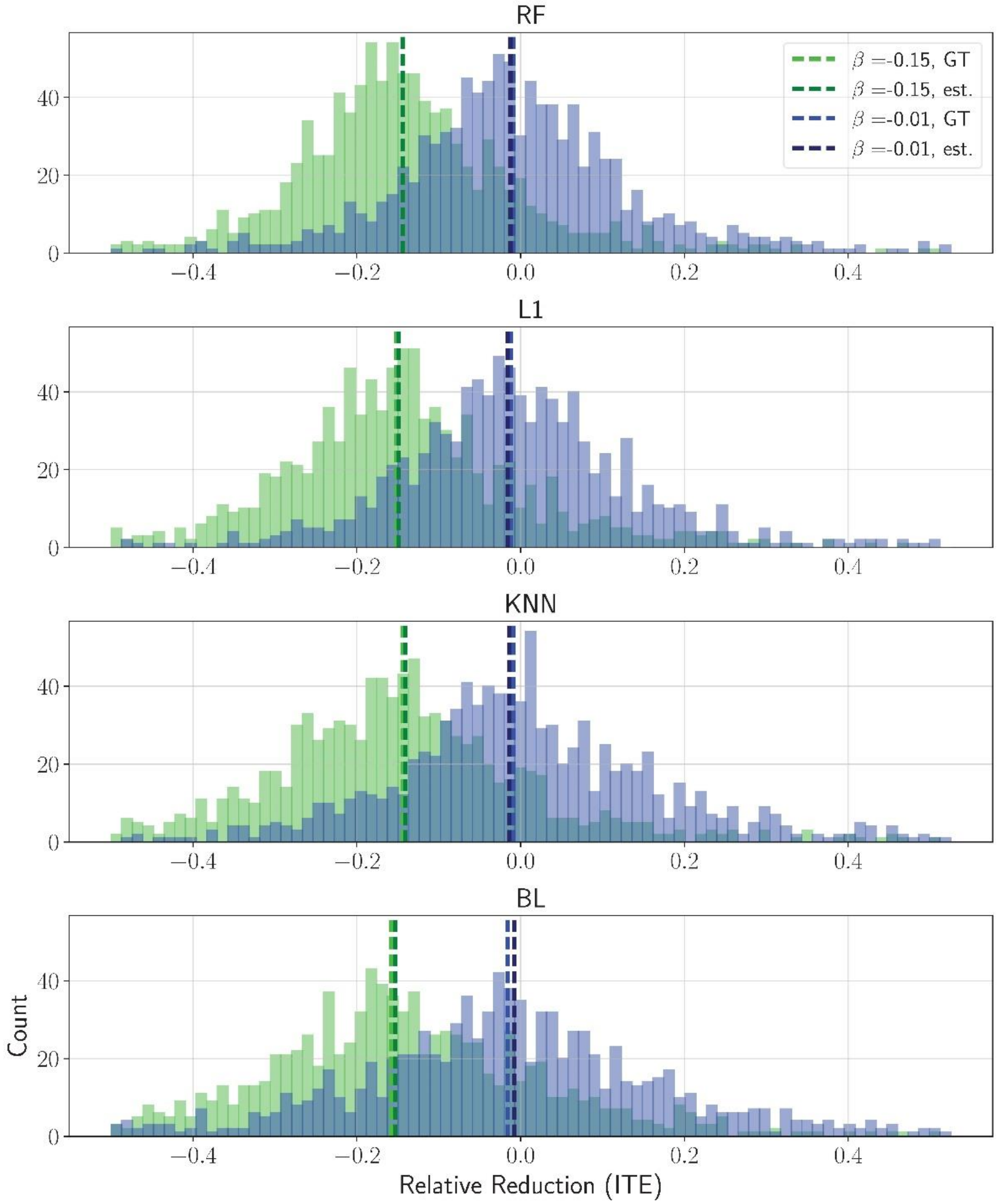
$\ldots = \beta_{\dagger} =: \beta \leq 0$, across all synthetic times and the covariate space for each user i, as a percentage of user i's mean counterfactual consumption. The one-sample reductions are varied around the mean reduction through Gaussian noise with an appropriately chosen standard deviation:

Since β is random in $\sigma^{ß}$, the realized ITE β is distributed according to β

which follows from the above equations and noting that $\{y_{iA}\}, t \in S_i$ are independent random variables. Using this semi-synthetic treatment data, one can evaluate the ability of the estimators to recover the generated ITE $\beta\mu_i$ (non-normalized) and β (normalized). The sample variance of the ITE estimation errors will again serve as a measure for the predictive power of an estimator, similar to the estimation error of placebo treatments.

Figure E-54 shows the distribution of estimated normalized ITEs $\{\ddot{\beta}_i\}, i \in I$ generated in the above equations across all users, for selected estimators, and for two different ground truth ITEs $\beta_i \in \{-0.01, -0.15\}$. Each ITE draw is obtained from a randomly drawn subset $M_i \subset S_i$, where we chose $|M_i|= 25$.

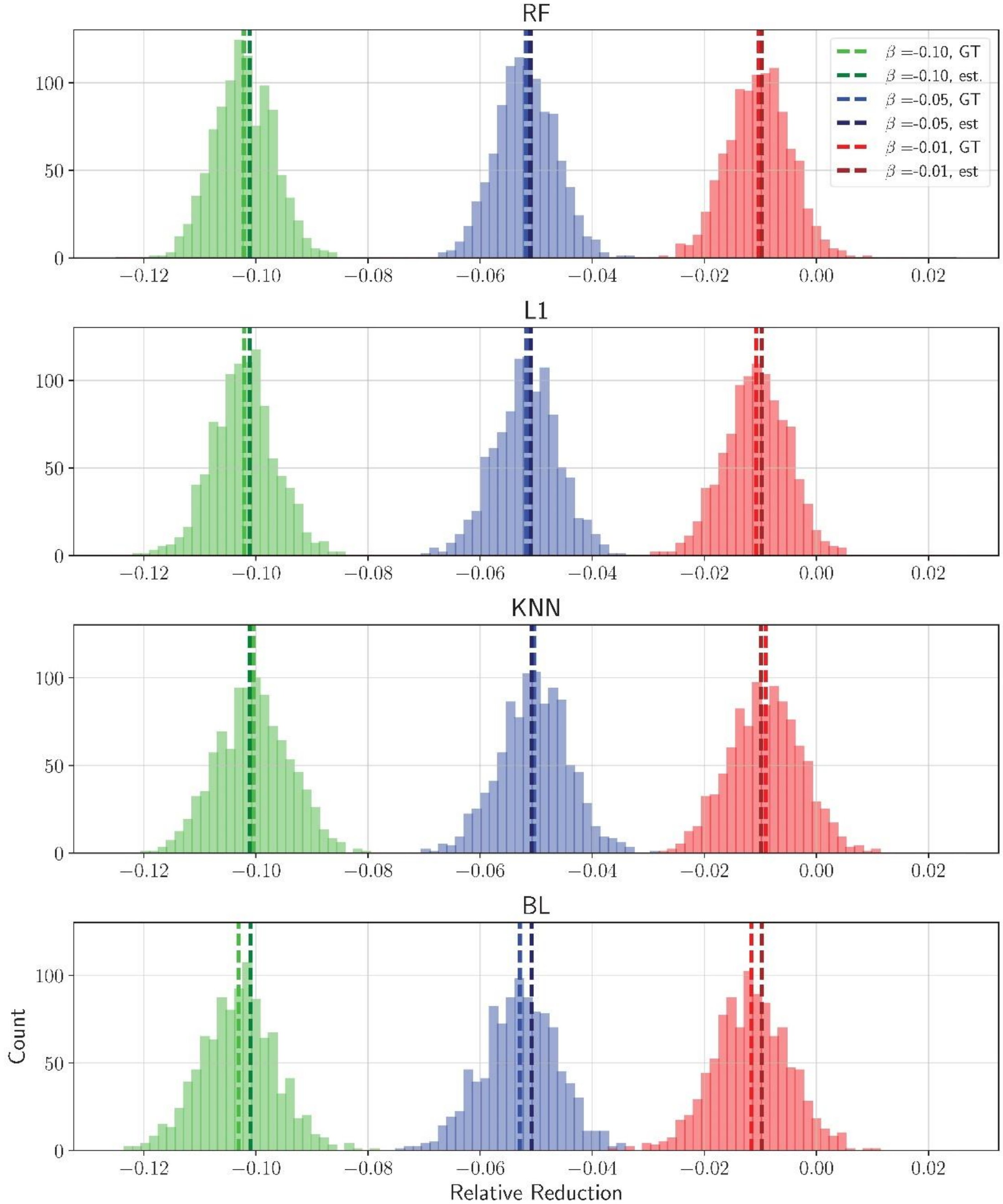
Figure E-54: Comparison of Synthetic ITEs and their Estimations for Various Estimators



Random Forest outperforms estimators (E1) – (E5) and the CAISO BL, as the histograms around the sample mean become wider as we move to the more inaccurate methods towards the bottom of the figure.

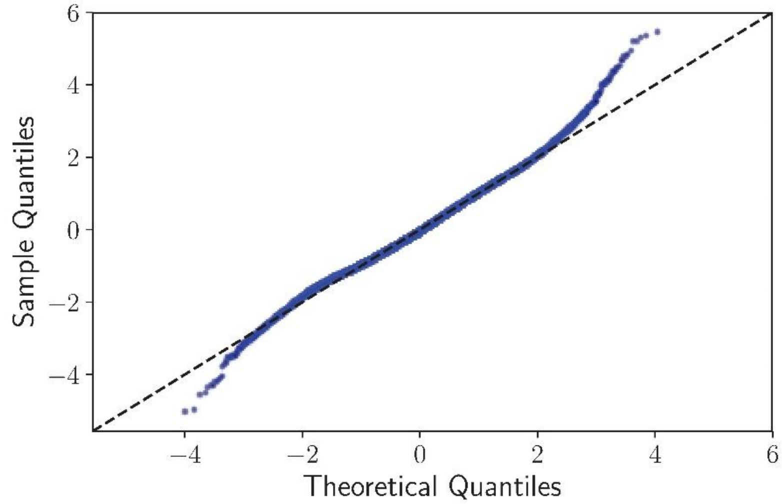
Given a ground truth ATT β, an atomic ATT estimate is obtained by taking the mean of the estimated ITEs (Figure 6.3.4) across all users. Repeating this process M times for each estimator on a newly randomly drawn subset $M_{\emptyset} \subset S_{\emptyset}$ for each iteration yields a distribution of ATT estimates. The ensuing sample standard deviations for $|M_{\emptyset}|$ = 25, M = 1000, and different ground truth ATTs $\beta \in \{-0.01, -0.05, -0.10, -0.15\}$ are shown in Table. For $\beta \in \{-0.05, -0.10\}$, Figure E-55 shows a histogram of the ATTs for each iteration as well as the empirical mean across all iterations. As in Figure 6.234, it can be seen that the sample variance increases as we move towards more inaccurate estimators at the bottom panels.

Figure E-55: Comparison of Synthetic ATTs and their Estimations for Various Estimators



Lastly, Figure E-56 shows a Quantile-Quantile-Plot for the distribution of the residuals estimated with Random Forest. We expect the residuals to be approximately normally distributed, as this is a crucial assumption to justify linear regression using the mean squared error as the loss function. It appears that the residuals are indeed approximately normally distributed around zero, but show a heavy-tailed property for larger deviations. This is sufficient evidence to utilize linear regression for the estimation of counterfactuals.

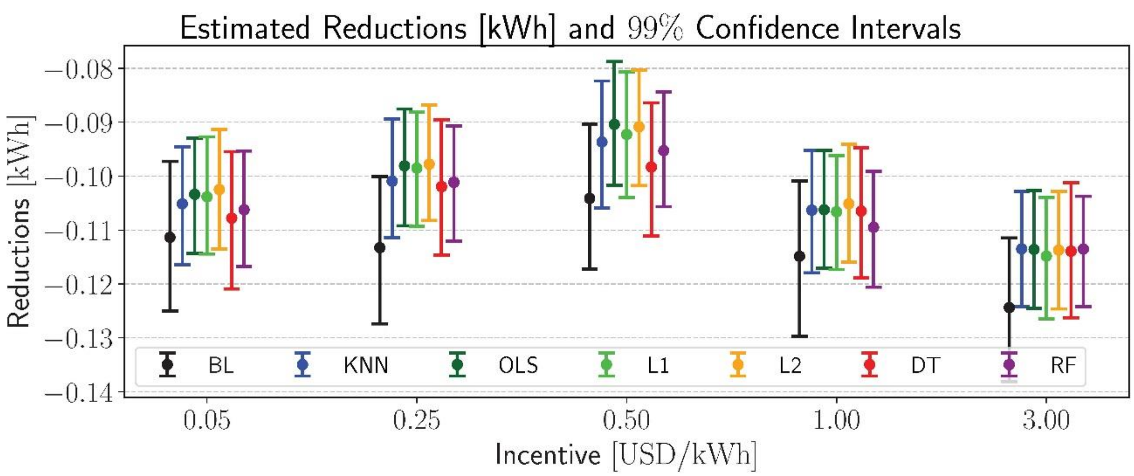
Figure E-56: Comparison of Synthetic ITEs and their Estimations for Various Estimators



6.4 Non-Experimental Estimation Results

Having validated non-experimental estimators (E1)-(E6) on placebo and semisynthetic data in the previous subsection, we can finally report estimates of the average treatment effect. Figure E-57 shows ATE point estimates and their 99% bootstrapped confidence intervals conditional on differing reward levels for all estimators as well as the CAISO BL. Due to the empirical de-biasing procedure (see previous subsection), the point estimates for estimators (E1)-(E6) are close to each other. BL appears to be biased in favor of the users, as it systematically predicts larger reductions than (E1)-(E6).

Figure E-57: Comparison of Synthetic ITEs and their Estimations for Various Estimators



The ATE averaged over the predictions of estimators (E1)-(E6) is -0.105 kWh / 11.5%. The intercept and the slope of the demand curve are -0.099 kWh / -0.013 kWh/USD, suggesting that users reduce an additional 0.013 kWh per dollar offered, which is only a small change. Due to the idiosyncratic nature of the CATE for r = 0.5 USD/kWh, the slope and the intercept have to be interpreted with caution. However, the results give rise to a notable correlation between incentive levels and reductions.

To compare the prediction accuracy of the estimators, Table E-15 reports the width of the confidence intervals for each method and incentive level. The inferiority of the CAISO baseline compared to the non-experimental estimators, among which RF achieves the tightest confidence intervals, becomes apparent. Therefore, in the remainder of this paper, we restrict all results achieved with non-experimental estimators to those obtained with RF.

Table E-15: Width of CATE Confidence Intervals (kWh) by Incentive Level

Estimation Method	0.05 USD/kWh	0.25 USD/kWh	0.5 USD/kWh	1.0 USD/kWh	3.0 USD/kWh
RF	**0.0211**	**0.0210**	**0.0212**	**0.0211**	**0.0205**
OLS	0.0214	0.0217	0.0230	0.0219	0.0219
L1	0.0217	0.0211	0.0233	0.0219	0.0219
L2	0.0221	0.0214	0.0214	0.0219	0.0218
DT	0.0255	0.0251	0.0247	0.0241	0.0250
KNN	0.0218	0.0219	0.0235	0.0227	0.0214
BL	0.0277	0.0273	0.0269	0.0289	0.0266

Figure E-58 plots ITEs for a randomly selected subset of 800 users who received at least 10 DR events in Phase 1, estimated with RF. Users are sorted by their point estimates (blue), whose 95% bootstrapped confidence intervals are drawn in black. Yellow lines represent users with at least one active smart home automation device. By marginalizing the point estimates over all users with at least 10 events, we obtain an ATE of -0.104 kWh (-11.4%), which is close to -0.105 kWh as reported earlier. The difference ensues from only considering users with at least 10 DR events. The 99% ATE confidence interval is [-0.115, -0.093] kWh.

Figure E-58: Estimated Individual Treatment Effects of 800 Randomly Selected Users with at least 10 Demand Response Events in Phase 1, Estimated with Random Forest (RF)

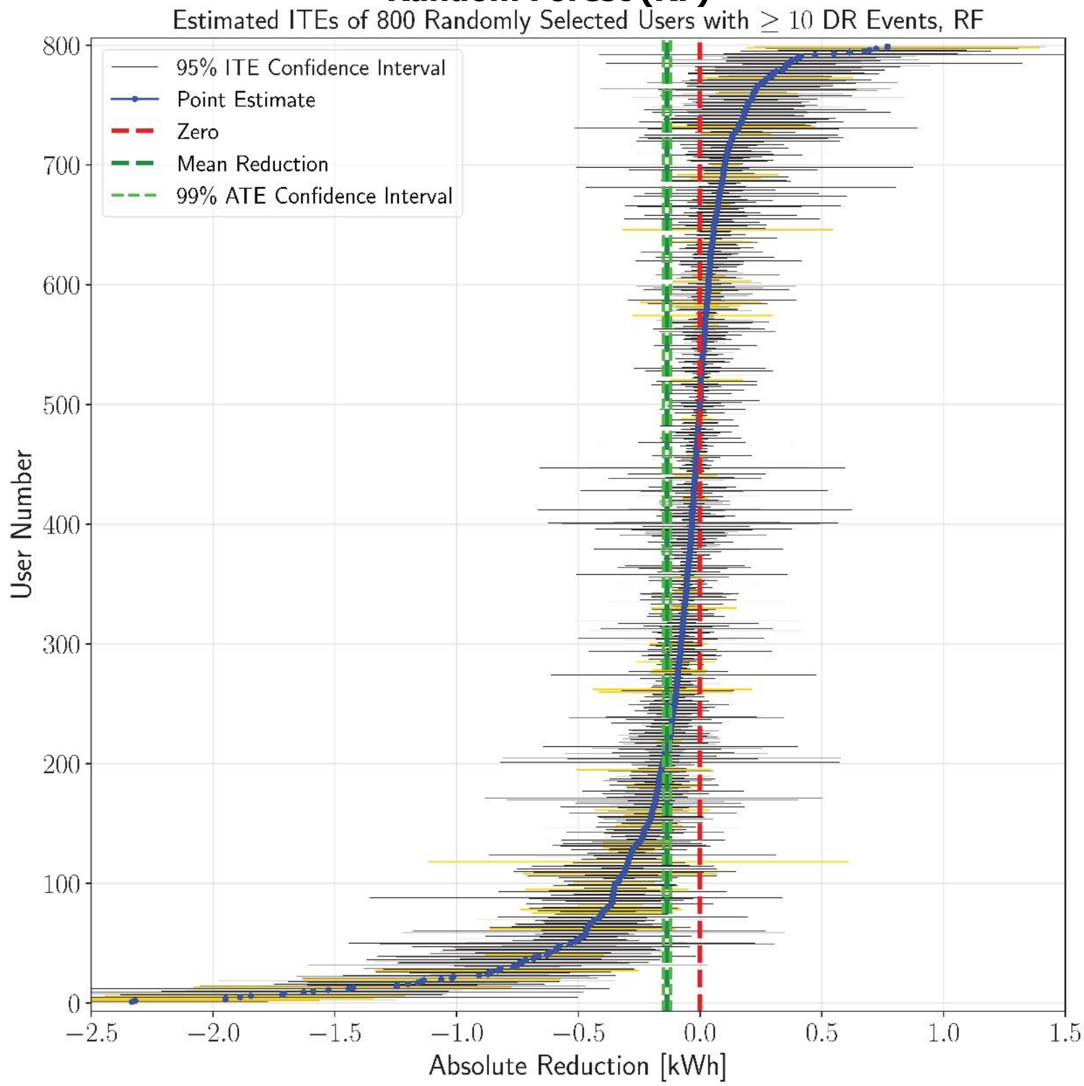


Table E-16 reports estimated ATEs for users with or without active smart home automation devices, which are obtained by aggregating the relevant estimated ITEs from Figure E-58. We notice larger responses as well as a larger percentage of estimated reducers among automated users.

Table E-16: ATEs Conditional on Automation Status for Users with at least 10 DR Events

	Number of Users	Percentage of Reducers	ATE (kWh)	ATE (%)
Automated Users	451	79.2	-0.279	-36.7
Non-Automated	4491	63.6	-0.087	-9.62
All Users	4942	65.0	-0.105	-11.5

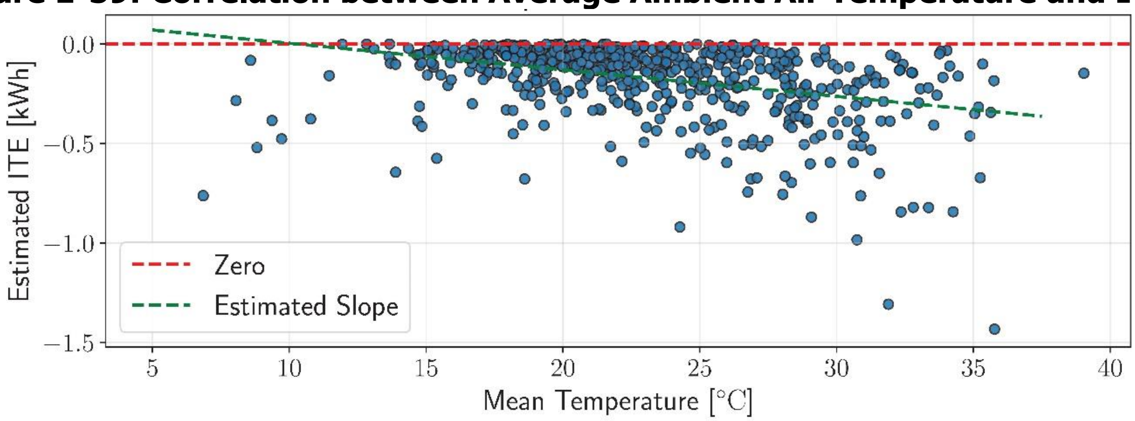
Next, Table E-17 reports the percentage of significant reducers for different confidence levels, obtained with the permutation test under the null hypothesis of no treatment effect. The results are obtained by using a permutation test. From Tables E-16 and E-17, it becomes clear that automated users show larger reductions than non-automated ones, which agrees with expectations.

Table E-17: Fraction of Significant Reducers among Sample of Size 4942

	$1-\alpha=0.90$	$1-\alpha=0.95$	$1-\alpha=0.99$
# Automated	225	205	159
% of Total	49.9	45.5	35.3
# Non-Automated	138 2	1162	829
% of Total	30.8	25.9	18.5
# All	160 7	1367	988
% of Total	32.5	27.7	20.0

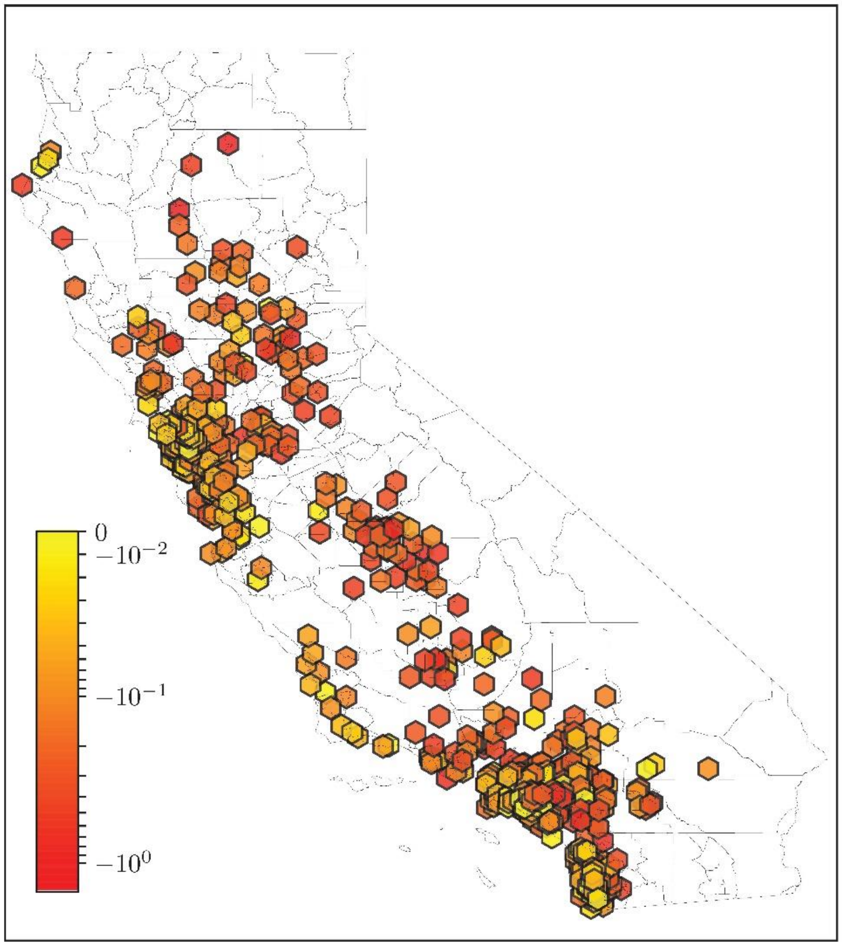
Larger reductions are estimated in warm summer months. To test the hypothesis whether or not there exists such a correlation, Figure E-59 scatter plots estimated ITEs as a function of the average ambient air temperature observed during the relevant DR events. This gives rise to a noticeable positive correlation of ambient air temperature and the magnitude of reductions. Indeed, a subsequent hypothesis test with the null being a zero slope is rejected with a p-value of less than 1e-9.

Figure E-59: Correlation between Average Ambient Air Temperature and ITEs



To support the results of Figure E-59, Figure E-60 color codes the CATT by geographic location. Each dot represents the average CATT among users for a particular ZIP code. The largest CATTs are found in the inland areas of California, which are considerably warmer than the coastal areas (in particular in the summer months).

Figure E-60: Conditional Average Treatment Effect by Geographic Location



6.5 Comparison of Estimation Methods

We now benchmark the results obtained from the best estimator (RF) to those from the fixed effects model described earlier. Figure E-61 compares the point CATEs by reward levels and their 95% confidence intervals. We notice that the point estimates are close to each other (-0.101 kWh) aggregated for fixed effects vs. -0.105 kWh for non-experimental estimate with RF, a difference smaller than 5%, a finding that suggests that our non- experimental estimation technique produces reliable estimates comparable to the experimental gold standard. The fact that the confidence intervals are notably tighter for RF corroborates this notion. As we have compared the prediction accuracy of various non- experimental estimators in Chapter 6.2, we can rule out the possibility of systematic underestimation of the variance in the experimental setting.

Lastly, Figure E-62 breaks out the estimates by month of the year. The point estimates are again close to each other, and it is visually striking that the width of the confidence intervals is noticeably smaller in the non-experimental setting.

Figure E-61: Estimated CATEs by Incentive Level with 95% Confidence Intervals

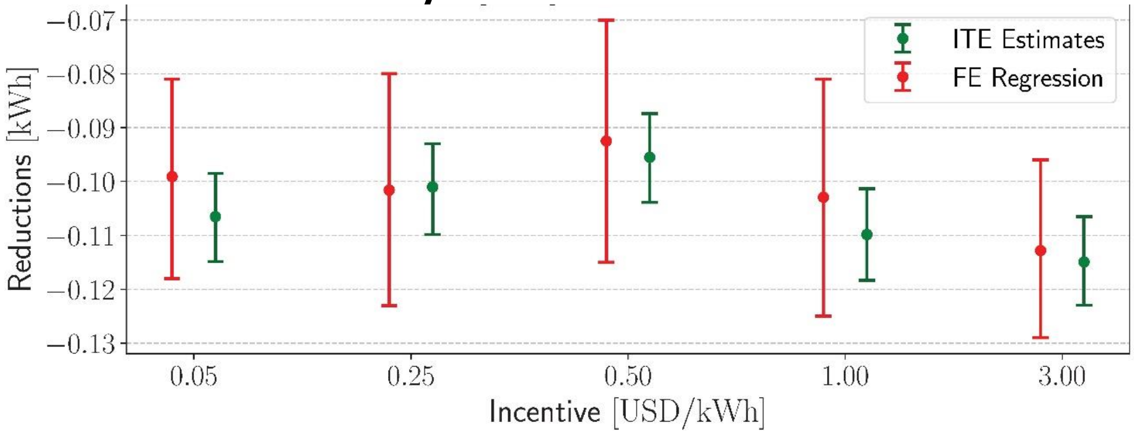


Figure E-62: Estimated CATEs by Month of the Year with 95% Confidence Intervals

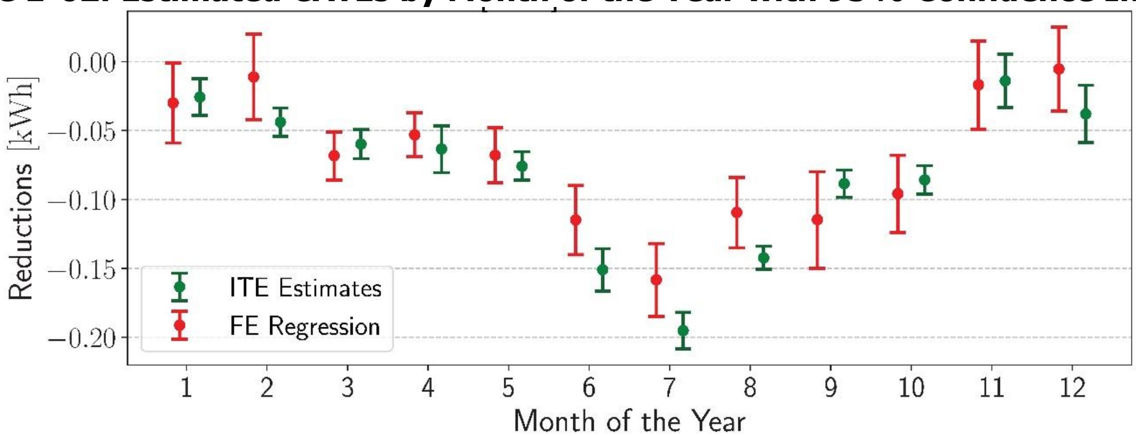
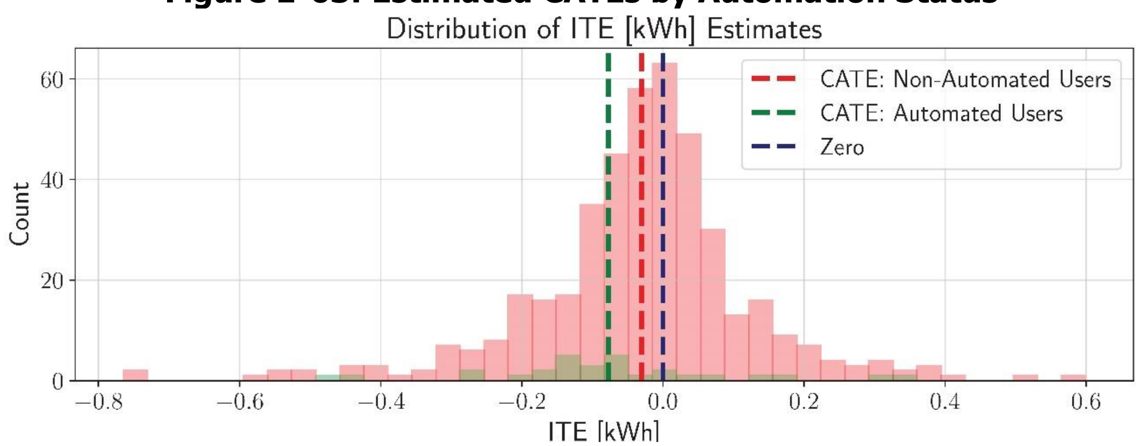


Figure E-63: Estimated CATEs by Automation Status



CHAPTER 7:
Survey

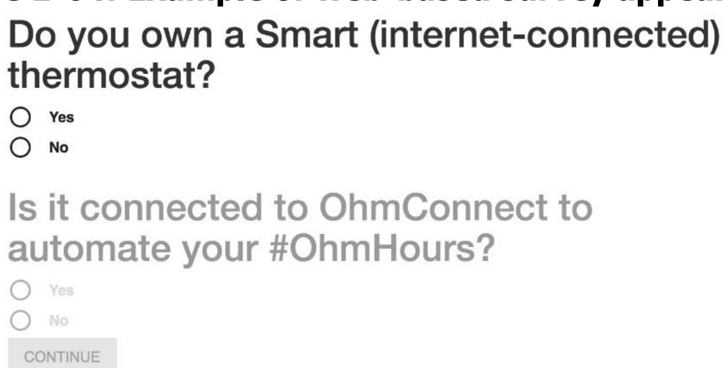
7.1 Survey Design and Questions

A web-based voluntary survey was offered to participants at the beginning of Phase 3 for Standard and Encouraged participants and two weeks after receiving their first #OhmHour for the Control Group. The survey was offered through an email with the following language:

> Subject: Tell us what you think and Earn $5 Plus a Change to Win $11 More
>
> Body: We're conducting a survey and we'd like your feedback. You'll get 500 points ($5) guaranteed when you complete this survey and you could earn an additional 1,100 points ($11) depending on your answers. The survey should take 5-10 minutes of your time.
>
> <Button with text “Get Started">

Figure E-64 shows a sample of how the web-based interface appeared. It was designed to look consistent with the OhmConnect product at the time of the experiment so that there would be less attrition. Each page shows a set of questions with conditional follow-ups.

Figure E-64: Example of web-based survey appearance



The complete list of Survey questions is provided in the following pages. Each “Page” refers to a separate web page. [IF YES] denotes a conditional follow-up question. Boxes indicate clickable answers and each page has a completion condition stated.

In total 672 households responded to the Survey.

Page 1. Do you own a Smart (internet-connected) thermostat?	Yes ☐ No ☐
[IF YES] Is it connected to OhmConnect to automate your #OhmHours?	Yes ☐ No ☐

Completion Condition: No or Yes + Yes/No

Page 2. Do you own an Electric Vehicle?	Yes ☐ No ☐
[IF YES] Is the charging station connected to OhmConnect to automate your #OhmHours?	Yes ☐ No ☐

Completion Condition: No or Yes + Yes/No

Page 3. Do you own any Smart plugs?	Yes ☐ No ☐
[IF YES] Are they connected to OhmConnect to automate your #OhmHours?	Yes ☐ No ☐
[IF YES] What do you have plugged into your Smart plug(s)?	
☐ Lighting	
☐ Heating or cooling appliance (fans, space heater, etc.)	
☐ Refrigerator or freezer	
☐ Cable box, television, or entertainment system	
☐ Pool pump or hot tub	
☐ Other	
[IF OTHER CHECKED] What do you plug into them?	*(text)*

Completion Condition: No or Yes + Yes/No + one checked box

Page 4. Do you currently own a centralized home automation system (i.e. Wink, Wiser Gateway, Smartthings, Insteon Hub)?	Yes ☐ No ☐
[IF YES] Is it connected to OhmConnect to automate your #OhmHours?	Yes ☐ No ☐

Completion Condition: No or Yes + Yes/No

Page 5. Do you currently use any of the following in your home? (check all that apply)

☐ Central air conditioning (AC)
☐ Wall-mounted air conditioning (AC) unit
☐ Central electric heating or electric floor heating
☐ Electric space heater
☐ Compact fluorescent (CFL) or LED light bulbs
☐ Hot tub
☐ Pool with an electric pump or heater
☐ None of the above

Completion Condition: At least one box checked

Strike the old "Are you the household member…" Page 6 question

Page 6. Out of the following, which would you rank highest as the way you most often save energy during #OhmHours? (check one)	
☐ Turn off lights	
☐ Turn off heating or cooling	
☐ Unplug or turn off small appliances (i.e. laptop, toaster, fans)	
☐ Unplug or turn off large appliances (i.e. TV, refrigerator)	
☐ Other	
[IF OTHER CHECKED] How do you save energy?	*(text)*

Completion Condition: At least one box checked

Page 7. Over the last 7 days, how often did you save energy in non-automated ways during #OhmHours? For example, turning off lights or appliances that are not plugged into automated devices.
❑ 1 (Never)
❑ 2
❑ 3 (About half the time)
❑ 4
❑ 5 (Always)

Completion Condition: One box checked

Page 8. Over the last 7 days, how often did #OhmHours help remind you to turn off appliances or lights that were left on accidentally? For example, you turned off lights in an empty room or the TV if it was unintentionally left on.
❑ 1 (Never)
❑ 2
❑ 3 (About half the time)
❑ 4
❑ 5 (Always)

Completion Condition: One box checked

Page 9. How much money is 100 points worth?
❑ $0.01
❑ $0.10

❑ $1.00
❑ $10.00

Completion Condition: One box checked

Page 10. How much was the last #OhmHour you received worth in points per kWh? Choose correctly and we'll give you 25 extra points!
☐ Zero points per kWh saved
☐ 5 points per kWh saved
☐ 25 points per kWh saved
☐ 50 points per kWh saved
☐ 100 points per kWh saved
☐ 300 points per kWh saved

Completion Condition: One box checked

Page 11. How many days ago was your last OhmHour? Choose correctly and we'll give you 25 extra points!
❑ Today
❑ Yesterday
❑ 2 days ago
❑ 3 days ago
❑ 4 days ago
❑ 5 or more days ago

Completion Condition: One box checked

Page 12. On average, how many points did you receive each #OhmHour over the *last* 30 days? We'll give you an extra 25 points if you come close to the correct answer!	*(integer)*

Completion Condition: An integer entered into the field

Page 13. On average, how many points do you *expect* to receive each #OhmHour over *next* 30 days? Remember, you'll get about 100 points per kWh saved below your forecast and lose points if you go over. We'll give you an extra 25 points if you come close to the correct answer!	*(integer)*

Completion Condition: An integer entered into the field

Page 14. Out of the following, which would you personally rank highest as a reason your household signed up for OhmConnect? (check one)
☐ Getting financial rewards for saving
☐ Environmental or sustainability concerns
☐ Donating rewards to a specific cause
☐ Entertainment/gamifying energy use
☐ Making my home Smarter and more efficient
☐ Grid reliability

Completion Condition: One box checked

7.2 Survey Sample Summary Statistics

Table E-18 shows summary statistics for the sample that completed the Survey and the full sample used in the Phase 1 analysis. In general, the demographic variables of the ZIP codes for the survey participants are not significantly different. However, it is apparent from the key outcome variables and temperature controls, that the respondents tended to consume less during the 90 days prior to enrollment and came from cooler areas with significantly more heating degree hours and significantly fewer cooling degree hours. Thus, it is likely that the responses are not representative of the sample more broadly.

Table E-18: Survey Respondent vs. No Response Summary Statistics

	No Response	Survey Respondent	p-value
Daily Consumption (kWh)	16.8	15.8	0.04
	(12.6)	(11.2)	
Max Consumption (kWh)	4.1	4.0	0.25
	(2.4)	(2.2)	
Mean Daily CDHs	42.7	21.5	0.00
	(50.5)	(36.2)	
Mean Daily HDHs	107.2	154.2	0.00
	(68.6)	(75.1)	
Pre-Enrollment Obs.	2045.9	2041.3	0.75
	(356.9)	(353.0)	
% HH Income < $25K	19.6	19.2	0.57
	(13.9)	(13.4)	
% HH Income > $200K	7.7	7.4	0.51
	(9.8)	(9.5)	
% Population Age 21-39	30.0	28.9	0.06
	(12.1)	(11.3)	
% Population Age 65+	12.0	12.8	0.03
	(8.1)	(7.7)	
% Population Black	5.8	5.5	0.43
	(8.5)	(8.3)	
% Population Asian/PI	15.5	15.2	0.73
	(18.1)	(18.0)	
% Population Latinx	30.3	30.2	0.96
	(22.6)	(23.2)	
Mean Family Size	3.2	3.3	0.62
	(0.5)	(0.5)	
% Family HHs	66.3	67.4	0.18
	(18.7)	(18.9)	
% HH Living w/ Children	46.5	46.1	0.58
	(15.8)	(15.5)	
% Population w/ Bachelors	21.8	21.8	0.88
	(12.5)	(12.5)	
Median Year Built	1974.2	1975.1	0.21
	(17.0)	(16.5)	
% Renters	49.3	46.8	0.04
	(26.6)	(26.2)	
% HHs w/ 3+ bedrooms	51.8	54.8	0.03
	(29.6)	(29.6)	
% Detached Units	53.3	55.7	0.12
	(33.0)	(33.3)	
% Electric Heating	28.1	27.2	0.23
	(17.5)	(17.6)	
Median Monthly Rent ($)	1331.6	1312.9	0.44
	(503.8)	(494.3)	
Households/Observations	4,819	672	

7.3 Survey Results

Pages 1-4 of the Survey ask about ownership of automatable technologies and whether the household has connected their technologies to OC's service. Figure E-65, Figure E-66, and Figure E-67 show the results for Smart Thermostats, Plugs, and Electric Vehicles, respectively. They break out take-up by the Encouraged and Non-Encouraged (Standard) Groups to show how ownership changed as a result of the rebate offer. In general, the Ownership rate was double to three times the rate of automation, indicating 10-15 percent of households who could be connected without purchasing. As a benchmark we include Electric Vehicles which also show an ownership gap, but do not show the effect of the Encouragement rebate as one would expect. Page 4 asked about home systems, but no Survey respondents answered yes to this question.

Figure E-65: Smart Thermostat Ownership and Automation

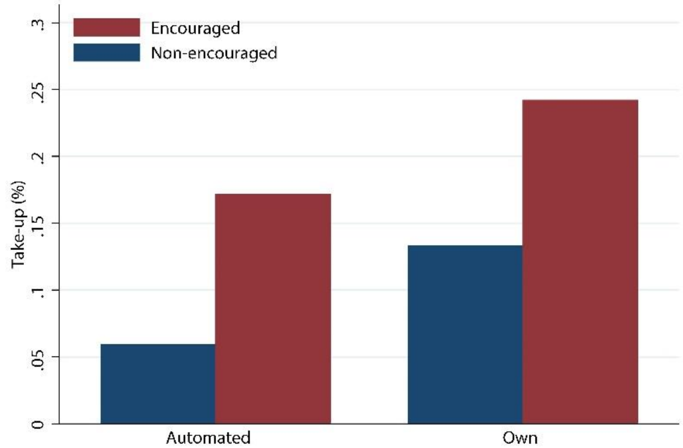


Figure E-66: Smart Plug Ownership and Automation

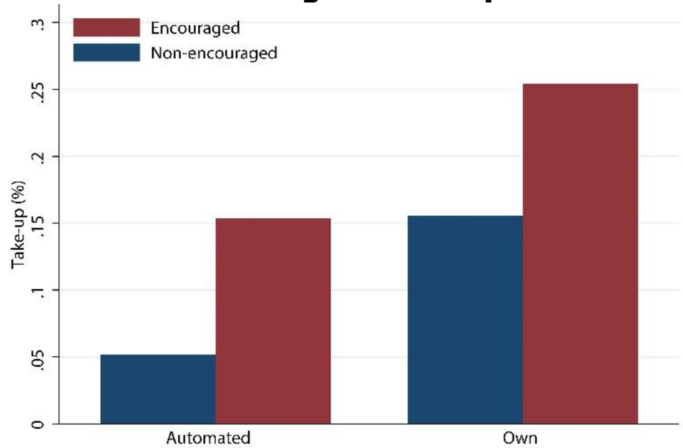
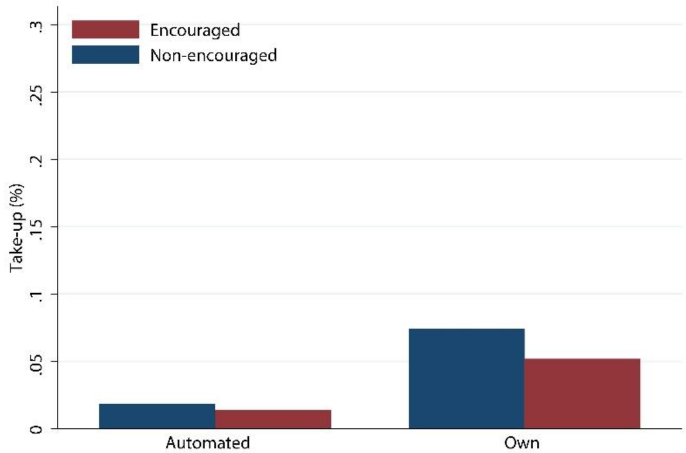
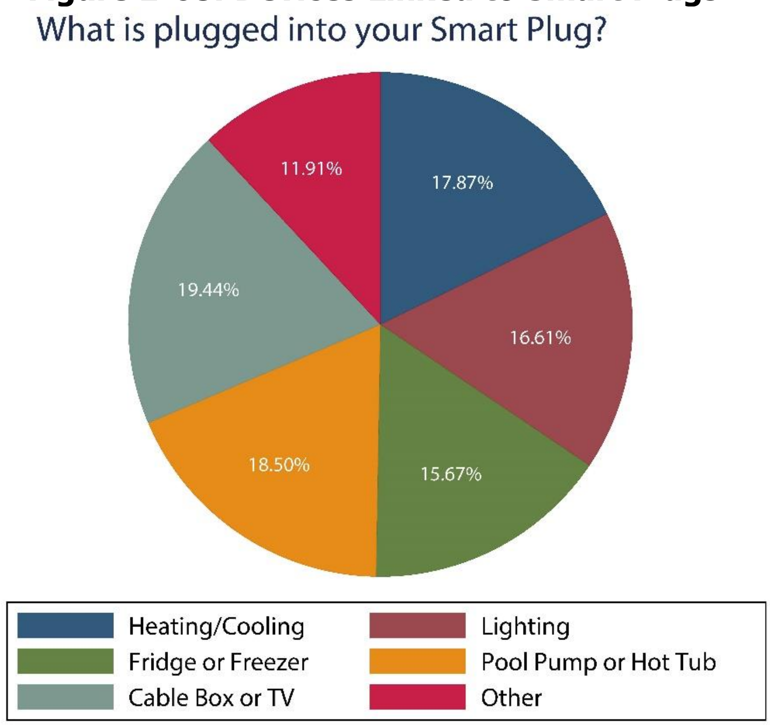


Figure E-67: Electric Vehicle Ownership and Automation



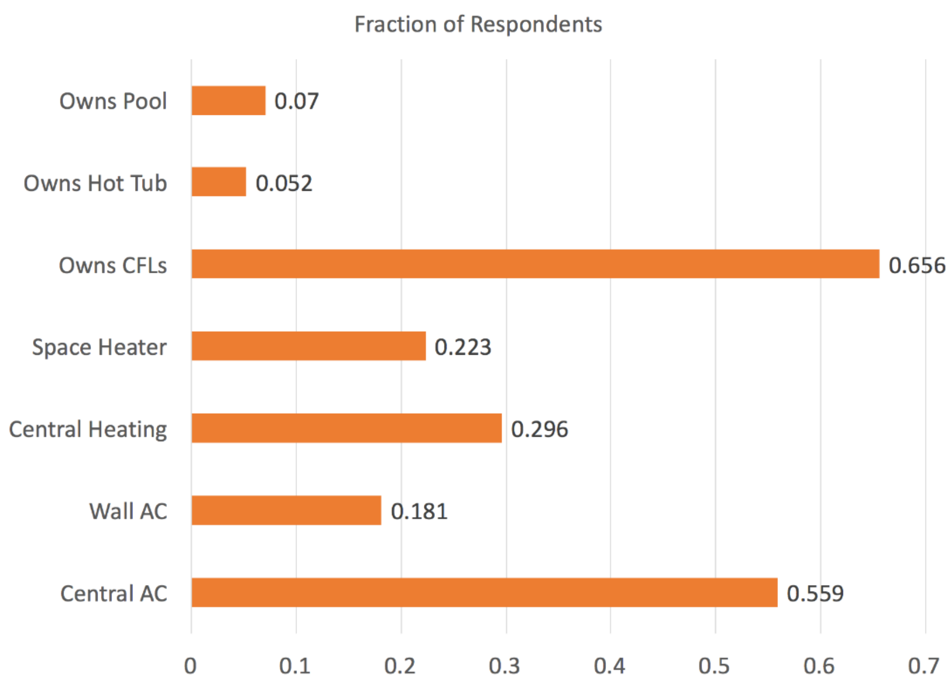
Page 3 also asked what individuals plugged into their smart plugs. Figure E-68 plots the results for households that owned Smart plugs.

Figure E-68: Devices Linked to Smart Plugs



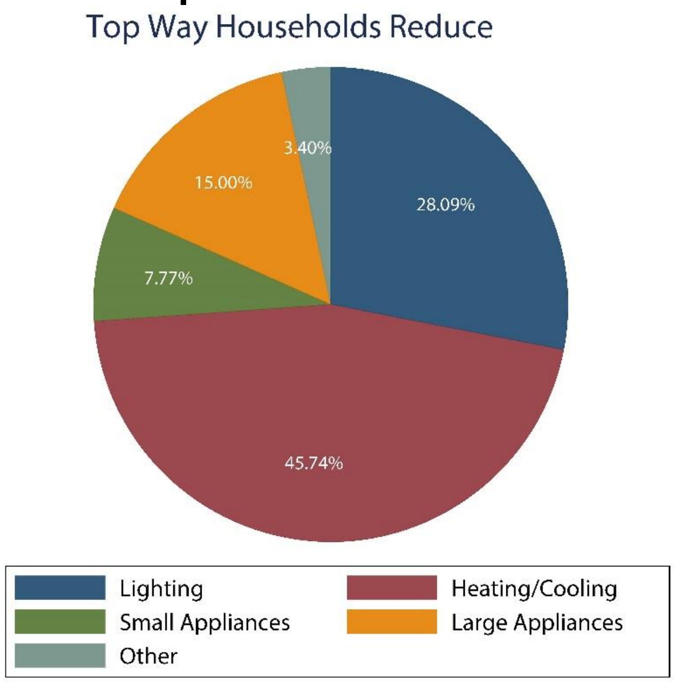
Page 5 asked about the fraction of respondents with various load sources. Figure E-69 reports the results of the question, showing the fraction of respondents answering yes to each load source.

Figure E-69: Types of Load Sources



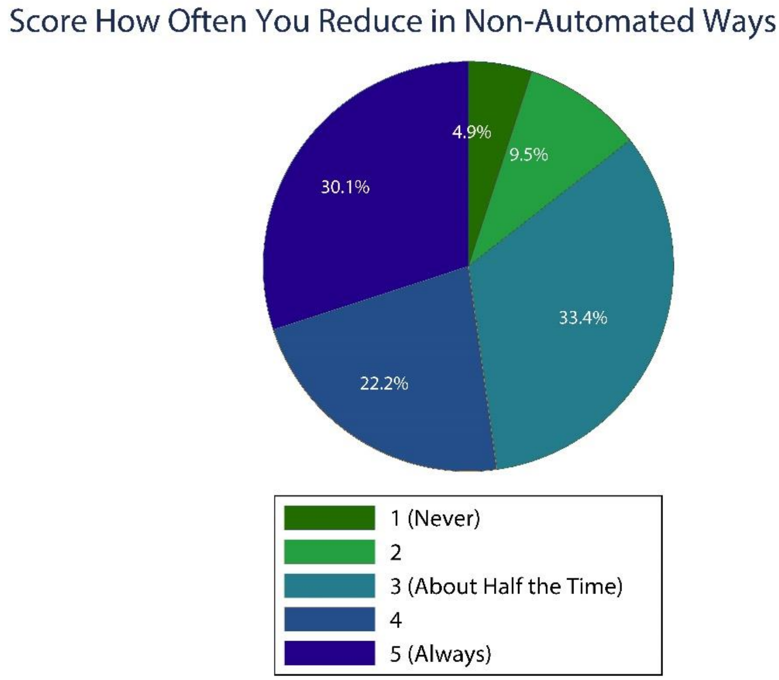
Page 6 asks respondents to rank the top way they reduce during DR events. The fraction is displayed as a pie chart in Figure E-70. The majority respond with heating and cooling which further validates the interpretation of the effect heterogeneity documented in Chapter 3.4 with temperature.

Figure E-70: Top Ranked Load Source for Reducing



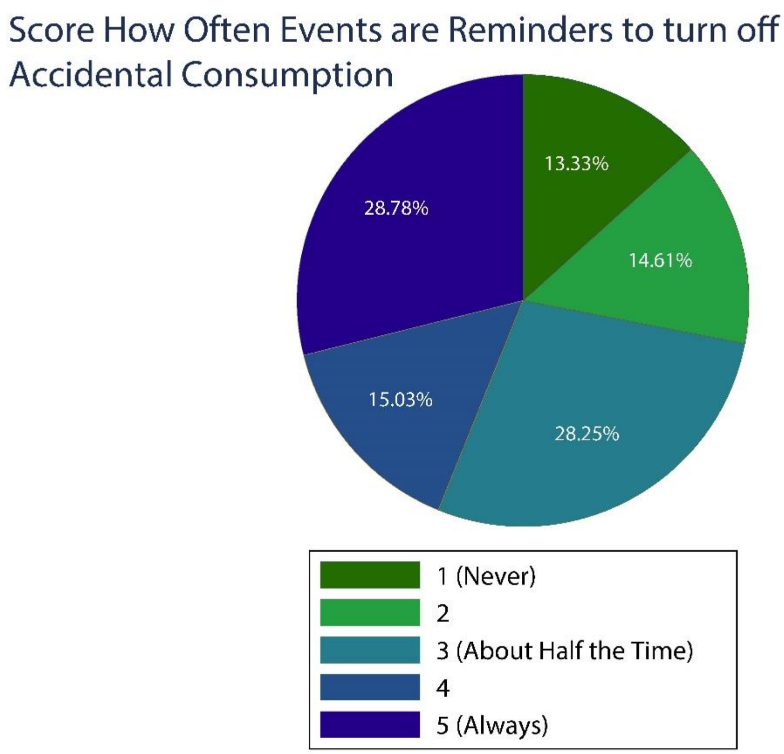
Pages 7 asked respondents to score on a scale of 1 (Never) to 5 (Always) how often in the last 7 days they recall reducing in non-automated ways. Figure E-71 plots the results by answer. The average was 3.6 with a standard error of 0.037, indicating the average consumer reduced in non-automated ways more than half the time. This is again consistent with the rates of automation take-up.

Figure E-71: Score How Often You Reduce in Non-Automated Ways



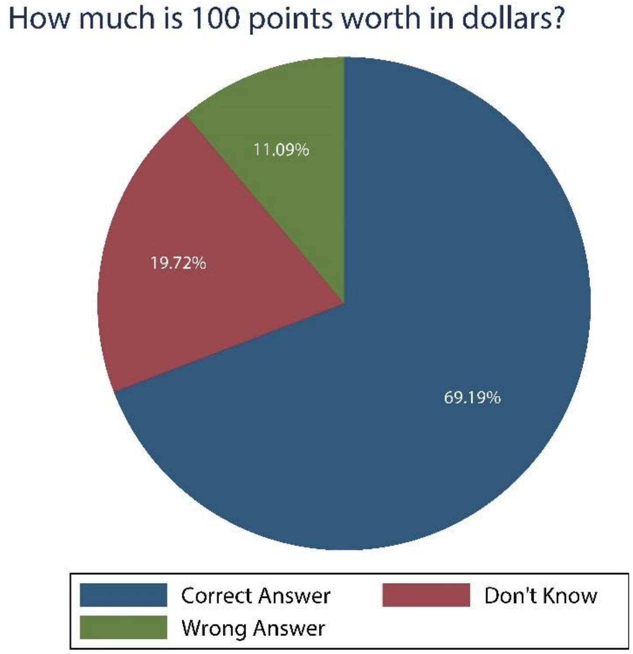
Pages 8 asked respondents to score on a scale of 1 (Never) to 5 (Always) how often in the last 7 days they recall DR events serving as reminders to turn off appliances they accidentally left on. Figure E-72 plots the results by answer. The average was 3.3 with a standard error of 0.045, indicating the average consumer was reminded to turn things off more than half the time. This suggests that much of the survey respondents are not being perfectly attentive in their electricity consumption.

Figure E-72: Score How Often DR Events Serve as Reminders to Turn Off Accidental Consumption



Pages 9-11 quizzed respondents on various aspects of the general and their personal experience with the product. Page 9 asked how much a point was worth. 69.2% answered correctly and 19.7% said they did not know. Figure E-73 plots the results.

Figure E-73: Correct Valuation of Point to Dollar Conversion



Page 10 asked what incentive level the household had last received. Only 18.9% got this answer correctly and 81.3% answered incorrectly.

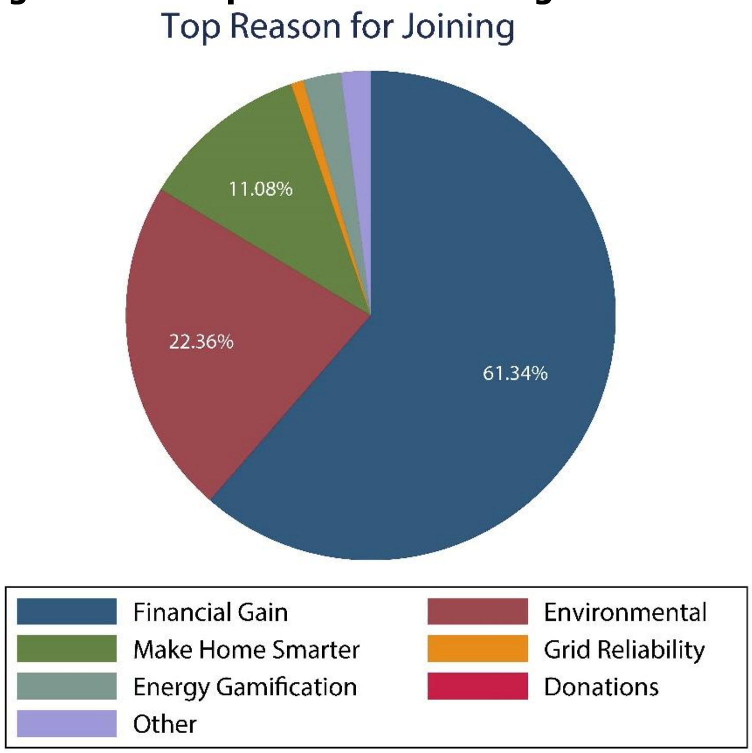
Page 11 asked how many days ago they had received their last #OhmHour. 31.5% of respondents answered correctly and 68.5% did not.

Page 12 asked households how many points they made in the last month. Actual points were strongly correlated with respondent's answers (r = 0.56), but generally did a poor job explaining the true variation (R-squared = 0.335).

Page 13 asked households how many points they expected to make in the next month. Actual points were very uncorrelated with respondent's answers (r = 0.05), and did an extremely poor job explaining the true variation (R-squared = 0.010).

Page 14 asked the top reason for respondents joining among a set of pre-specified choices. Figure E-74 plots the results. Financial Gain is the dominant choice with over 61% of respondents choosing it as the top reason. The second highest is environmental motivations, coming in at 22.36%.

Figure E-74: Top Reason for Joining OhmConnect



CHAPTER 8: Economic Valuation

This section attempts to put an economic value on the DR events studied in this project. We look at the effect of different incentive levels sent during Phase 1 and quantify the kWh reduced by points paid for the sample and for groups of automated and non- automated users. We also examine what the energy return is to the automation technologies deployed by the rebates.

These valuations reflect assumptions made based on the best information available to the authors of this report. The manner in which DR events were called for the grant may differ from how they would be called by OC in the future or when not required by the grant to call messages in a specific fashion. Further, we evaluate the wholesale market value from the energy markets and do not include any value provided by capacity contracts. We also do not have solid information on the cost to OC of calling DR events. Thus, the numbers here are incomplete and should be interpreted as such.

8.1 Value of Events to Household

First we calculate the private value of Phase 1 events to the participant. The value to the household can be broken down into the points paid for reductions and the avoided retail electricity costs. Since we find little evidence of spillover reductions or increases in non- event periods, we only consider the value during event periods. We assume an average price of electricity of $0.16/kWh because we could not identify participant's billing data. For high users, this may be an underestimate and for low users this may be an underestimate due to the tiered rate structures many customers are likely to be on. Still this provides a rough approximation of the dollar value the customer saved.

Table E-18 reports the dollar savings for different segments of the experimental sample and different scenarios. Column 1 reports the average dollar savings for a household over the 90 days of Phase 1 DR events which randomized the price levels, giving an average incentive level of $0.96/kWh. Thus, the first number is the average kWh reduction of 0.120 as reported in Table 3.4.1 multiplied by $0.96/kWh plus $0.16kWh and then multiplied by 25 because that was the average number of messages sent during the period.

Columns 2 and 3 of Table E-18 calculate the average savings for a household who experienced two hypothetical experiments where they were exposed to incentives that reflected wholesale market price variation 100 times within a year. In column 2 we consider incentives reflective of average prices and in column 3 we consider incentives reflective of the highest prices as an upper bound to the savings if events were called during these periods. The California Independent System Operator (CAISO) that operates the wholesale market makes clearing prices in the Day-Ahead Market (DAM) and Real- Time Market (RTM) available to the public. We scraped information on DAM and RTM prices from CAISO's website for the three major Demand Load Aggregation Point nodes for the hour of DR events, the mean price per month, as well as the top 10 highest priced hours for each month. We average across these values as they are generally fairly consistent. We use the DAM prices in column 2 of Table 8.1.1 which gives an average of $0.03/kWh, but the numbers are similar for the RTM prices. We use RTM prices for the upper bound because they have higher peaks and the annual average is $0.34/kWh.

Lastly, recall that OC shared 80% of its revenue from its wholesale market operations with the customer and retained 20% for its business operations. Thus, we multiply these prices by 0.8 to get the savings for the customer and assume the responsiveness of customers is generally the same regardless of the incentive level shown.

Column 4 reports the number of households in the sample to provide a sense of the scale of savings for the experimental sample. For example, the aggregate savings for Phase 1 totaled $14,945 for the 4,448 households enrolled in the program and the maximum value of an experiment that targeted the hours with the highest wholesale market prices would have saved $21,350.

Table E-19 Savings for Phase 1 and Hypothetical Programs

	Experimental Events (90 days)	Average Wholesale (Annual)	Maximum Wholesale (Annual)	Number of Households
Full Sample	$3.36	$1.82	$4.80	4,448
Non-Automated	$1.96	$1.06	$2.80	4,108
Automated	$7.28	$3.95	$10.40	340
Auto Adopters	$17.84	$9.68	$25.48	96
1st Quartile (lowest) (0kWh – 0.34kWh)	$2.66	$1.44	$3.80	1,132
2nd Quartile (0.34kWh – 0.56kWh)	$2.94	$1.60	$4.20	1,119
3rd Quartile (0.56kWh – 0.90kWh)	$3.36	$1.82	$4.80	1,103
4th Quartile (highest) (0.90kWh – 5.96kWh)	$6.13	$3.33	$8.76	1,094

Row 1 of Table E-19 reports the average savings across all households in the sample who were not assigned to the control group. Rows 2-8 report breakouts by various subsamples using the treatment effect estimates for each respective segment. Rows 2-3 report the averages for non-automated and automated households using the non-causal decomposition of Phase 1 event responses. Row 4 reports the averages based on the causal effect of adopting automation as estimated by the encouragement design. Rows 5- 8 report the estimates by consumption quartile.

Generally, the calculations show that savings are reasonable for Phase 1 of the experiment, but these incentive levels represent higher than average prices. The two hypothetical programs show more modest annual savings for the frequency of events called. Automated households do save more, and adopting automation provides substantial savings, even with the lower assumed wholesale prices. Also the largest savings accrue to the largest consumers with the top quartile representing a significant portion of the savings due to the right skewed nature of the electricity consumption.

8.2 Value of Events to Wholesale Energy Market

To calculate the complete social value of the events, we require more information than was available at the time of this report. We provide the formula here to underscore the distinction between what we calculate and the complete social value measure. The complete social value measure would be:

$$Social\ Value\ of\ kWh = Consumer\ Value - Cost\ Reduction$$

where

$$Cost\ Reduction = Procurement\ Value + Capacity\ Value + Externality - DRP\ Cost$$

The externality component is the marginal cost of the electricity generation. This component is the unpriced costs of generating electricity that comes from burning fossil fuels which cause air pollution. We use an estimate of the marginal externality cost of electricity from Holland, Mansur, Muller, and Yates (2016), a peer-reviewed academic article. Holland et al (2016) report the damages caused by electric vehicles on the emission of five air pollutants for the California Independent System Operator (CAISO), taking into account a variety of detailed factors driving electricity generation. We convert the numbers in Table 1 to kWh using 0.32kWh/mi, the electricity to mile conversion for the car described. The numbers vary for hours 9AM-12PM from \$20/MWh to \$26/MWh with an average of \$24/MWh. We use the latter number which represents an avoided externality cost of \$0.024/kWh. While this does not count for hour to hour variation within the year, it provides an approximation of these additional benefits.

The DRP Cost of a kWh is also not observed by the researcher, but was communicated in conversation to be about \$3 per customer per quarter for OC, coming out to a cost of approximately \$0.10 per household event. Thus, it does not scale with the kWh so there is a slight abuse of notation in the Cost Reduction metric above. We reconcile this in our actual calculations below. It was noted by OC that this cost was for the experimental DR events and may not be reflective of the company's costs more broadly.

The cost to the utility is comprised of the energy cost and the marginal cost of transmission and distribution. The latter is likely to be small for the kWh quantity of the DR studied here, but for larger customer populations and during critical grid events it may be more substantial. Quantifying these costs is beyond the scope of this evaluation so we leave them open to assumption by the reader.

This leaves the cost of energy which can be roughly categorized as procurement and capacity costs. Despite there being a market for backstop capacity in CAISO, capacity costs are difficult to quantify due to the degree to which much of it is also functionally procured via bilateral energy contracts. Instead, we focus on the procurement costs which can be more readily quantified with same data from CAISO's wholesale energy market that was used in Chapter 8.1. Importantly, these prices do not represent the capacity value of the overall electricity procurement cost and these prices may be lower than they would otherwise be without the separate procurement of capacity. Another point worth noting is that the DR events called were timed to be biddable into the RTM, but it is unclear the degree to which OC participated in this market.

We quantify the value of procurement costs avoided as follows:

$$Procurement\ Value = Wholesale\ Price - Point\ Payout$$

Note that this is not the wholesale price OC received and the point dollar value does not capture the fact that some participants may not cash out their points. It also does not capture any additional rewards OC gives the user. Lastly, for the top 10 price hours we must assume the event responses are the same, which may not be the case. With these caveats, it does represent the approximate dollar value and an upper bound for the marginal unit of electricity avoided by calling these DR events. To summarize, the value we report here is the dollars provided in terms of procurement value and externality cost avoided:

$$Report\ Value\ for\ Event = kWh\ Reduced * (Procurement\ Value + Externality)$$

We multiply this by the number of events within the time period considered. The reader should note that this value omits many of the other costs mentioned above and does not consider the long-run impacts of reducing demand which would cause the need for less generation.

Table E-20 summarizes the short-run dollar value per household of a set of 5 point per kWh DR events. We assume a frequency of 100 per year because this was the approximate frequency with which events were called during the experiment and we use our estimates in Chapter 3 for the kWh reduction per household. We average over geographic areas which may misrepresent the specific nodal pricing, but provides an approximate value based on the broader market picture. We use DAM and RTM during events and then use the top 10 RTM prices to provide the upper bound.

Columns 1-3 show the price variation we consider. Generally, prices during events were fairly close to the mean price in the market over the entire sample. These number as considerably smaller than the highest prices from the RTM which can reach \$300- 500/MWh or \$0.30-\$0.50/kWh. The first number in row 1 of column 4 indicates the DAM value of energy procurement and externality reduction of about 33 summer events comes out to \$0.12 per household. This was calculated as 33*(0.18997)*(0.04418+0.024-0.05) = 0.12. The numbers in column 5 perform a similar calculation with the actual RTM price. The numbers in column 6 show the value if the highest prices are used. The treatment effect for kWh reductions is smaller for non-summer events so the dollar value for these 67 events is lower despite sending more events. The final row sums these two numbers, giving an average value of 100 DR events per household to be \$0.13-\$0.17 with a hypothetical upper bound of \$3.96. This highlights the importance of targeting the highest value hours to achieving the greatest impact. Lastly, the value calculated scales proportionately with responsiveness so for the largest consumers who respond nearly twice as much as the average, the value approximately doubles. Thus, targeting larger consumers and responders will likely increase the average value of the program.

Table E-20: Short-Run Value of 5 point per kWh Messages per Household

	Avg. Wholesale Event Price		Maximum Wholesale	Value from DAM	Value from RTM	Upper Bound on Value
	DAM $/MWh	RTM $/MWh	RTM $/MWh			
Summer ~33 events	$44.18	$39.06	$405.09	$0.12	$0.08	$2.40
Non-Summer ~67 events	$36.56	$33.17	$345.51	$0.05	$0.05	$1.56
All Hours ~100 events	$39.10	$35.13	$365.37	$0.17	$0.13	$3.96

We do not calculate the value for 25, 50, 100, or 300 point per kWh messages. All of these calculations yield negative values because the procurement cost is generally below these reward levels. This highlights the impracticality of sending messages with higher incentive levels in the CAISO region, but may be different in other markets with higher wholesale prices.

8.3 Automation Adoption Payback Periods

As a final exercise, we calculate the payback periods to the private consumer of adopting and the optimal rebate based on our social value calculations in 8.2. This provides information on how much savings are garnered to the consumer privately and what rebate would equal the social value of households adopting automation.

Table E-21 reports the numbers for private payback periods assuming the full cost of 2 Smart plugs is $80 and the Smart Thermostat is $198. Other more expensive thermostats were also available, and the payback periods will be longer for those models. We consider the two 100 events per year hypothetical experiments with average wholesale prices from the DAM during actual events and the highest 10 prices per month from the RTM to calculate savings to the household and assume that electricity prices remain constant going forward. Increases to electricity prices will reduce the payback periods as savings are more valuable. We assume a 6% discount rate and use the estimates on the kWh effect of adopting automation from Chapter 3. While these pool the effect of thermostats and the effect of plugs, we cannot disentangle the two. Thus, these numbers may not be perfectly representative of the technology. Lastly, the estimates only consider energy savings and do not count any additional comfort or utility the household gets from the new technology.

Our calculations in Table E-21 indicate that a customer paying a full cost of $80 and facing prices consistent with when the DR events in the experiment were called will garner discounted savings equal to the upfront cost in 12.3 years. If events are targeted during the highest price periods, the payback period is cut down to 3.6 years, a 71% reduction. Thermostats are much more expensive and for the actual wholesale prices, there is no positive payback period because discounted future savings are less than $198. Conversely, for the highest price periods, the payback would be around 11.2 years, which is generally within the lifespan of a thermostat. The final row considers a program that calls 50% more events to show how these additional savings would decrease payback periods assuming the responses were the same.

Table E-21 Payback Periods for Adopting Automation at Full Cost

	2 Smart Plugs Actual Prices	2 Smart Plugs Upper Bound	Smart Thermostat Actual Prices	Smart Thermostat Upper Bound
Up-Front Cost	$80	$80	$198	$198
Wholesale Prices	Actual DAM	Highest RTM	Actual DAM	Highest RTM
Payback Period for 100 events	12.3 years	3.6 years	None	11.2 years
Payback Period for 150 events	7.1 years	2.3 years	33.9 years	6.5 years

Lastly, we consider the discounted social value of the plugs using the same hypothetical experiments and assuming a lifespan for the devices. This provides some idea of the magnitude of rebate that would equal the social value as calculated in Chapter 8.2

Table E-22 reports these estimates. Again these numbers assume the response is the same for 2 plugs and one thermostat, which is likely not the case, but the research design cannot disentangle the effect of each individually. Nonetheless, the discounted social value of procurement is nontrivial for the 5 and 12-year lifespans we consider. The differences between the actual and the upper bound come from the higher procurement costs avoided if hours are called when the RTM prices are the highest. Two caveats are that these calculations assume the externality remains $0.024/kWh and if the electricity generation mix gets cleaner over the lifespan, these benefits will be overstated. Similarly, if wholesale prices fall in the coming years, the savings will be overstated.

Table E-22: Discounted Social Value Rebate Calculations

	2 Smart Plugs Actual Prices	2 Smart Plugs Upper Bound	Smart Thermostat Actual Prices	Smart Thermostat Upper Bound
Device Life Span	5	5	12	12
Wholesale Prices	Actual DAM	Highest RTM	Actual DAM	Highest RTM
Rebate (Discounted Social Value)	$23.76	$146.73	$46.80	$288.98

CHAPTER 9: Conclusion and Outlook

To conclude, this report documented several findings that should be interesting to regulators, DRPs, and academic researchers.

First, households respond to hour-ahead DR events by reducing their consumption on the order of 12-14%. This shows the potential for very short notice DR events to generate reduction.

Second, households appear to be insensitive to the variable pricing as studied here. This suggests without innovation to the messaging, there is little ability for varying incentive levels to marginally change household consumption. Future work should try to understand if conveying relative value or providing information on the dollar value of a kWh could change the insensitivity.

Third, this report shows heating and cooling load are the primary drivers for responses. This confirms the previous literature and the general motivation of many residential DR programs to get customers to reduce cooling load in the summer.

Fourth, offering rebates can increase take-up of automation technologies and this automation causes significantly larger responses to DR events. Future work should explore whether the generally low take-up rates were due to the rebate design studied here or if there are other costs to consumers when considering adoption.

Fifth, targeting events using an initial set of interventions can dramatically lower the cost of the program. Future work should understand the persistence of these types of targeting strategies.

Lastly, the social value of the DR events calculated here was small, but this number could change if DR events were targeted more effectively to high price hours. Future work should include more elements of the social value calculation and understand if targeting events is feasible.

REFERENCES

[Breiman et al. 1984]: L. Breiman, J. Friedman, C. Stone, and R.A. Olshen, "Classification and Regression Trees," CRC Press, 1984.

[CAISO 2014]: California Independent System Operator Corporation (CAISO): "Fifth Replacement FERC Electric Tariff," 2014.

[Hastie et al. 2009]: T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning. Springer New York, 2009.

[Holland et al. 2016]: S. Holland, E. Mansur, N. Muller, and A. J. Yates, American Economic Review, vol. 106 no. 2 pp.3700-3729

[Palensky and Dietrich 2011]: P. Palensky and D. Dietrich, "Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads," IEEE Transactions on Industrial Informatics, vol. 7, no. 3, pp. 381-388, 2011.

GLOSSARY

Term	Definition
#OhmHours	The proprietary term used by the implementer of the experiment, OhmConnect, Inc., for a demand response event.
2SLS	Two-staged least squares (2SLS) – A two-step regression analysis technique that extends ordinary least squares estimation to deal with correlation between the independent variable and the error term. The first stage of the analysis fits a model to isolate the "good" variation in the independent variable that is not correlated with the error term. The second stage estimates the regression model using these fitted values.
ATE	Average Treatment Effect (ATE) – The average effect or impact of an experimental treatment or intervention on an outcome for the population of interest. For example, the average reduction in hourly electricity consumption per household caused by OhmConnect calling a demand response event.
ATT	Average Treatment Effect on the Treated (ATT) – The average effect or impact of an experimental treatment or intervention on an outcome for those who receive the treatment. For example, the average reduction in hourly electricity consumption per household caused by a household receiving demand response event.
BL	Baseline (BL) – A forecast for hourly electricity consumption that is constructed using a statistical model in order to determine what consumption would have been in the absence of a demand response event.
CAISO	California System Independent System Operator – The non-profit independent system operator of the wholesale market for electricity that serves California.
CATE	Conditional Average Treatment Effect (CATE) – The average treatment effect (see ATE above in glossary) for conditional on treating a subset of the population. For example, the average reduction in hourly electricity consumption per household caused by OhmConnect calling a demand response event for households located in the Pacific Gas and Electric service territory.
CATT	Conditional Average Treatment Effect on the Treated (CATT) – The average treatment effect on the treated (see ATT above in glossary) for conditional on treating a subset of the population. For example, the average reduction in hourly electricity consumption per household caused by a household receiving a demand response event for households located in the Pacific Gas and Electric service territory.
DAM	Day-Ahead Market (DAM) – The wholesale electricity forward market operated by the California Independent System Operator in which market participants bid to provide electricity one day ahead of the delivery time when it will be produced and consumed.

Term	Definition
DR	Demand Response (DR) – The term used to represent a host of strategies meant to reduce the demand for electricity during periods when it is valuable to the grid.
DRP	Demand Response Provider (DRP) – An entity participating in the wholesale market or contracting with utilities to provide demand response. For example, OhmConnect, Inc. is a demand response provider.
DT	Decision Tree – The method used for predicting Individual Treatment Effects by learning a set of decision rules on observable characteristics which determine the target value.
E/NE/C Treatments	Encouraged (E) / Non-Encouraged (NE) / Control (C) – The names given to the treatment groups for Phase 1 of the experiment.
HT/LT/NT Groups	High-Targeted (HT) / Low-Targeted (LT) / Non-Targeted (NT) – The names given to different treatment groups for Phase 2 of the experiment. The High- and Low-Targeted groups were sent exclusively high and low incentive levels whereas the Non-targeted group received all incentive levels.
ITE	Individual Treatment Effect (ITE) – The effect of an experimental treatment or intervention on an outcome for an individual user or household over time.
IV	Instrumental Variable (IV) – The econometric technique used to isolate "good" variation in an independent variable when its raw value is correlated with the error term of the outcome variable. The strategy finds an instrumental variable (an "instrument") that is correlated with the independent variable of concern, but not the error term of the outcome variable. For example, the automation rebate offer was correlated with the decision to adopt automation, but not electricity consumption because it was randomly assigned.
KNN	k-Nearest Neighbors Regression (KNN) – A non-parametric analysis technique used for regression.
kWh	kilowatt-hour (kWh) – A standard unit of measurement for power used in electricity consumption.
L1	LASSO Penalty (L1) – A penalty term added to non-zero features in a Machine Learning method to promote sparsity of the model.
L2	Ridge Regression (L2) – A penalty term similar to L1 added to non- zero features in a Machine Learning method to promote sparsity.
MAPE	Mean Absolute Prediction Error (MAPE) – A metric used for evaluating the accuracy of a Machine Learning Method.
OC	OhmConnect, Inc. (OC) – The third-party demand response provider that implemented the experiment.

Term	Definition
OLS	Ordinary Least Squares (OLS) – The standard regression analysis technique used to fit the relationship between a dependent variable and a number of independent variables.
PG&E	Pacific Gas and Electric (PG&E) – One of California's three investor- owned utilities.
Points	The currency used on the OhmConnect platform that can be cashed out for financial rewards at the rate of: 1 point = $0.01.
RCT	Randomized Control Trial (RCT) – The term used for randomized experiment where one group of the sample is assigned to a control group that does not receive any treatment.
RF	Random Forest (R) – A collection of Decision Trees with the goal of minimizing the prediction error by using multiple predictions of the same target value.
RTM	Real-Time Market (RTM) – The wholesale electricity spot market operated by the California Independent System Operator in which market participants bid to provide electricity five minutes ahead of the delivery time when it will be produced and consumed.
SCE	Southern California Edison (SCE) – One of California's three investor-owned utilities.
SDG&E	San Diego Gas & Electric (SDG&E) – One of California's three investor-owned utilities.