Take the Load Off: Effort and Technology as Determinants of Electricity Demand Response*

Megan R. Bailey† David P. Brown‡ Blake Shaffer§
Frank A. Wolak¶

October 2026

Abstract

As electricity systems transition to greater shares of renewable energy, flexible demand becomes vital for grid stability. We test whether smart technologies can unlock this flexibility or if human behavior remains the key barrier. Our field experiment randomizes peak event timing across households enrolled in one of three demand response programs: fully automated, app-enabled, and manual. Automated households reduced peak consumption five times more than those requiring any action. Simply providing smart technology made no difference when effort was still required. Our results suggest that time and attention—not technology—are the main constraints to realizing greater demand-side flexibility.

*We thank the Editor, John List, and two anonymous referees for their numerous helpful comments and suggestions, as well as Jim Bushnell, Lucas Davis, Meredith Fowlie, Koichiro Ito, Katrina Jessoe, Ryan Kellogg, Robert Metcalfe, David Rapson, Mar Reguant, and attendees at the 2022 Electricity Camp in the Rockies, 2022 Berkeley Energy Camp, 2022 Canadian Resource and Environmental Economics Annual conference, 2023 ASSA sessions, the University of Alberta, 2023 Berkeley POWER conference, 2023 NBER Spring Environment and Energy Economics Meeting, University of California Davis for useful feedback. We are grateful to our partner electric utility company for sponsoring and managing the field experiment on which this study is based. This work is supported by excellent research assistance from Mallika Sharma. We acknowledge funding from our utility partner, the University of Calgary, the Canada First Research Excellence Fund as part of the University of Alberta's Future Energy Systems research initiative, the University of Calgary's Global Research Initiative, and the Social Sciences and Humanities Research Council Insight Development Grant. This experiment is registered in the AEA RCT Registry under ID #AEARCTR-0008680 and the UCalgary Conjoint Faculties Ethics Board (REB21-0161).

†Haskayne School of Business and the School of Public Policy, University of Calgary megan.bailey@ucalgary.ca
‡Department of Economics, University of Alberta dpbrown@ualberta.ca
§Department of Economics, University of Calgary blake.shaffer@ucalgary.ca
¶Department of Economics, Stanford University and NBER wolak@stanford.edu

1 Introduction

> *"While the electricity system of the past involved forecasting demand and dispatching supply, going forward grid operators will increasingly find themselves forecasting supply and dispatching demand."*
>
> –Jeff Dagle, Pacific Northwest National Laboratory

The energy transition is rapidly changing electricity systems around the world as supply decarbonizes and new sources of demand are electrified (e.g. heating, transportation). However, the growth of variable renewable generation sources, such as wind and solar, has raised reliability challenges for power grids, where supply and demand must balance at every instant and at every location. These market dynamics highlight the potential for flexible demand to play a critical role in facilitating the transition to a low-carbon electricity supply. Despite its promise, questions remain as to how to elicit such flexibility from notoriously inelastic electricity consumers.

One solution, long advocated by economists, is to send dynamic price signals to customers to reflect the time-varying nature of marginal electricity system costs (Boiteux, 1960; Kahn, 1970; Joskow and Wolfram, 2012). However, when facing time-varying pricing, customers face tradeoffs between the financial benefits, discomfort from adjusting consumption, and the cost of taking action which includes the time, effort, attention, and cognitive burden of interpreting and responding to price signals, making such first-best policies unrealistic (Schneider and Sunstein, 2017). In addition to the onus of having to gather (and understand) their electricity price schedule and usage, consumers face an even more fundamental obstacle: The reward for adjusting consumption may simply not be worth their opportunity cost of time (Becker, 1965). Consumers may be understandably reluctant to allocate their scarce time when the individual private gains are small—despite potentially large social gains in the aggregate. This "rational inattention" (Sallee, 2014) can lead to low price responsiveness and is reflected in inertia in consumer decisions.

In this paper, we examine the role of effort and attention as barriers to flexible residential electricity demand. In particular, we investigate the role that emerging technologies and the requirement to respond to events actively versus passively can play in reducing these barriers. Partnering with a large electric utility in Canada, we recruited approximately 1,000 households to a demand response experiment that lasted 17 months. Participants received individually randomized notifications of "peak events", roughly 3 to 4 per month, whereby they received financial rewards

for reducing electricity consumption during the 3-hour window of each event.

Participants were assigned to one of the three demand response programs that differed in the provision of enabling technology and whether active versus passive responses to events were required. Households in the most basic **Manual program** received information on their real-time consumption as well as notifications of peak events via the Utility's App and had to respond *actively* by manually adjusting their consumption, as the name suggests. Households in the **Tech program** received the same information but also had app-enabled load (electricity demand) controllers installed on their baseboard thermostats, hot water heaters, and/or electric vehicle (EV) chargers. These controllers allowed them to respond *actively* to events but with the ease of doing so remotely via an App. Finally, households in the **Central program** had the same load control technology installed and information provision as those in the Tech program, but with the key difference that the Utility automatically reduced their consumption in response to an event. In this sense, their response was *passive*. Even if they exert no effort, or pay no attention to the event, households in this program will have their consumption reduced and can earn rewards. Households could override the automatic reductions, but, in contrast to the Tech program, they needed to actively make an effort to *not* reduce consumption by pushing a button on their App.[1] To serve as a baseline, we also have a subset of never-treated households who did not receive peak events.

Our treatment of interest is the effect of peak events on electricity consumption and, importantly, how that differs across participants in different programs. Accordingly, we randomize our treatment events, i.e., peak event notifications, at the level of household-day. That is, each participant received a unique randomized schedule of peak events over the 17-month period of the study.[2] This granularity of randomization differs from typical "between-subject" designs, where randomization occurs at the level of group allocation, with participants subsequently receiving common treatments. Our design incorporates what is often called a "panel experiment" or "within-subject" design (Charness et al., 2012; Bojinov et al., 2021; List, 2025). Within-subject designs are far less common than between-subject designs, in part because of the complexity involved in implementing an experiment with personalized

[1]There are a number of recently developed programs that include utility-managed load-control for various appliances including hot water heaters (Wattersaver, 2023), thermostats (PG&E, 2023), electric vehicles (DTE Energy, 2022), and solar-plus-storage systems (Spector, 2020).

[2]The schedule is blind to the participant prior to the notification that occurs 21 and 2 hours preceding an event.

randomization.

The within-subject design allows us to estimate household-specific treatment effects, providing estimates of the full distribution of responses across households to peak events for each demand response program. This offers unique insights into the heterogeneity in price responsiveness across households, letting us further explore the role of effort and technology as determinants of effective electricity demand response. The within-subject design also generates additional variation through the individually randomized timing of peak events, increasing the statistical precision of our treatment effect estimates.

Our results provide four key insights. First, we find that households in all three programs respond to financial incentives but the requirement to respond actively is a major barrier to performance. Using both within- and between-subject variation, we find participants in the Central program reduced consumption, on average, by 26.2% during events, as compared to 4.8% and 5.3% in the active-response-requiring Tech and Manual programs, respectively. Having to take an action to reduce consumption, even one as small as pushing a button on an app to respond to an event, results in one-fifth of the demand response. Second, technology alone is not sufficient. Households in the Tech program, despite receiving remote control capabilities, performed similarly to those in the Manual program. Simply providing consumers with smart home technology to monitor and control their energy use appears insufficient to drive significant consumption changes. Third, the magnitude of the Central program's response reinforces the considerable potential of demand flexibility as a resource for electricity grids. Fourth, the Central program's automation during peak events is well tolerated, with an event opt-out rate of approximately 4%, and low attrition rates comparable to those in other programs.

We find that the distributions of household treatment effects, estimated solely based on within-subject variation, differ significantly across the programs. Participants in the Central program are symmetrically distributed around a mean reduction of 24%. However, participants in the Tech and Manual programs form a large mass around 0%, with a small subset of high performers driving mean program reductions of approximately 5%. Our within-subject design also allows us to examine the consistency of individual household treatment effects by exploring the variance of responsiveness across events within each household. We find that the estimated standard errors on the treatment effects are similar across the three programs, with slightly smaller values for the Central program. This suggests that households' re-

sponses to events are broadly consistent. That is, the Central program households are consistently large responders, while the Tech and Manual households are in large part consistently low responders. This consistency has important implications for demand response as a reliable resource for electricity systems.

Using our within-subject design, we decompose the consumption response into an extensive margin (*"Should I respond?"*) and an intensive one (*"If so, by how much?"*). Using a principal stratification approach, we find that the Central program has a large effect on the extensive margin: 80% of households in the Central program are identified as responders, compared to 20% and 18% in the Tech and Manual programs, respectively. On the intensive margin—the magnitude of response conditional on responding—the Central program again outperforms the active-requirement programs. Central program responders reduce consumption by 31%, which is statistically larger than the 27% and 23% reductions for Tech and Manual, respectively. The Central program not only boosts the likelihood of responding, it also increases the conditional magnitude of response.

We explore potential mechanisms driving our results by combining the household-level estimated treatment effects with detailed data on participants' interactions with the utility's App. We provide evidence suggesting that differences in performance are associated with differences in time allocation and attention. We find that when households do not interact with their App on event days, the average household-level treatment effect is roughly 2% for the Tech and Manual programs and 22% for the Central program. When households do interact with their App, the reductions increased to 6% for the Tech and Manual programs, and 25% for the Central program. These results emphasize that while all programs increased responsiveness when households interacted with the App, the Central program's "headstart" from not having an active response requirement is associated with the largest effects.

The app interaction data provides further insight into the connection between effort and performance during the demand response events. Placing all the treated households into performance quartiles based on their estimated household-level treatment effects, we note that Central participants dominate the top quartile. 74% of top quartile households came from the Central program, compared to 13% each for the Tech and Manual programs. Looking deeper into the relationship between a household's quartile performance and their app interaction frequency (i.e., the share of events during which the household interacted with the app), we note two key findings. First, app interaction is positively correlated with performance. Those who

delivered larger reductions tended to interact with the app more frequently, across all programs. Second, among households that achieved top quartile performance, those in the Tech and Manual programs had to interact twice as frequently as those in the Central program. High performance was possible in the active programs but required considerable effort.

We also investigate how price affects the degree of responsiveness by randomizing peak event incentives into regular and "high" reward events, with the latter offering roughly double the financial reward for a given consumption reduction. We find no statistically significant evidence that higher prices motivate greater consumption reductions over the range of prices tested.[3] In our context, while the doubling of prices is meaningful relative to wholesale power prices, the resultant impact on individual households, per event, is admittedly small; price sensitivity cannot be ruled out at more substantial price ranges. Nevertheless, this observed result suggests a story of responsiveness in which the costs of actively responding act as the key barrier, rather than a smooth response to price signals.

Our results point to the need to take an active action as a key barrier to responsiveness. This is consistent with a notion of adjustment/hassle costs (or "ordeals") resulting in a strong default effect. Default effects have been considered in a range of settings, including retirement savings (Chetty et al., 2014; Bernheim et al., 2015), organ donations (Abadie and Gay, 2006), subscription decisions (Einav et al., 2025), and health insurance (Brot-Goldberg et al., 2023; Shepard and Wagner, 2025). In this literature, the work most closely related to ours is Fowlie et al. (2021), who look at opt-in versus opt-out default effects at the household's choice of selecting time-varying retail electricity pricing plans. Our paper complements this work by focusing on default effects in the consumption response to events, among households already enrolled. Our key contribution beyond the existing literature is the finding of significantly greater responsiveness when consumption reductions are made the default, or passive, action in response to demand response events. Requiring customers to take action—even with the provision of technology that makes the associated cost as minimal as remote control with a mobile phone app—is no match for the power of demand response that is initiated on the consumer's behalf.

There is an emerging literature that explores automation of key household ap-

[3]This finding is similar to that of Prest (2020) who estimates the price-responsiveness of households facing time-of-use prices in Ireland (without automated technologies) and finds that households are responsive to time-of-use price signals, but the magnitude of response does not increase with higher prices.

pliances for consumers to reduce time- and effort-related costs. Evidence shows that automation of smart thermostats and electric vehicles can facilitate short-run demand responsiveness when combined with pricing (Harding and Lamarche, 2016; Bollinger and Hartmann, 2020; Burkhardt et al., 2023; La Nauze et al., 2024; Bailey et al., 2025a,b; Blonz et al., 2025; Garnache et al., 2025; Metcalfe et al., 2026). Our paper builds on this work by investigating the role of price, technology, and effort as determinants of demand response, using automation in a wider range of smart home technologies. In particular, by considering two programs, both with enabling technologies but only one including automation, we are able to isolate the relative impact of automation versus the technologies themselves. In addition, our within-subject design allows us to dig deeper into the mechanisms underlying these results. Our findings suggest that there is a threshold effect in which even a seemingly small amount of effort required to respond serves as a sufficient barrier to limit demand response.

Recent work finds that technology on its own may not deliver its full stated potential. For example, consumers may override important settings of demand response enabling technologies, reducing the anticipated benefits (Brandon et al., 2022). Consistent with this work, we find our Tech program performs no better than the Manual program on average. That is, given the ability to remotely control large appliances (e.g. thermostat settings), Tech households fare no better than consumers who require a more manual action. The key takeaway is that to reach their full potential, enabling technologies must account for how consumers actually engage with technology—including their reluctance—as a barrier to responsiveness.

The rest of the paper proceeds as follows. Section 2 describes our experimental design and data. We begin with descriptive analysis in Section 3, followed by our formal estimation framework in Section 4 and estimation results in Section 5. Section 6 presents empirical evidence to understand the mechanisms behind our results. Section 7 discusses the external validity of our experimental setting, and Section 8 concludes.

2 Experimental Design and Data

2.1 Overview

We partnered with a large Canadian electric utility (hereafter referred to as the "Utility") to create three demand response programs that vary in terms of the enabling

technology provided to customers, as well as household versus utility-initiated electricity demand changes on specific devices. We are primarily interested in how customers in each program adjust their electricity consumption in response to "peak events", times during which the Utility asks consumers to reduce consumption and rewards them financially for doing so.

A unique feature of our study is the fact that we randomized the timing of peak events at the household-level. Each household in the experiment has their own unique schedule of events. We leverage the randomization of these events (as well as the richness of our data) to estimate household-specific treatment effects. These features allow us to look at the distribution of effects within each program to better understand what is driving average program-specific results.

2.2 Treatment Events

Customers in each of the demand response programs received notifications of peak events through an electricity consumption management phone App offered by the Utility. Peak events had the possibility of occurring at one of two time periods: morning (7am to 10am) or evening (5pm to 8pm). The schedule and timing (morning or evening) of events were unknown ex-ante to the customer, each receiving a unique, randomized schedule of events over the course of the experiment. Consequently, households could not predict the day or event time when they would receive a peak event. Households received two event notifications – 21 and 2 hours before the event – that included an offer for households to receive financial rewards for reducing electricity consumption during the peak event period, relative to their household-specific baseline.[4] Further, because events were randomized in time for each household, they are not correlated with other drivers of household electricity consumption.

"Event types" were also randomized and were one of two pricing levels, "normal" and "high", with rewards increasing in the latter for large reductions. High peak events were only possible during evening periods. During normal events, households could receive \$1 for a 10% reduction, \$2 for a 30% reduction, or \$3 for a 50% reduction. During high peak events, households could receive \$1 for 10%, \$3 for 30%, or \$6 for

[4]Baselines were calculated based on a household's average consumption during the relevant event time window over the last five weekdays prior to the event, excluding days where events occurred. Customers did not know how the baseline was calculated to avoid the potential for customers to "game" their baseline (Wolak, 2007). 21 hours was selected as the longest notification period to avoid consumers having knowledge of a pending event while their baseline was still being set.

50% reductions.[5] By randomizing the pricing levels, we are able to estimate the effect of greater price incentives on household consumption decisions.

The incentive amounts translate to payments ranging from $1.11 to $2.22 per kWh of electricity reduced, for the average household.[6] These incentives are in the range of wholesale price caps used in a number of jurisdictions in North America.[7] From the household's perspective, the payments for a systematic 10%, 30%, or 50% reductions would be in the range of $3, $7, or $12 per month, respectively. This reflects 2%, 5%, or 8% of the average single family household's electricity bill of $156 per month.

Events occurred randomly on weekdays, excluding holidays. Households typically experienced two "normal" and one "high" event per month. This schedule was altered in the summer months of July and August when peak demand is typically lower in this part of Canada. During these months, households did not experience "high peak" events. Events started on February 22, 2022, and continued until June 30, 2023, resulting in 30,502 household-event days in our experiment.

Event notifications provided information on the time of the event and the financial incentives for the different demand reduction levels. Once consumers received the 21-hour notifications, they could also see event details in the App itself. See Appendix B.1 for examples of the notification and in-App event messages. Households' rewards for consumption reductions during events were displayed in the App at a two- to three-day lag. The App also gave households a summary of their total rewards to date. See Appendix B.2 for details on each program's in-App experience.

2.3 Demand Response Programs

Our initial sample consists of all households that first downloaded the utility's App. The App provides customers with household-level hourly consumption posted at a one-day lag. From the pool of approximately 9,000 households that joined the App, we identified 1,661 eligible households based on factors such as being located near

[5]Report dollars are in Canadian dollars. CAD$1 equals approximately USD$0.75 or EUR€0.69 as of June 2023, the end point of our experiment.

[6]The average household consumes 1.8 kWh in each hour in our sample. A 10%, 30%, and 50% reduction translates to a 0.54, 1.62, and 2.7 kWh reduction over the three-hour event, respectively. Consequently, for a 50% reduction during a normal peak event, we compensated households $\frac{\$3}{2.7}$ = $1.11 per kWh. For a 50% reduction during a high peak event, compensation was $\frac{\$6}{2.7}$ = $2.22 per kWh. The other percent reductions lie between these two cases.

[7]Examples include the wholesale price cap of CAD$1.00/kWh in Alberta (Brown and Olmstead, 2017), USD$3.50/kWh in the Mid Continent Independent System Operator that operates in the Midwest United States (IRC, 2017), and USD$5/kWh in Texas (Smith, 2022).

a major metropolitan city and household characteristics.[8] Eligible households were randomized into one of three demand response programs or two never-treated groups. Email invitations were sent to eligible households randomized into the demand response programs, with several reminders. See Appendix A.1 for a complete description of the recruitment and assignment process.

Table 1. Summary of Household Programs

Programs	DR Control	Load Controller	Price Incentive	Real-Time Usage Info
Central	Utility	✓	✓	✓
Tech	Household	✓	✓	✓
Manual	Household		✓	✓
Info	Household			✓
Control	Household			

Notes. DR Control represents whether demand response to events is controlled entirely by the household or by the Utility for the load-controlled devices. Load Controller denotes whether the household has load controller equipment installed. Price Incentive reflects if households receive peak events and rewards for reduced demand during events. Real-Time Usage Info denotes whether households receive real-time household-level consumption information. ✓indicates categories that are applicable to each program.

Table 1 summarizes our demand response programs and never-treated groups. Households in the **Manual** program earned financial rewards for demand reductions during peak events but did not have any load controller equipment installed by the Utility to manage their consumption via the App. They had to respond to events manually.[9] Households were sent a device that allowed them to monitor their real-time electricity usage information in the Utility's App.

The **Tech** program differs from the Manual program in that the Utility installed load controller equipment on one or more of the household's electric hot water heaters, baseboard thermostats, and level 2 EV chargers, to enable remote electricity consumption ("load") reductions. This equipment allows households to see device-specific

[8]Eligible households were those that were in or near a major metropolitan city for which it was feasible for utility-partnered electricians to install load control equipment; homeowners; not in condos or apartments; had at least one month of historical electricity consumption data as of September 2021; and had at least one controllable electrical device (level 2 electric vehicle charger, electric baseboard heater as the primary heat source, and/or an electric hot water tank).

[9]We can observe if a household independently installs its own equipment and links it to the Utility's App. Only 3 (out of 242) Manual households installed equipment independently. These 3 all installed smart thermostats that allow the monitoring and remote control of the household's electric baseboard heaters.

electricity consumption and turn on and off devices remotely via the App. While the Tech program households are equipped with load control technology to ease their effort in responding to events, they still must take an active action to do so via their phone's App. See Appendix B.2 for detail on the experience that participants had with the App, by program. Both the Manual and Tech programs allow us to test the efficacy of an active-response-requiring approach to demand-side flexibility—one without and one with enabling smart technology, respectively.

The **Central** program received the same equipment installed in their homes as the Tech program. However, during an event, the passive setting for Central program participants was for the Utility to manage their load-controlled devices by reducing electricity consumption. That is, without any active response, the Central program participants would reduce consumption via demand management initiated by the Utility. For example, during events, their thermostat temperature is automatically reduced, EV charging is delayed, and/or the water heater is turned off. Central program households needed to actively choose *not* to respond to an event (i.e., select out of utility management) by pushing a button on their App. The Central program allows us to examine the efficacy of centralized demand management and, as compared to the Tech program, the difference between having to take active action during an event versus passively responding by conceding control to a third party.[10] It is important to note that the Central, Manual, and Tech programs had symmetric information about their real-time usage which could be observed within the App.

Finally, we have two groups of households that serve as *never-treated* baselines throughout the study. One is a **Control** group that receives no intervention or messaging regarding the experiment after joining the utility's App. Another is an **Info** group that is identical to the Control group, but these households have access to real-time consumption information for their home via the App after they accepted and installed a device provided by the Utility. Both of these never-treated groups do not receive peak events or financial incentives. We passively monitor their consumption.

While households outside of the Control group needed to accept being in each program, the final sample of participating households are representative of those that utility companies are likely to be interested in until such time that demand response

[10] Customers in the Central program received advanced notifications of events just as other program participants, and the language of the notifications reminded them that the Utility would manage their devices. See Appendix B.1 for event notifications for each program. They had the option to opt-out of utility management for all devices ahead of an event or for individual devices during an event. These options are described in Appendix B.2.

programs serve as the default.

2.4 Acceptance

Table 2 summarizes the number of households invited and the acceptance rates for each program offer. Acceptance rates among all programs were high. In particular, the acceptance rate for the Central offer was 42%, and only marginally different than the acceptance rate for the Tech offer (48%). An equality of means test between these two values yielded a p-value of 0.072. Compared to Tech and Central, the Manual program had a statistically significantly higher acceptance rate of 59%, followed by the Info-only group at 68%.[11] Finally, Control had 100% acceptance because their participation was not subject to an offer. The acceptance rates of the Central and Tech programs were lower than the others due in part to the need for load controllers to be successfully installed in households that accepted these offers.[12]

Table 2. Program Acceptance by Program

	Central	Tech	Manual	Info	Control
Invited	423	382	409	259	188
Accepted	177	184	242	177	188
Percent Accepted	42	48	59	68	100
	(49)	(50)	(49)	(47)	–

Notes. "Invited" reflects the number of households invited to participate in the experiment, by program. "Accepted" is the number of households that accepted our offer and made it through the equipment installation process (as applicable, by program). "Percent Accepted" displays acceptance rates relative to the number of households invited; standard deviation for the Accepted indicator variable are reported in parentheses.

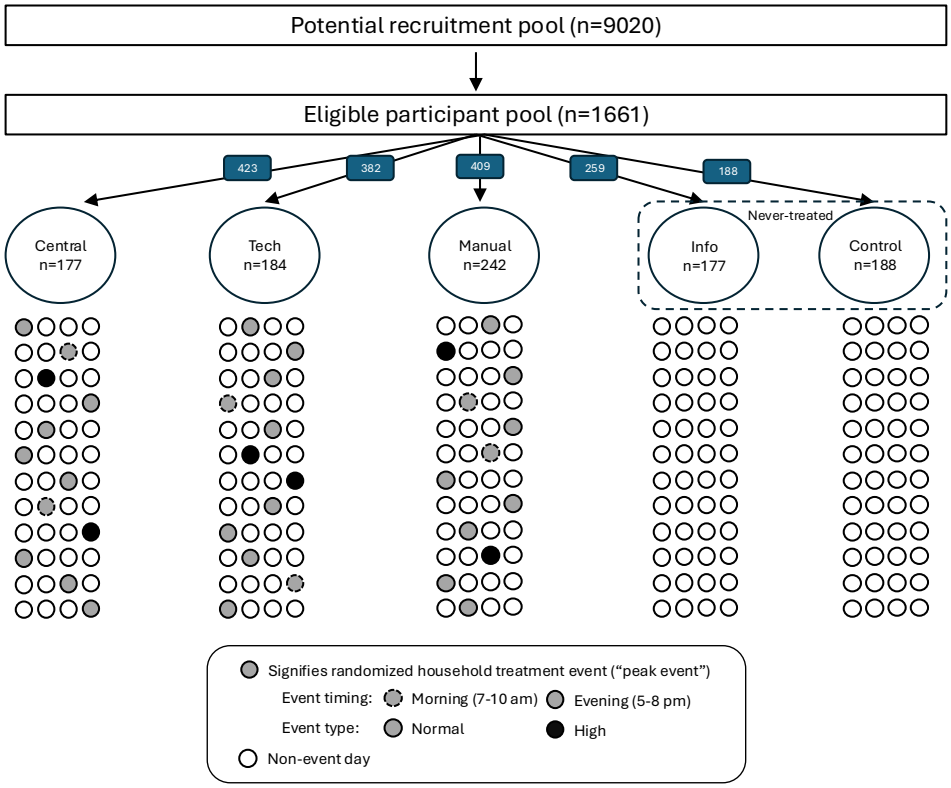
We take the similarity among final acceptance rates between the Central and Tech programs as the first set of evidence that we can compare our estimated treatment effects between these programs. While the Manual program had a higher final acceptance rate, concerns that the Manual program participants systematically differ from those of the other two programs are mitigated based on a comparison of observable characteristics across programs in our final sample, as are concerns about differences in the composition of households invited to each program. See Appendices

[11] Like the Manual program, the Info group required actively accepting the offer to join the experiment and installing a device that facilitates the monitoring and reporting of real-time consumption.

[12] We observed unsuccessful installation at households that initially accepted these offers due to, for example, households never responding to subsequent inquiries to receive and install equipment or households not being in compliance with local electrical codes to allow installation of devices.

A.2 and A.3 for a detailed discussion of balance using pre-treatment data. We address concerns over selection further in Section 4.4.

Figure 1. Experimental Design Summary



Notes: Dots represent individual household-days. Treatment events are indicated by filled dots, which are randomized at the household level, i.e. running vertically in this diagram. Dashed contours on treatment events indicate morning events, solid contours indicate evening events. Grey fill indicates "normal" event types and darker fill indicates "high type" (i.e. larger incentive) events.

Figure 1 provides a summary of our experimental design, starting with our initial recruitment and eligible participant pools, and subsequent allocation and acceptance into our demand response programs and never-treated groups. Next, we illustrate our within-subject design with a column of dots for each household (limited to four per program here, solely for exposition), with each dot representing a day in the experiment and shaded dots indicating randomly allocated peak events. Event variants, such as normal versus high reward days and morning versus evening events, are indicated based on the color and line type of the shaded dots. This figure highlights how the randomly allocated peak events occur on different days for the treated households,

although the distribution of events (and variants) is similar across households. We leverage this within individual randomization in our identification strategy outlined in Section 4 to identify the average treatment effect of peak events on the population of households that accepted participation in our demand response programs. We further evaluate if responses vary by peak event type (i.e., morning, normal evening, high evening) by program.

2.5 Data Description

For all households in our experiment, we track hourly household-level consumption (in kWh) from October 1, 2020 until June 30, 2023. We also have information on a number of household characteristics, such as household appliances, that were provided through survey responses as a necessary condition to enter the first phase of our recruitment process. In addition, the Utility provided supplementary household information, including the type of dwelling (e.g., single-family/duplex, row home) and an approximate geographical location.

We have access to hourly consumption (in kWh) for installed load controllable devices at households enrolled in the Central and Tech programs. We are also provided time-stamped information on household interactions with the Utility's App at the daily level for the Central, Tech, and Manual program households. The App data also include opt-out behavior for the Central program households allowing us to see if they override automated control during peak events. We use these supplementary data to better understand the types of behavior and consumption adjustments made during peak events.

We complement the detailed household-level data with demographic information from the 2016 Canadian Census (Statistics Canada, 2021a). We are provided a household's Census Dissemination Area (CDA) identifier; the CDA is the most granular geographical unit for which all Census information is provided publicly. We collect hourly weather information to control for environmental factors that impact electricity consumption, including temperature and humidity at three stations that are geographically representative of the households located in our study.[13] These data were accessed at Environment and Climate Change Canada.

In the last month of the experiment (mid-June 2023), we conducted an additional survey that contained questions on participants' experience with their respective demand response program. We provide a subset of questions from the survey in

[13]We match the households in our sample with their closest weather station.

Appendix D.1. We use these survey responses in our analysis in Section 6.4.

3 Descriptive Results

We begin our analysis with descriptive evidence that participating households reduce their electricity consumption during peak events and show how this response differs across demand response programs. Figure 2 provides average hourly household-level consumption for the Central, Tech, and Manual programs for the entire sample period during non-event (solid lines) and event (dashed lines) days. The shaded regions reflect the relevant event hours for each event type.

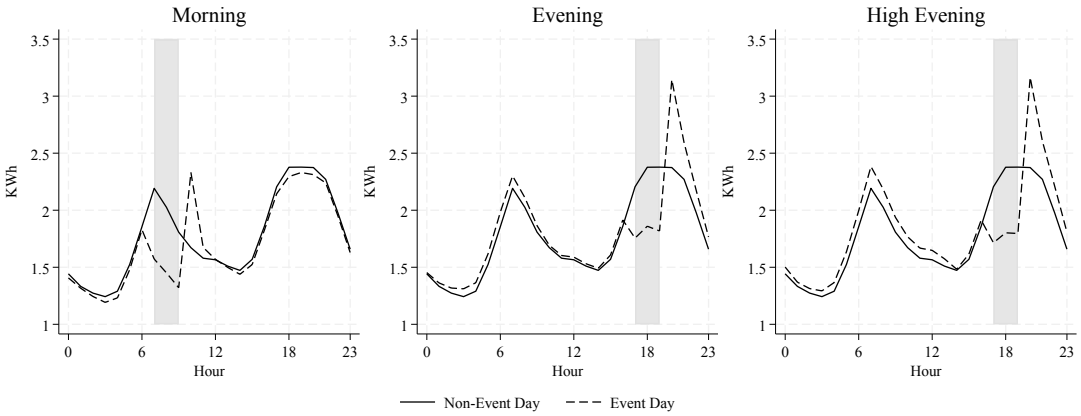
Figure 2a demonstrates that the Central program had a large reduction in average consumption during events regardless of the event type. After each event, we observe a large spike in consumption. This "snap-back" is consistent with the devices turning on immediately after the event (e.g., to reheat the water tank and/or home, or restart EV charging).[14] Comparing High Evening to Evening event consumption, we see no discernible difference in response to this higher reward.

Figures 2b and 2c demonstrate that the Tech and Manual programs show negligible changes in average consumption patterns during events. This limited observable response for the Tech program arises despite the fact that this program has access to the same equipment as the Central households. However, unlike the Central program where changes are automated, the Tech households must actively engage with the App or device to turn off the same appliances during an event.

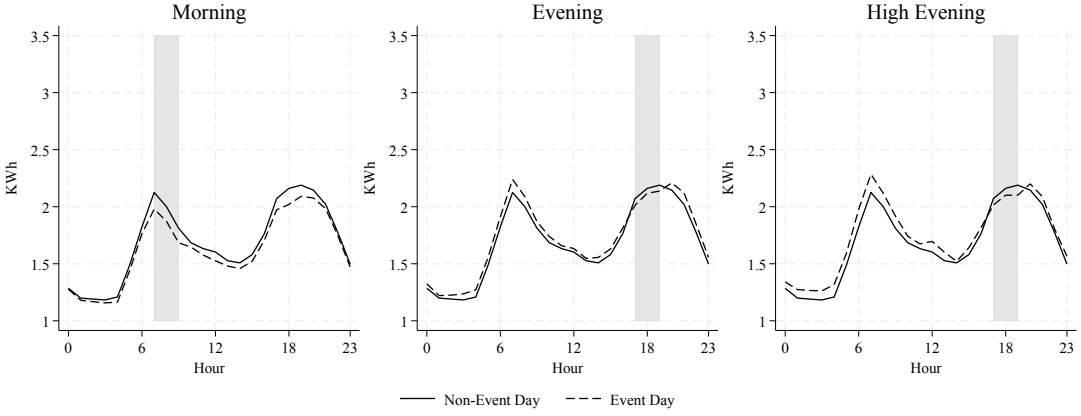
Taken together, these descriptive results suggest that the Central program has a considerably larger response to each event type. Further, we see no visual evidence of greater performance for greater financial rewards. Rather, the largest difference appears to be whether the household is in a passive versus active response-required program.

[14] Because we are interested in event-induced demand flexibility during peak event hours, not energy efficiency, we are unconcerned with demand being shifted to another time. However, based on conversations with the Utility, the new problematic peaks in demand created by the observed snap-back could be mitigated by the Utility staggering the beginning and/or end of the load-controlled event across households or only partially adjusting the demand levels on controllable devices. Managing the "snap-back" or "shadow peak" from demand response is an important area for further research. See, e.g. Bailey et al. (2025a); Turk et al. (2025) for further discussion of this issue.

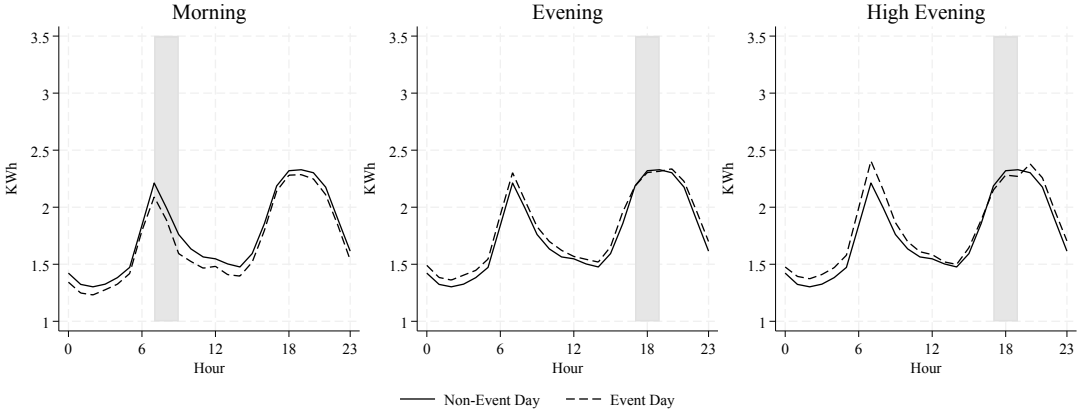
Figure 2. Average Household Consumption



(a) Central Program



(b) Tech Program



(c) Manual Program

Notes: Figure plots mean household consumption by hour and demand response program on weekdays, on event days and non-event days over the period February 1, 2022 - June 30, 2023. Event days are separated by type: Morning, Evening, and High-Evening. The shaded area represents the relevant 3-hour event period.

4 Empirical Framework

We begin by presenting our program-level regression specification that includes all households and programs, using random variation in peak event timing both across and within individuals. We then present a regression specification that leverages our within-subject design to estimate household-specific responses to peak events. This provides insights into response heterogeneity within a demand response program. We conclude by discussing the assumptions underlying our identification strategy and summarize a series of extensions and robustness checks to evaluate them.

4.1 Program-Level Regressions

We estimate the average treatment effect of peak events by comparing electricity consumption for households facing a randomized peak event to those who, contemporaneously, are not, using the following model at the household i and hour t level:

$$ln(c_{it}) = \sum_{j \in \{C,T,M\}} \left(\beta_j \, \text{Program}_{ji} \cdot E_{it} + \gamma_j \, \text{Program}_{ji} \cdot Spillover_{it} \right) + \alpha_i + \tau_t + \delta X_{it} + \varepsilon_{it} \quad (1)$$

in which $ln(c_{it})$ is the log of household electricity consumption, Program_{ji} is a categorical variable for which demand response program the household i is enrolled in (e.g., Central [C], Tech [T], or Manual [M]), and E_{it} is the household-specific event indicator that equals one if the household is (randomly) assigned an event in hour t. $Spillover_{it}$ is a vector that includes pre- and post-event indicator variables that equal one if hour t is two hours prior to and after a peak event, and zero otherwise.[15] These variables capture potential pre-event adjustments or post-event "snap-back" effects as devices turn on. A key advantage of this model is that it allows us to readily test for differences in responsiveness to events across the three demand response programs. We use the log of household electricity consumption on the left-hand side to account for the right-skewed nature of consumption.[16]

We include α_i, household fixed effects, which control for time-invariant household characteristics. We also include τ_t, an hour-of-sample fixed effect, which controls for time-varying factors that impact consumption. Household electricity consumption

[15] We considered a range of pre- and post-event period lags from 1 to 4 hours. Our results are robust as we vary the spillover window.

[16] Our results are robust to functional form; our key conclusions remain with a linear specification.

varies with local weather conditions (especially due to thermostat settings). To control for this, we include X_{it}, a vector of hourly weather controls that include the relative humidity and cooling degrees and heating degrees above and below 65° F (18.33° C). Since consumption behavior may vary with weather conditions in a non-linear way, we include a flexible functional form with a polynomial up to the third degree for each weather-related covariate. ε_{it} is a mean zero error term. We cluster standard errors at the household level.

We also consider a version of this regression specification where the event indicator variable, E_{it}, is adjusted to be a categorical variable for the three potential event-types: Morning, Evening, and High Evening. This analysis allows us to evaluate if households' responses to events differ by time-of-day and the financial reward for responding, in the case of Evening compared to High Evening events. In this extension, we have separate pre- and post-event spillover control variables for the hours surrounding morning and evening event windows.

Our main parameters of interest are β_j for $j \in \{C, T, M\}$, which measure the change in household-level electricity consumption during peak events for each of the Central, Tech, and Manual demand response programs. Because of our log-linear specification, we transform our estimates to report the percentage change in hourly consumption during an event via $100 \times (exp(\widehat{\beta}_j) - 1)$.

Our analysis compares event time consumption to non-event-time consumption of households in the same program, other programs, and never-treated households. This relies on the assumption that the households in the never-treated groups and other programs provide a valid counterfactual on non-event days. We examine the sensitivity of our estimates to the choice of non-event comparators, beginning with the most restrictive set—households within the same demand response program only—then augmenting the comparator group to include the never-treated households, and finally comparing against all treated and never-treated households. This robustness to different comparator groups is assessed in more detail in Section 5.3.

4.2 Household-Level Regressions

A unique feature of our setting is our ability to estimate household-level treatment effects from *within-subject* variation by leveraging the randomized event timing. This allows us to examine heterogeneity and distribution of event responsiveness across households, and to explore factors associated with household-level responsiveness that speak to the role of attention and effort/time allocated to responding to events.

Separately for each household i in the demand response programs, we estimate a household-specific treatment effect using the following model:

$$ln(c_{it}) \;=\; \mu_i + \beta_i\, E_{it} + \gamma_i\, Spillover_{it} + T_t + \delta_i X_{it} + \eta_{it} \tag{2}$$

where, analogous to above, $ln(c_{it})$ is the log of household electricity consumption, E_{it} equals one when household i has an event in period t and zero otherwise, $Spillover_{it}$ is a vector of indicator variables that equal one for two hours pre- and post-peak events and zero otherwise, and X_{it} includes the same set of temperature controls as the specification in Equation (1). In this specification, T_t is a set of time fixed effects that includes day-of-week, hour-of-day, and year-month to capture time-varying factors that impact consumption.[17] η_{it} is an error term, and we use heteroskedastic robust standard errors.

The regression in Equation (2) estimates a separate $\widehat{\beta}_i$ for each household in our experiment. The household-level regression compares consumption during event hours to non-event hours within the same household, conditional on time-based fixed effects and weather variables. Similar to the approach in our program-level estimation, we also consider a specification that allows for differential responses by event type (i.e., Morning, Evening, and High Evening). In this extension, we have separate pre- and post-event spillover control variables for the hours surrounding morning and evening event windows. We transform our household estimates to report the percentage change in hourly consumption during an event via $100 \times (exp(\widehat{\beta}_i) - 1)$.

4.3 Identification

Our framework falls within the literature on experiments that leverage "within-subject" variation (Charness et al., 2012; Bojinov et al., 2021; List, 2025). These designs involve a treatment of interest (i.e., peak events) that varies in time within a given subject. This differs from designs that use a *common* one-time treatment and between-subject variation to identify treatment effects. List (2025) articulates the key identifying assumptions that arise in this context and categorizes them as

[17] We cannot include an hour-of-sample fixed effect in the household-level regression because it would absorb the variation we are using for identification in this specification. Appendix C.6 presents results where we estimated our program-level specification in Equation (1) using this less granular set of fixed effects. The results closely reflect the estimates reported below. Appendix C.6 also demonstrates that our results are robust to alternative more granular calendar fixed effects at the year-month-hour-of-day level controlling for household-specific consumption schedules that are unrelated to peak events.

balanced panel, *temporal stability*, and *causal transience*. We address each in turn below; the corresponding empirical tests and robustness checks associated with each are presented in Section 5.3.

First, the *balanced panel* assumption requires subjects to remain in the study for all periods of the experiment and their outcomes to be observable. Attrition that impacts the balanced panel assumption can arise through two primary channels in our setting: (i) households explicitly request to leave the experiment or (ii) households close their utility accounts because they move. In either case, they are unenrolled from our experiment and their data are no longer observable.

Second, *temporal stability* rules out time-varying factors that are unrelated to the peak event itself, but could affect a household's estimated response to a peak event. Examples in our setting could include seasonal occupancy changes, changes in work-from-home patterns, and/or the adoption of new appliances (e.g., an EV) or the removal of appliances. Such changes can lead to adjustments in electricity usage in peak event and non-peak event hours, possibly differentially, leading to the within-subject design effectively comparing a different household pre- and post-change.

Third, the *causal transience* assumption in our setting requires that the effect of each peak event does not persist over time. This assumption implies that there are no "carryover" effects from previous treatments, the sequence of events does not impact behavior, and there is no sensitization (or desensitization) effect that could arise by being exposed to multiple treatments.[18]

In addition to these within-subject-specific assumptions, our experimental framework has several additional features relevant for identification. By randomizing events within an individual, events are not correlated with other drivers of household electricity consumption. We include weather controls to ensure that our estimated average treatment effects can be interpreted as weather-agnostic.

4.4 Estimand and Compliance

Our empirical framework provides estimated average treatment effects to peak events by program for households that accepted the demand response program offers. We define compliance as a household being assigned a randomized peak-event notification, irrespective of whether the household responds to the notification and/or permits

[18] List (2025) categorizes different types of within-subject designs depending on their threats to the temporal stability or causal transience assumptions. We view our analysis as having a relatively low threat of temporal stability, but a higher difficulty in satisfying all aspects that are entailed within the causal transience assumption, landing our analysis into the *Overt* design category.

automation in the Central program. The extent to which households respond to a peak event notification, and permit centralized utility control where applicable, is our primary feature of interest. Consequently, there is no non-compliance of our treatment of interest for households that are enrolled and remain in our experiment, and our design produces well-identified estimates of responsiveness among the population of participants. We address questions on the generalizability of these estimates in Section 7 on external validity.

5 Empirical Results

5.1 Program-Level Treatment Effects

Figure 3 provides the estimated average response to events by program as a percentage change in household-level consumption using the specification in Equation (1). We observe an average 26% reduction in consumption during events for the Central program. In contrast, the Tech and Manual programs reduced demand by approximately 5% on average during events.[19] Both of these effects are statistically different from zero. Even though the Tech program had the same equipment as the Central program, it demonstrated a much lower response to events. Additionally, the average response for the Tech and Manual program are not statistically different.[20]

Recall that participants in the Central program have the ability to opt-out of events using the App. We observe an opt-out rate of only 4% for these participants. When taken together with the results for the Tech and Manual programs, this low opt-out rate suggests that the large reductions for the Central program are driven in large part by allowing utility management of their devices during events.[21]

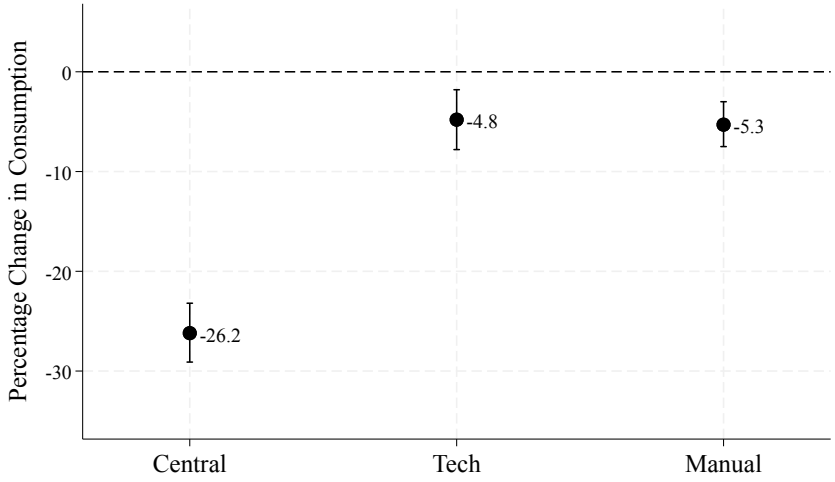
These results align with the descriptive data presented in Section 3 that suggest households in the Tech program did not use the load controller equipment to the same extent as the Central program. These results indicate that the installation of technology in the Tech program that enables remote control of household devices is

[19]The estimates from our Tech and Manual programs at roughly 5% are comparable to recent findings from Bernard et al. (2026) who find an average effect of 6.8% amongst over one million households who signed up for a behavioral demand response program in the United Kingdom.

[20]These main program-level results are estimated using all households for non-event comparators, including never-treated households, but are robust to varying the comparator households, including a restrictive version that only includes within-program households. Detailed robustness results can be found in Appendix C.3.3.

[21]Notably, when consumers opted out of central management, they generally did so for their thermostats: 90% of device opt-outs occurred by households adjusting thermostats during events.

Figure 3. Average Estimated Treatment Effects of Participants by Program



Notes: The reported results are program-specific marginal effects calculated from estimating $\widehat{\beta}_j$ in (1) for $j \in \{C, M, T\}$. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

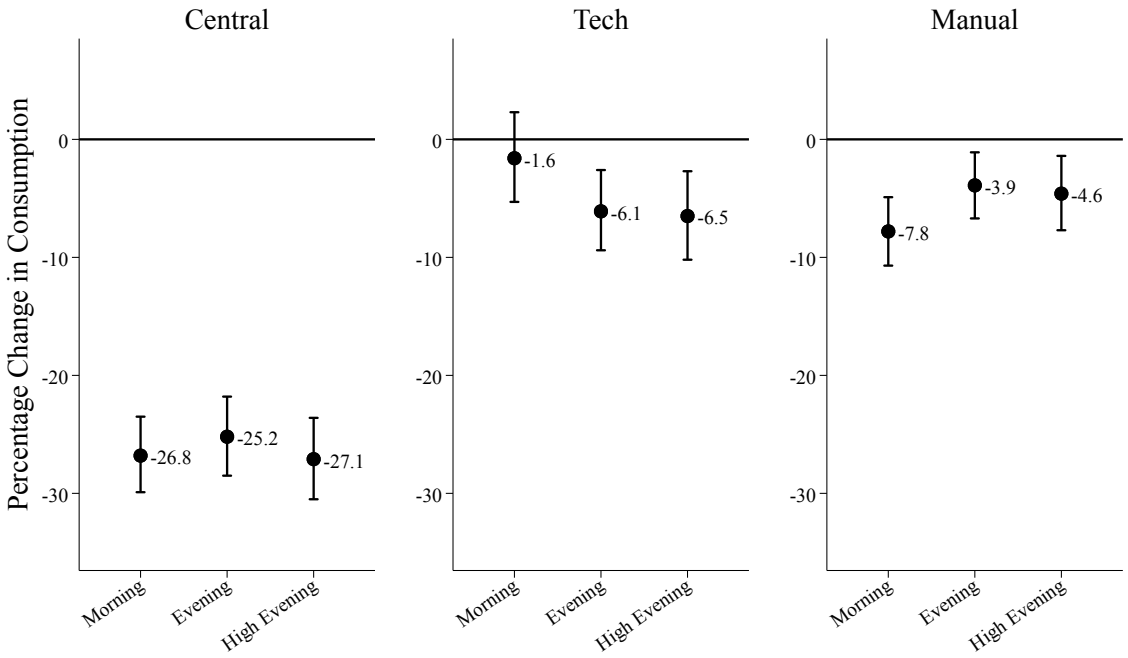
likely to be insufficient to resolve the time allocation/effort barrier that limits the provision of demand flexibility. In contrast, centrally managed demand response that flips the default from an active to passive response, thus requiring minimal-to-no effort by the consumer, results in large demand reductions. We present additional evidence using device-level usage data in Section 6.3 to support these findings.

In Appendix C.13, we present results for the pre-event and post-event spillover effects. Consistent with the descriptive results in Figure 2, we find evidence of a sizable "snap-back" post-event for the Central program. There is no evidence of spillover effects for the Tech or Manual programs. Using device-level data, we find the largest post-event increases in usage arise from the Central program's hot water heaters followed by thermostats, with no evidence of a snap-back from EVs.

In addition to randomized event timing, we also randomize event types, varying both the time of the event and the financial reward for reductions. This allows us to estimate how consumers respond to different price incentives and event times. Figure 4 presents the estimated response to events allowing for differential responses by event type. For the Central program, we see a large demand reduction for all event types, with an approximate 27% reduction during morning events, 25% during evening events, and a 27% average reduction during high evening events. This indicates that the Central households allowed central management of demand during both morning

and evening times. It also indicates that they were not distinctly more responsive to the greater incentives offered during the High Evening events.

Figure 4. Average Treatment Effect of Participants by Program and Event Type



Notes: The reported results are program- and event type-specific marginal effects calculated from estimating $\widehat{\beta}_j$ in (1) for $j \in \{C, M, T\}$, adjusted to allow for event-type interactions with the program indicator variables $Program_{ji}$. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

During the Evening and High Evening events, the Tech program reduced its demand by approximately 6 to 7%, while the Manual program had a 4 to 5% estimated reduction in demand during these event types. These effects are statistically different from zero. The Evening and High Evening Tech and Manual program effects are not significantly different from each other, when compared within each event type.

Looking at Morning events, we find the Tech program response is not statistically different from zero. This differs significantly from the Manual program, which had an average estimated reduction of 8% during the Morning events. However, these results are sensitive to the inclusion of more granular fixed effects. When we include flexible household by hour-of-day fixed effects, both programs' estimated Morning event responses are significantly different from zero, but not different from their estimated Evening and High Evening event responses (see Table C17). The other within and

across demand response program conclusions drawn from Figures 3 and 4 summarized above persist under this alternative specification.

We now consider the effect of the magnitude of incentives on responsiveness. For all programs, the change in consumption during High Evening events, where the reward for reductions was larger, is slightly more negative than for regular Evening events. However, none of the differences were statistically significant. This result is in line with the idea that the main barrier to demand responsiveness may have less to do with the scale of financial rewards and more to do with the hurdle of the opportunity costs of time/effort.[22]

That said, participants were only eligible to receive $1 more for achieving a 30% reduction in electricity use during a High Evening vs. Evening event. The reward for achieving a 50% reduction was doubled ($6 vs. $3). It is possible that a larger scaling of incentives could induce a greater response. However, given our rewards fall in the range of wholesale price caps observed in practice, it is unlikely that incentives provided in a real-world setting would be considerably larger than the amounts we provided.

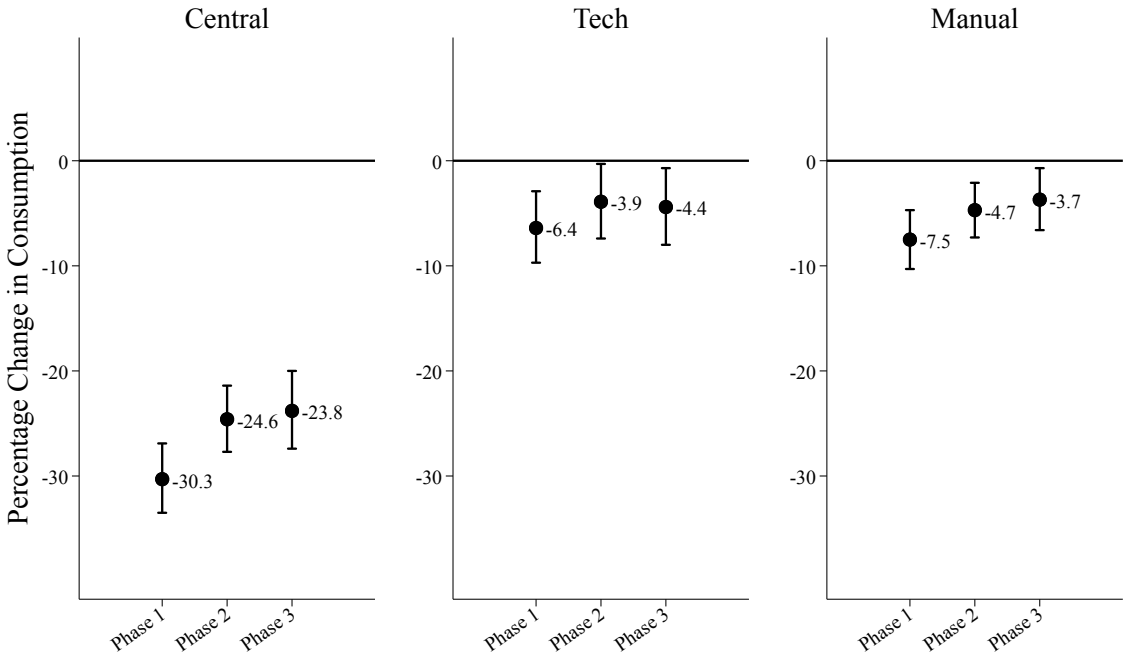
Next, we investigate if there are heterogeneous responses to events over the duration of our 17-month experiment. Figure 5 presents the estimated peak event treatment effects, allowing for differential effects across three phases of the experiment corresponding to roughly 6-month intervals that capture an early, middle, and end-phase of our experiment. For all demand response programs, we see a tapering in the estimated peak event response over time. For the Central program, there is a reduction from an estimated 30% effect during Phase 1, to approximately 23 to 24% during the latter two phases. For the Tech and Manual programs, we observe a similar tapering in the estimated peak event responses, with effects also stabilizing after Phase 1 to an approximate 4 to 5% reduction in demand. However, these differences are only significantly different from Phase 1 for households in the Manual program.[23]

While there is evidence of a modest reduction in responsiveness over time across

[22]This result persists despite observations raised by List (2025) that in a within-subject experiment, sensitization can lead to within-subject designs having elevated price-elasticities compared to a between-subject design. Further, the rewards in our experiment were quite salient as they were published within the App's "rewards screen" after each event (see Figure B8). One interpretation of these results is that it further illustrates the importance of automation to overcome key barriers to demand response in a setting with small-payout decisions.

[23]Figure C2 provides month-specific estimates of peak event responses by demand response program. These estimates parallel the findings in Figure 5, finding a modest reduction in the estimated response to peak events in the first couple of months of our experiment, with a lower degree of precision due to the large number of estimated coefficients.

Figure 5. Average Treatment Effect of Participants by Program and Phase



Notes: The reported results are program-specific marginal effects calculated from estimating $\widehat{\beta}_j$ in (1) for $j \in \{C, M, T\}$, adjusted to allow for Phase interactions with the program-event indicator variables $Program_{ji} \cdot E_{it}$. We consider three phases: (1) February 2022 - July 2022, (2) August 2022 - January 2023, and (3) February 2023 - June 2023. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

each of demand response programs, our key conclusions persist. For any phase of our experiment, the Central program's estimated peak event response is roughly 4 to 5 times larger than the Tech and Manual programs, with the latter two programs achieving similar estimated peak event demand reduction levels. For the Central program, the decline could be driven in part by behavioral changes where households were undertaking additional effort to reduce demand above the automated devices during the first phase of the program, or possibly for technical reasons related to the loss of connectivity of load controllers over time. For the Tech and Manual programs, the modest decline in the estimated response could be driven by fatigue, reduced effort to respond, and/or increased inattention.

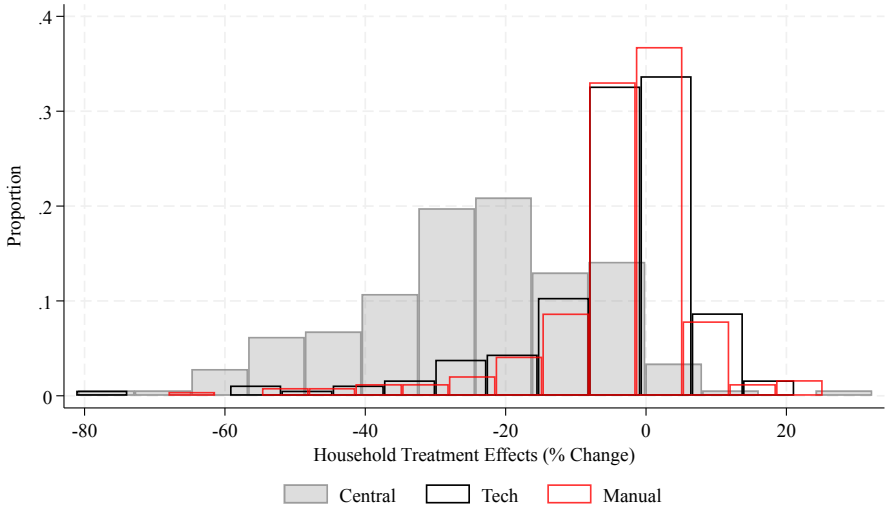
The causal transience assumption outlined in Section 4.3 requires that there are no sequence-dependent effects of peak events. While there is evidence of a modest change in the estimated treatment effects in the first couple of months, the richness in

our empirical framework allows us to empirically disentangle these effects and provide distinct estimates of the event treatment effects over time (see Section 5.3 for further discussion).

5.2 Household-Level Treatment Effects

Figure 6 presents histograms of household-level treatment effects by program, estimated as per Equation (2). We observe that the Tech and Manual programs' treatment effects are tightly distributed near zero. On average, households in the Tech and Manual programs reduce their consumption during events by 4.6% and 4%, respectively. This corresponds closely to the estimated program-level treatment effect in the previous section. Only 20% and 18% of the estimated treatment effects are negative and significant at the 5% level for the Tech and Manual programs, respectively.

Figure 6. Household-Level Estimated Treatment Effect Histograms by Program



Notes. The reported results summarize the distribution of estimated household-level treatment effects obtained from estimating specification (2), separately by demand response programs. The histogram presents the proportion of observations that fall within a specific bin's range. We present the marginal effects on the x-axis to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$.

Despite the fact that the Tech program has enabling devices, the *distribution* of their estimated treatment effects closely resembles that of the Manual households. In both programs, we observe a long left tail, indicating that there is a subset of households exhibiting large responsiveness. This is striking, as even though the Tech

and Manual programs have similar mean responses when looking across all households, we might expect that the Tech program would contain a subset of households with higher household-specific event responses given their ability to remotely control their hot water heaters, EV chargers, and/or thermostats during events.[24] There is only modest evidence to support this; the Manual households' estimated effects are more tightly distributed near zero. We dig into the heterogeneity across households, particularly the small subset of high performers, in more detail below.

Figure 6 also re-confirms the larger response from households in the Central program. The average household-level treatment effect is a reduction of 24% during events for households in the Central program, closely reflecting the estimated effects in the program-level regressions. There is a range of household estimates in the Central program, consistent with the fact that there is variation in controllable devices and heterogeneity in the use of these devices during event hours. However, only the largest estimates in the Manual and Tech programs achieve reductions that are in the range that is typical in Central. In contrast to these programs, 80% of the Central household-level estimates are negative and precisely estimated.[25]

Figure 7 plots the distribution of the standard errors of the estimated household-level treatment effects separately by demand response program. These results demonstrate that while the Central program has a slightly lower standard error on average, the distributions are similar across all three programs. The similarity in the precision of the estimates, combined with the average treatment effects noted above, suggest that the Central households are consistently large responders to events, while the Manual and Tech households are consistently low responders.[26] These findings provide important policy implications because they suggest that in addition to having larger estimated reductions in consumption during events, the Central program provides a reliable source of demand-side flexibility. Appendix C.8 provides addi-

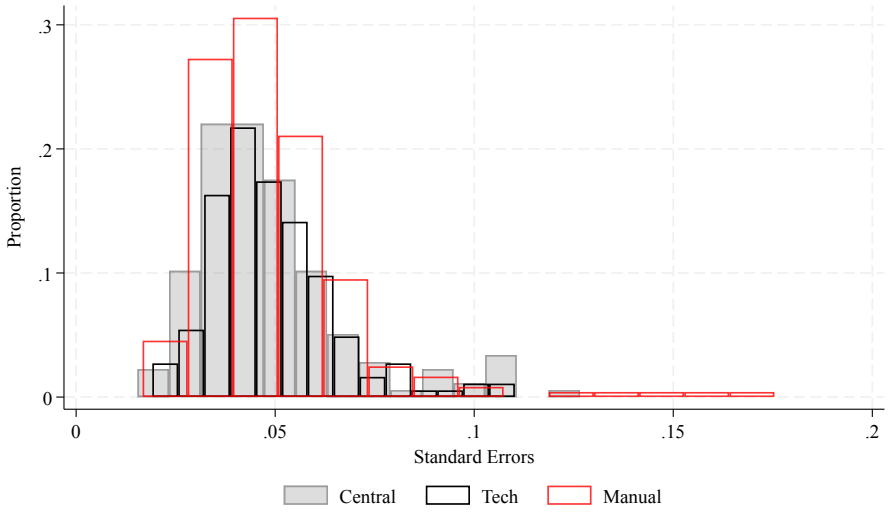
[24] We note that we cannot rule out the possibility that Manual program households were using other non-Utility Apps to respond to peak events. However, as will be shown in Section 6.3 below, we present evidence that Tech program households were not broadly using the Utility's App to respond to peak events by adjusting usage on the installed load controllers. The comparability of the Tech and Manual estimated demand reductions suggests that households in either program were not broadly leveraging App-enabled appliances to respond to peak events.

[25] Appendix C.7 estimates household-level treatment effects by program and event type. We continue to find no evidence of a larger response to the elevated incentives during high peak events.

[26] There are several large standard error estimates for the Manual program. This small subset of households has an average consumption reduction of 8.4% during events, more than double the average household treatment effect for the Manual program of approximately 4%. This imprecision could reflect inconsistent responses across events, or it could also arise from an evolution in responsiveness as the households gain experience with the program.

tional empirical evidence demonstrating the consistency in the responses to events by demand response program throughout our experiment.

Figure 7. Standard Errors of the Household-Level Estimated Treatment Effects by Program



Notes. The reported results plot the standard errors of the estimated household-level treatment effects from specification (2), separately by demand response programs. The histogram presents the proportion of observations that fall within a specific bin's range.

5.3 Identification, Selection, and Robustness Results

In this section, we summarize the results of empirical tests and robustness checks that address each of the key assumptions outlined in Section 4.3, as well as potential selection concerns. We present additional details on each and an array of robustness checks and sensitivity analyses in Appendix C.

Balanced Panel

Attrition in our 17-month experiment is relatively low and gradual with rates in the range of 4.5% - 7% across our programs and never-treated groups (see Table 3 and more detail in Appendix C.2). We find no evidence of statistically different attrition rates across the programs. With few exceptions, attrition arose from households that moved and closed their utility accounts rather than requesting an exit from the program.[27] We interpret this form of attrition as being independent from the

[27]This can be directly observed by the 6.9% attrition rate among the pure Control households who experienced no intervention and we simply received their data.

households' experience with the program. To evaluate the implications of attrition on the balanced panel assumption, in Appendix C.2 we run versions of our analyses with our full sample and the balanced sample that removes households that attrited. Our results are robust across the two samples.

Table 3. Final Attrition Rates by Program

	Central	Tech	Manual	Info	Control
Final Attrition (%)	6.21	4.89	4.55	6.78	6.91

Notes. Final attrition (%) represents the percentage of initially enrolled users that unenrolled by the end of our experiment.

Temporal Stability

We extend our household-level regression by considering a specification that includes fixed effects at the year-month-hour-of-day to further absorb time-varying factors that impact not just the level of consumption, but also the patterns of consumption over the day within a household. Although these specifications do not directly capture discrete unobservable events that could occur within a month, they help absorb large changes that could occur at a point in time and lead to persistent changes in average household-level consumption differentially throughout the day (Appendix C.6). Additionally, household-specific time-varying changes that could affect temporal stability are likely to be idiosyncratic in nature and therefore would not be expected to occur differentially across demand response programs. As a result, such changes are less likely to confound the identification of differences in peak event responses across programs. Finally, we extend our regression specifications to allow for heterogeneous treatment effects over time (Figure 5; Appendix C.1.3). While this does not directly control for threats to temporal stability, it provides insights into how the estimated responses to peak events have changed over time. The main result, that Central households reduce consumption by roughly 5 times that of Tech and Manual households, persists across all periods of the experiment.

Causal Transience

There can be carryover if households systematically adjust their consumption in event hours even on days when there is no peak event after they start to receive them. This could be broadly characterized as a form of habit formation. In Appendix C.1.1, we test for this behavior by evaluating whether households adjusted their consumption

during the event windows on non-event days in the post-treatment period, relative to the never-treated households, and find no evidence of persistent carryover behavior. Another concern could arise if households responded differently to an event based on the number of days between being exposed to an event, or if the prior event was a morning, evening, or high-evening event. In Appendix C.1.2, we find no evidence of differences in peak event responses on these two sequence-based margins.

Finally, households could change the extent to which they respond over the course of the experiment. For example, households could begin by responding strongly to peak events, but their response could decline over time as their willingness to exert effort/attention to events declines, or if there are any technical challenges with the installed equipment. The analysis of responsiveness by phase of the experiment in Section 5.1 (Figure 5) shows a modest reduction in responsiveness between the first and second phases, with effects stabilizing thereafter. Appendix C.1.3 supplements this with household-level dynamic effects and an analysis of event opt-out behavior over time for the Central program. Again, while there is some evidence of a modest change in the estimated treatment effects in the first couple of months, the effects stabilize throughout the remainder of the experiment.[28]

Selection

Our estimates reflect the average treatment effect of peak events among households that accepted a program offer. Thus, a relevant concern is selection into the programs, rather than non-compliance affecting identification (Section 4.4). Appendix C.3 presents several analyses to address concerns over sample selection that can arise because of differential acceptance of demand response program offers.

First, we perform an Intention-to-Treat (ITT) analysis to estimate the effect of program invitation on event-time consumption, including households that did not accept the invitation to participate (Appendix C.3.1).[29] Because households that did not join a program did not subsequently receive the (randomized) peak events, we assign them a distribution of "synthetic" events matching that of participating households. As expected, the ITT analysis attenuates our main estimated average treatment effects downward; it effectively assigns a noisy zero estimated treatment

[28] In regards to the extent to which Central event opt-out behavior varies over time, with the exception of the first week of events, we find limited evidence of systematic changes in opt-out behavior over time. Rather, opt-out behavior is more closely tied to seasonality, likely reflecting the fact that device opt-outs primarily occur via baseboard thermostat overrides.

[29] More specifically, the ITT estimates the overall impact of a program offer on demand response to events and leverages the randomization of households into the initial program offers. It therefore does not suffer from the same selection concerns.

effect to the households that were invited but did not join a specific program offer. While the ITT estimate has a different economic interpretation than our main specification, our key conclusions comparing the relative magnitude of the estimates across programs persist: the Central program peak event response is statistically and economically significantly larger (by at least three to four times) than the Tech and Manual programs, which have similar magnitudes to each other.

Second, we perform a bounds analysis that serves as a lower bound on the gap between the Central program estimate and those from the other programs (Appendix C.3.2). We estimate what the average response to peak events for the Central program would be, assuming their acceptance rate was equal to that in the Tech or Manual programs, but the additional households that would have joined to equalize those rates had zero response to peak events. This puts downward pressure on our estimated average peak event response for our counterfactual Central program, leading to an estimate of -19% (-23%) when adjusted by higher Manual (Tech) acceptance rate. This compares to the point estimate of -26% in Figure 3. Despite the rather extreme assumption that those who did not enroll would show *zero* responsiveness if they otherwise had enrolled, differential selection rates cannot explain away the differences in our main estimates across the programs.

Finally, our program-level analysis compares event-time to non-event-time consumption of households in the same program, other programs, and never-treated households, relying on the assumption that these groups provide a valid counterfactual on non-event days. In Appendix C.3.3, we vary the set of non-event comparators to test the robustness of our results. We start with the most restrictive set: limiting the analysis to be solely within each demand response program and excluding the never-treated households. This restrictive specification ensures we are only comparing behavior across participating households within the same program. Next, we augment the comparison group to include never-treated households, i.e. those in the Control and Info groups, as non-treated comparators. Finally, we pool all participants and never-treated households to run a regression comparing those receiving a peak event to all households not receiving a peak event at the same time. We find our results to be robust to these specifications, suggesting that potential differential selection into programs is not driving our key conclusions.

Statistical Inference

Throughout our analysis, we conduct a large number of statistical tests. This can raise concerns of false positives because of the large number of null hypotheses being

tested. In Appendix C.4, we carry out multiple hypothesis testing corrections to address these concerns using False Discovery Rate (Benjamini et al., 2006) and the more conservative Familywise Error Rate (List et al., 2019, 2023). Our levels of significance and qualitative conclusions remain.

To evaluate the potential role of increased statistical power, in Appendix C.5 we follow List et al. (2024) and include baseline household-level information on pre-treatment consumption patterns and household appliances. Consistent with their results, we find that the inclusion of this information increases precision of our estimates. Although, the estimates and precision lead to the same conclusions as in our main analyses.

6 Mechanisms

Our results have focused so far on estimated treatment effects by demand response programs. Our main results suggest that active versus passive response and the associated differences in effort and attention play a key role in delivering demand flexibility. Participants in the Tech and Manual programs, both requiring active responses to financial incentives, reduced electricity much less on average during events, with similar behavior across the two programs. Our household-level estimates parallel these findings, but find that there is heterogeneity in responsiveness within each program.

Several natural and important economic questions arise from these analyses: What drives these differences across the three programs, and heterogeneity within each program? Is it driven primarily by differences in the proportion of households responding, or by a divergence in the amount responding households adjust their consumption? What behavior underlies the differential responses to events of the Central compared to the Tech programs, given they have the same load-control technologies? What devices are associated with the largest reductions?

In this section, we present results to address each of these questions. We begin by leveraging a principal stratification framework to disentangle differences across programs into two components: differences in the likelihood of responding to peak events at all and, conditional on responding, differences in the magnitude of demand reductions. Second, we leverage App interaction data during events to analyze the relationship between the frequency of engagement with the App and household-specific estimated treatment effects. This provides insights into the role of attention and ef-

fort, and how it relates to the heterogeneity in event responsiveness we observe in our household-level estimated treatment effects. Third, we use device-level usage data to better understand the adjustments behind the observed demand reductions. These device-specific results also provide intuition into the potential private costs (e.g., discomfort) that was induced by these adjustments. Fourth, we present data from an exit survey that included self-reported income and the stated value of event participation to look at the correlation between these measures of a household's opportunity cost of time/willingness to exert effort and its estimated response to peak events.

6.1 Decomposing the Consumption Response

Our headline estimates combine two distinct margins for the household's consumption decision. First, upon receipt of a peak event signal, the household must decide whether to respond at all ("the extensive margin"). Second, conditional on responding, the household must decide by how much to respond ("the intensive margin"). The ultimate consumption response is the product of these two margins. A larger reduction in demand during peak events can arise because a larger fraction of households respond, responding households reduce demand by a larger amount, or both channels simultaneously. This distinction is especially relevant in our setting because the Central program generates a substantially larger average response than either the Tech or Manual programs. The question is whether this gap is driven primarily because automation raises the likelihood of responding at all, or whether it allows responding households to reduce consumption by more. Our study design allows us to disentangle these margins and decompose the gap between Central's large average reductions relative to the two more active-response-requirement programs.

To do so, we leverage the principal stratification framework (Imbens and Angrist, 1994; Frangakis and Rubin, 2002). Principal stratification classifies households by their latent response type, where type is defined by post-treatment behavior under each possible program assignment. The key empirical difficulty is that these types are not directly observed: for any given household, we only observe behavior under the program to which it was randomly assigned and to which it accepted, not the full set of counterfactual response decisions under the other programs. The principal stratification framework is a way of using the experimental design to recover information about unobservable behavior that is policy-relevant. In our setting, the relevant post-treatment behavior is whether a household exhibits a consistent reduction in demand during peak events (i.e., they are a "responder" to peak events).

To conceptualize this in our setting, define $R_i(j) \in \{0,1\}$ to indicate whether household i would be a responder to peak events under program j (for $j \in \{C,T,M\}$). Let $Y_i(j)$ denote the household's treatment effect under program j, measured so that more negative values correspond to larger reductions in demand. The estimated average treatment effect for program j combines two components: the probability that a household responds under that program and the magnitude of response conditional on responding: $E[Y_i(j)] = Pr(R_i(j) = 1) \times E[Y_i(j)|R_i(j) = 1]$.[30]

The principal stratification framework is useful because it gives a structured interpretation for these two margins. We assume that the effort required of household i to respond to a peak event declines as one moves from Manual to Tech and Tech to Central, and therefore impose a monotonicity assumption $R_i(C) \geq R_i(T) \geq R_i(M)$. This assumption rules out response patterns such as households that would respond in Manual but not in Tech, or in Tech but not in Central.

Under this assumption, each household belongs to one of four latent response strata, summarized by the vector $(R_i(C), R_i(T), R_i(M))$. These are: (i) never-responders (0, 0, 0), who would not respond under any program; (ii) always-responders (1, 1, 1), who would respond under all three programs; (iii) Tech-induced responders (1, 1, 0), who would begin responding once moved from Manual to Tech; and (iv) Central-induced responders (1,0,0), who would respond only under the automated Central program.

We leverage this partition to link observed cross-program differences in response shares to latent types. Let π_A, π_T, π_C, and π_N denote the population shares of always-responders, Tech-induced responders, Central-induced responders, and never-responders, respectively. Under the monotonicity assumption, when looking across all participants in our experiment, the observed share of responders in each program has the following interpretation: $Pr(R_i(M) = 1) = \pi_A$, $Pr(R_i(T) = 1) = \pi_A + \pi_T$, and $Pr(R_i(C) = 1) = \pi_A + \pi_T + \pi_C$.

This relationship implies that the observed responder share in Manual can be interpreted as the proportion of always-responders (π_A); the increase in the responder share from Manual to Tech reflects the share of Tech-induced responders (π_T); and the increase from Tech to Central reflects the share of Central-induced responders (π_C). The remaining share of the population consists of never-responders (i.e., $\pi_N = 1 - \pi_A - \pi_T - \pi_C$). Although we do not observe each household's full latent response

[30]For simplicity, this equation assumes non-responder households ($R_i(j) = 0$ for $j \in \{C,T,M\}$) have an expected zero treatment effect and therefore does not include the term $Pr(R_i(j) = 0) \times E[Y_i(j)|R_i(j) = 0]$.

type directly, under random assignment and monotonicity, the principal stratification framework provides us with a framework to infer the population shares of these latent types from cross-program differences in responder shares.

With this framework in mind, the next step is to leverage our empirical setting to classify the proportion of households that are responders in each demand response program, representing the extensive-response margin, and linking this to the principal strata discussed above. Because we do not observe the underlying latent behavior, we do not observe whether a household would have been a responder under a specific program. To overcome this challenge, we implement an empirically-driven approach to define if a household is a consistent responder to peak events by leveraging our household-specific estimated peak event responses from Section 5.2. In particular, we define a responder as a household whose estimated treatment effect during peak events is negative and significant at the 5% level. This definition captures both the direction of the estimated household-level response and the degree of statistical evidence that the response differs from zero. We interpret negative and significant estimates as evidence of a consistent average reduction in demand across peak events.[31]

We then estimate the conditional response to peak events among households categorized as responders. More specifically, we compute the average household-level estimated treatment effect by demand response program among households deemed to be responders. Taken together, this provides an estimated decomposition of the average program-level estimated treatment effect into the underlying extensive- and intensive-response margins.

Table 4 presents the results of our decomposition exercise. First, we begin discussing the extensive margin of likelihood of being a responder. Under the principal stratification framework and the monotonicity assumption above, the 18% responder share in Manual is interpreted as the population share of always-responders to peak events. The Tech share of 20% indicates that there is only a 2 percentage point increase in proportion of households classified to be responders, suggesting that the Tech-induced responder group is limited. The largest jump arises as we move to the Central program, where there is an additional 60 percentage point increase reflecting the share of households classified as central-induced responders. The remaining 20%

[31]This classification rule should be interpreted as an empirical proxy for the latent response indicator $R_i(d)$ that is not directly observed. Our empirical responder classification is a proxy constructed from household-level estimates and is used to approximate the distribution of these latent response types rather than a full identification exercise of households' principal strata.

Table 4. Responder Shares and Average Peak Demand Reduction Among Responders

	Central	Tech	Manual
Share of Responders	0.80	0.20	0.18
Conditional Response (%)	-31.29	-27.11	-22.93

Notes. The reported results present the average household-level estimated treatment effects from estimating specification (2), focusing only on households deemed to be responders. Responders reflect households whose coefficient estimates were negative and significant at the 5% level. We present the treatment effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$. Share of responders represents the percentage of households within a demand response program that are classified as responders.

of households can be interpreted as never-responders.[32]

Second, we determine the magnitude of response, conditional on responding, by calculating the mean estimated household-specific treatment effect among households flagged as responders separately by program. We find these conditional responses to also be larger for the Central program (31%) than Tech (27%) and Manual (23%).[33] The difference between Central and both Tech and Manual is significant, while that between Tech and Manual is not. This pattern is suggestive of a stronger intensive margin response under Central.

Taken together, these results suggest that Central appears to outperform the other programs on both margins: it induces substantially more households to exhibit a consistent reduction in demand during peak events, and the households that do respond under Central reduce demand by more on average. Of the two margins underlying the superior reductions of the Central program, the dominant factor appears to lie on the extensive margin of being a responder. This suggests that the first-order advantage of Central is that it expands the share of households that respond at all, likely facilitated by the lower threshold on the time, effort, and attention cost of responding.

In Appendix C.9, we report results considering alternative significance thresholds to define responders. In addition, we use a different approach that defines responders

[32] These could reflect households that opted out of events, had limited usage of load-controllable appliances during peak event windows, and/or whose responses are too small or too noisy to satisfy the empirical responder definition.

[33] Our conditional response results allow us to compare our estimates more directly with Bernard et al. (2026) who focus on the average change in consumption during peak events among households who first "opt-in" to respond to an event—thereby directly flagging themselves as responders. Their estimate of a 23.1% reduction amongst responders compares favorably with our estimates from the similarly behavioral Tech (27%) and Manual (23%) programs.

based only on the magnitude of their household-level estimated treatment effects (i.e., if their estimates achieve certain demand reduction thresholds). While the precise quantitative results vary, we find that the rankings of the share of responders and the conditional peak event reductions among responders across the three demand response programs persist.

6.2 App Interactions

In this section, we use data on households' interactions with the Utility's App to better understand the potential mechanisms driving our results. We begin by allowing our household-level treatment effects to vary by whether or not the household interacted with the App. We then break down how a household's performance during peak events varies by the intensity and types of App interactions.

The App data tell us when users are interacting with the App on a given day, as well as more details on which features of the App they are accessing. For the Central and Tech households, interacting with the App allows them to control connected devices. For the Central program, the App can be used to opt out of automatic load control before or during events. In all programs, the App allows the household to observe the details of upcoming events 21 hours in advance, detailed information about household consumption in real-time, and performance in previous events.[34]

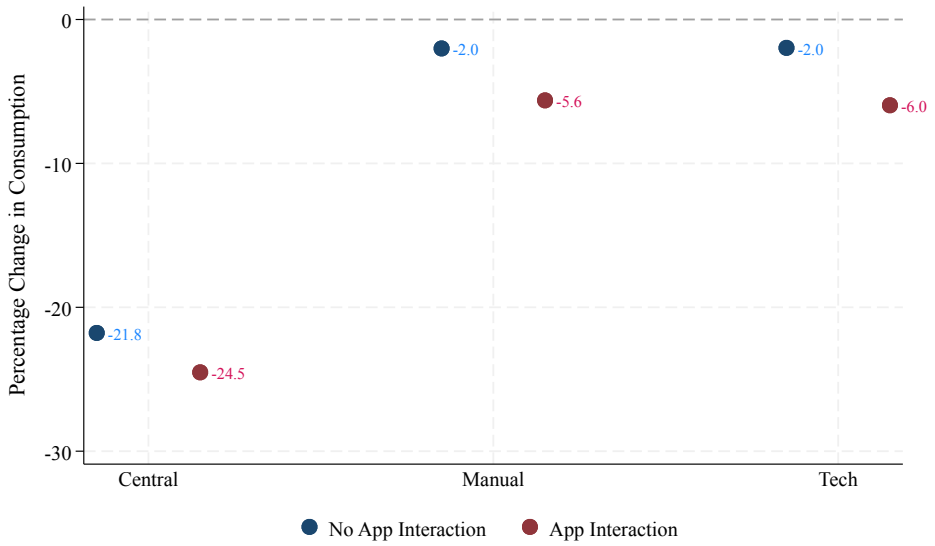
While households could be aware of an upcoming event without opening the App through push notifications sent to their phones, App interactions are a strong indication that the household is aware of the demand response event and serves as a proxy for time spent/effort to respond. Despite the lower average response to events for the Tech and Manual households, Figure 6 demonstrates there is a small subset of high performers in the Tech and Manual programs that have large negative estimated treatment effects. We leverage the App interaction data to evaluate if these high-performers differ in their use of the App during events.

We run a specification of Equation (2) that interacts E_{it} with an indicator variable App Interact$_{it}$ that equals one if the household interacts with the Utility's App on the relevant day and zero otherwise. A key benefit of this approach is that it allows us to quantify how a specific household's estimated treatment effect varies by whether or not they interacted with the App on an event day. This helps overcome the sample selection challenge that would arise by running an analogous regression using

[34] In addition to actively opening the App on their phone, the App opens and logs an interaction if a user clicks on the event push notification sent to their phone.

all households within a given demand response program. With such a regression, it would not be possible to disentangle whether the different treatment effects arise because the household interacted with the App on an event day or whether households that interact with the App are unique in the way they respond to events. However, even in our regression analysis, it is important to acknowledge that App interactions are not random within a household. They can reflect days in which the household is home and able to respond, or days where they are more attentive to usage decisions for unobservable reasons.

Figure 8. Average Estimated Household-Level Treatment Effects by App Interaction



Notes. The reported results are average household-level treatment effects by whether or not the household interacted with the App on an event day. This represents the specification in equation (2), adjusted to interact the event indicator (E_{it}) with an App Indicator$_{it}$ variable that equals 1 when the household interacts with the App on the event day and zero otherwise. All specifications include fixed effects at the year-month, day-of-week, and hourly levels.

Figure 8 reports the average estimated household-specific treatment effects, allowing for heterogenous treatment effects by whether or not the household interacted with the App on a given day. For the Central program, the average estimated household-level treatment effect is -21.8% when the household does not interact with the App, increasing in magnitude to -24.5% when they do interact with the App during an event day. In contrast, for the Manual and Tech programs, the average household-level treatment effect is -2% when they do not interact with the App, increasing in magnitude to roughly -6% when they interact with the App.

These results provide two key insights. First, there is a positive relationship between App interactions and demand reductions. Households who were more attentive to the App on event days achieved higher reductions. This differential effect is slightly larger in magnitude for the Tech and Manual programs.[35] Second, there is a considerable difference in the "no interaction" estimates between Central program participants and those in the other programs (-21.8% vs. -2%). This is consistent with the Central program having a roughly 20 percentage point "headstart" over the other programs, even on days when they appear inattentive to the events and less likely to be taking active action to respond to them. This "headstart" is likely key to the overall finding of greater demand response by the Central program participants, all of whom had a default setting to passively respond to events.

The results reported in Figure 8 summarize average estimates of the distribution of household-level treatment effects. Underlying these effects are varying degrees of significance. For the Central program, approximately 68% of the household-level peak event coefficient estimates are significant both with and without App Interactions. In contrast, we see a more than doubling in the proportion of significant household-level event coefficients, with only 11% (12%) of Tech's (Manual's) estimates being significant when there is No App Interaction on the event day, more than doubling to 23% (25%) on days with App Interactions.[36] These results further emphasize the "headstart" of the Central program, and the added time/effort associated with significant reductions in demand for the Tech and Manual programs.

We also have detailed data on the types of pages/tabs the households interact with in the App (i.e., General Interactions, Energy Usage Dial, Devices Tab, and Advisor Tab). This provides insight into the types of actions households may have

[35] For the Central program, the larger effect suggests that on average, when Central households interacted with the App on event days, they took additional action to reduce consumption beyond just permitting automation (e.g., by delaying use of other non-automated appliances). These results are consistent with the observation of low opt-out rates reported in Section 5.1, indicating that Central households were not systematically interacting with the App to override automation.

[36] For the Tech and Manual programs, while there is an increase in the presence of a significant event response when household interact with the App (compared to the null of no change), the estimated treatment effect is only significantly different on days when the household interacts with the App from days when it does not for 18% and 23% of these households, respectively. These results are suggestive of a subset of households that are actively interacting with the App and reducing event consumption in these programs, a point addressed in Table 5 further below. For the Central program, while we do not see an increase in the proportion of households that reject the null of no peak event response, 24% of households have significantly different effects on days where they do interact with the App compared to non-App interaction days. The moderately larger average peak event response in Figure 8 suggests that a subset of Central households took additional action on these days, possibly by reducing demand of other non-controlled appliances.

taken to respond to events. Table 5 summarizes the average daily App interaction frequency for each demand response program separated by performance quartile across all participants. The performance quartiles are determined by ranking household-level treatment effects over all demand response programs. This separates households into categories of whether or not they are high or low-performing households during our experiment. We summarize the count of the households and the percentage of households enrolled in each demand response program that fall within a specific performance quartile.[37]

Table 5. Average Daily App Interaction Frequency by Program and Performance Quartiles

Program	Performance Quartile	Household Count (%)	General Interactions	Energy Usage Dial	Devices Tab	Advisor Tab
Central	1	111 (63%)	0.32	0.30	0.18	0.26
	2	44 (25%)	0.19	0.18	0.12	0.15
	3	15 (8%)	0.21	0.20	0.15	0.16
	4	7 (4%)	0.15	0.15	0.11	0.11
Tech	1	20 (11%)	0.54	0.49	0.32	0.42
	2	43 (23%)	0.30	0.27	0.18	0.23
	3	53 (29%)	0.23	0.22	0.11	0.17
	4	68 (37%)	0.13	0.12	0.07	0.09
Manual	1	20 (8%)	0.59	0.57	0.07	0.49
	2	64 (26%)	0.18	0.17	0.03	0.15
	3	83 (34%)	0.15	0.15	0.03	0.12
	4	75 (31%)	0.15	0.14	0.02	0.12

Notes. The reported results provide the daily frequency of App Interactions by program and performance quartile. Performance quartiles are determined using the household-specific estimated treatment effects obtained from estimating specification (2). Household Count represents the number of households that fall within each performance quartile. The percentages report the percentage of households within a program that falls within each quartile. General Interactions reflect any interactions with the App. Energy Usage Dial displays the energy dial in the App that provides data on real-time usage. The Devices Tab displays a household's connected devices and allows households to adjust the use of the installed devices. The Advisor Tab reports information on upcoming events and historical performance on past events.

Table 5 demonstrates that the majority of Central households (63%) fall within the top quartile of performance. In contrast, only 11% and 8% of Tech and Manual households are in the top quartile. This is consistent with our results above that

[37]Figure C17 presents the share of participants in each performance quartile by demand response program. This figure illustrates that the vast majority of the top quartile performers are in the Central program (74%), while the lower quartiles consist largely of the Tech and Manual households.

the Central program has a significantly greater demand response than the Tech and Manual programs, which show similar results.

Within each program, we see a reduction in the frequency of daily General Interactions as we move down the performance quartiles. This suggests that higher-performing households are more attentive and/or spend more time to monitor and react to events. In fact, across all columns, the frequency of App interactions in the top quartile is significantly greater (at the 5% level) than the frequency in the second-highest quartile.

We see a considerably higher frequency of App interactions among the Tech and Manual households that achieve the top quartile of performance, interacting with the App on 54% and 59% of days on average, respectively. In contrast, Central households in the top quartile only interact with the App on 32% of the day. The top quartile of the Central program interacts significantly less with the App than the top quartile of the Tech and Manual programs (at the 5% level). This is true for each App interaction type in the table. This suggests that achieving this high threshold of performance required considerable more time and attention for the two active programs as compared to Central—a key barrier for many participants.

Looking across the App categories, households interacted with the Energy Usage Dial the most, followed by the Advisor Tab. This suggests that when households used the App, they often monitored their real-time consumption and the details of upcoming events and/or their past performance in the Advisor Tab. The top quartile performers in the Tech and Manual programs stand out in both of these categories, with the frequency of their interactions with both the Energy Usage Dial and Advisor Tab being roughly 2 to 4 times those of the bottom 3 quartiles.

Tech households in the highest-performing quartile interacted with the Devices Tab in the App at a significantly higher frequency than all other households. This suggests that a (small) subset of households in this program were using the installed devices to achieve larger demand reductions. However, looking at the lower performance quartiles in the Tech and Central programs, households in the Tech program interacted with the Devices Tab about as much as those in the Central program.[38]

Overall, these results suggest that households in the Tech and Manual programs

[38]In addition to the installed devices from our experiment, a subset of households had other devices linked to the App. These were primarily smart plugs linked to lighting in the home. This helps explain why we observe interactions with the Devices Tab for the Manual program. Only 3 Manual households installed and linked devices to the App during our experiment, all three were thermostats for electric baseboard heaters.

had to allocate a higher degree of their time/effort to achieve relatively high demand reductions and the associated rewards. In contrast, the Central program participants had to invest less time, effort, and attention in responding to events than others to generate the same scale of demand reductions.

6.3 Device-Specific Estimates

In this section, we use data on hourly device-level consumption for installed load controllers at households in the Central and Tech programs.[39] This provides insights into the types of adjustments underlying the changes in household-level consumption, and what this suggests for households' private cost associated with these adjustments.

Figure 9 reports results from estimating the specification in Equation (1) separately for the Central and Tech programs, replacing household-level hourly consumption with device-specific electricity usage. For Tech, we see small and insignificant reductions in hot water and heating. The point estimate for EVs is larger, suggesting potentially more willingness to adjust this device, although the estimates are imprecise due to smaller sample size.[40] These results are consistent with the relatively small average peak event reductions using household-level data for the Tech program reported above.

In contrast, for the Central program, we observe large estimated average percentage reductions for all devices, with a near complete reduction in water heater usage during the 3-hour peak events.[41] These results are consistent with households in the Central program permitting broad automated control of these devices during peak events. Appendix Figure C13 presents results for a linear specification and demonstrates that the largest average demand reduction in kWh also arises from water heaters, followed by EVs and thermostats.

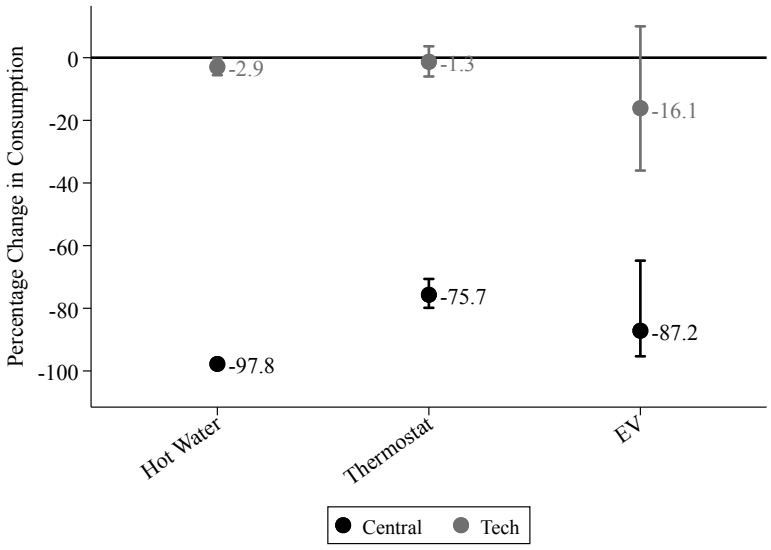
It is important to acknowledge that these estimated reductions in consumption from household devices can come at a private cost (e.g., household discomfort). While we do not observe precise measures of discomfort, Central households have the ability to opt-out of peak events for all connected devices or specific devices using the Utility's App. We observe a low opt-out rate of approximately 4% across all events, and

[39] We do not observe device-level consumption data for the Manual program.

[40] Appendix C.10 reports the number of households with each device type, illustrating a relatively small number of EV load controllers in our experiment.

[41] Focusing on electricity usage of hot water heaters in Central households, the point estimate on the peak event indicator is precisely estimated and implies a narrow percentage reduction 95% confidence interval $[-98.59, -96.44]$. Appendix Figure C10 provides descriptive evidence to support the near-zero consumption from hot water heaters during peak events for Central households.

Figure 9. Device-Specific Estimated Treatment Effects - Central and Tech Programs



Notes. The reported results are *device-specific* marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$, replacing the household-level consumption with device-specific consumption. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

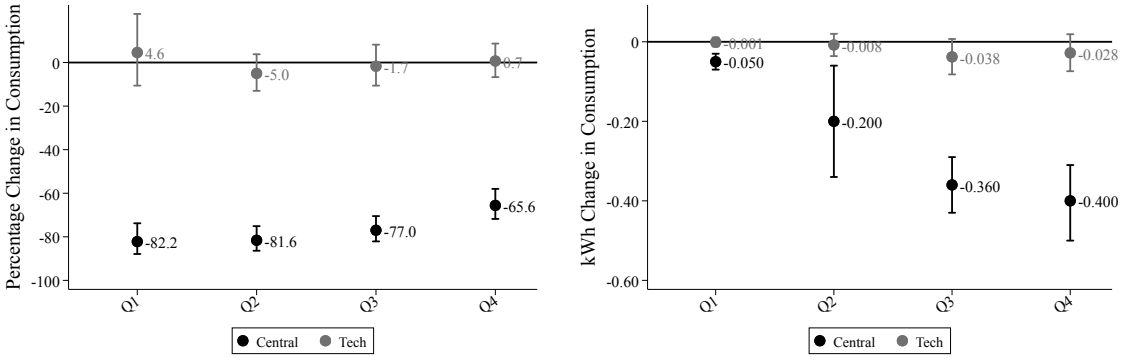
approximately 90% of these arise from adjustments to thermostat settings during events or pre-event opt-outs. The minimal opt-outs on hot water heaters and EVs, and their large estimated treatment effects, are suggestive of limited discomfort, or at least insufficient disutility to lead households to act to override automation.[42,43]

While thermostat opt-outs were modest, and automated thermostat adjustments were limited to a 3-degree Celsius temperature deviation, the relatively higher opt-out rates for thermostats are suggestive of a higher private cost of automation for this device. In our setting, automated adjustments to electric thermostats during peak events were limited to adjustments to heating loads. To further understand this margin, Figure 10 focuses on thermostat electricity usage and presents estimated peak event treatment effects, permitting differential effects by heating degree quartiles separately for a log and linear specification. Tech program households continue to

[42]These results are consistent with findings in the literature that find limited override of automation of electric vehicle charging (Bailey et al., 2025a) and a large engineering literature documenting the thermal storage potential of electric hot water heater tanks for short-duration (1 - 4 hour) peak events with limited loss of heat (Atikol, 2013).

[43]Appendix C.10 presents supplementary device-level results, including evidence that there is no evidence of a significant difference in device-specific responses by event type.

Figure 10. Thermostat Estimated Treatment Effects by HDD Quartile - Central and Tech Programs



(a) Log Specification

(b) Linear Specification

Notes: The reported results are baseboard thermostat-specific marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$, replacing the household-level consumption with baseboard thermostat-specific consumption. We allow the event indicator to have differential effects by Heating Degree quartiles. We consider a log and linear specification for the dependent variable. For the log-specification, we present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. The linear specification reports $\widehat{\beta}_j$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

exhibit limited evidence of an adjustment to thermostat usage during peak events.

Figure 10a illustrates that there is a large response among Central program households that is declining in percentage change terms as we move to higher heating degree quartiles. However, Figure 10b shows that underlying these changes is a strictly increasing reduction in kWh of thermostat usage, with minimal kWh reductions at the lowest quartile (Q1) reflecting days with limited heating needs. These results suggest that even on the coldest peak event hours, the average estimated treatment effect among Central household's thermostat usage is large. The declining percentage reduction in demand is driven by two possible effects: (i) the 3-degree Celsius temperature deviation limit is reached and the device turns on periodically to stay within this limit and/or (ii) there is a higher rate of opt-outs on thermostats as heating degrees increase. We have data that can directly inform the extent of the latter. The opt-out rate increases only modestly in absolute value as we move from the first quartile of heating degrees (3%) to the fourth quartile (6%), suggesting that even on the coldest days, the majority of households do not find automated control of thermostat settings sufficiently uncomfortable to override.

6.4 End-of-Sample Survey

In this section, we use data collected from a survey at the end of our experiment (in June 2023) to further understand the potential mechanisms driving our results. Households who participated in the Central, Tech, and Manual programs were sent a survey and asked questions to better understand their opportunity cost of time/effort and stated net benefit to respond to events. We correlate households' responses to these survey questions with their estimated household-level treatment effects. These results should be interpreted with the understanding that they are based on stated-preference questions among those who responded, and households' responses could have been informed based on their desire to justify their behavior during the peak events. Nevertheless, we report these results as additional suggestive evidence of the mechanisms at play in our analysis.

There were two key survey questions to assess households' willingness to exert effort/opportunity costs of time.[44] First, we asked participants their annual household income level and gave them several income categories (Less than \$50k, \$50-99k, \$100-149k, \$150-200k, and over \$200k). In the context of Becker's (1965) household production model, a household has a finite amount of time to allocate to tasks. The time spent allocating to a specific task, which in our setting includes the time/effort to respond to peak events, is weighted with the opportunity cost of undertaking other activities, such as working and earning income (e.g., the wage rate in the model). In this framing, a household's income serves as a key component of its opportunity cost of allocating effort to reduce consumption during peak events.

Second, we asked a stated-preference question to assess the extent to which participants felt that participating in events was worth their time: "For the events you noticed, how often was it worth your time to participate by attempting to reduce your electricity consumption?" Respondents choose one of the following: (1) Never, (2) Sometimes, (3) About half the time, (4) Most of the time, or (5) Always. This question encompasses an arguably broader stated-preference assessment of a household's perceived net benefit of allocating effort, time, and attention to responding to peak events.

We observe a relatively high response rate to the survey, with 75%, 71%, and 69% of households in the Central, Tech, and Manual programs completing the survey. Respondents were paid \$20 upon completion of the survey. The results presented

[44] Appendix D.1 provides a more detailed summary of the exit survey.

below focus on the households that completed the survey.[45]

For each household i that responded to the exit survey, we use the survey responses and household-level estimated treatment effects to estimate the following regression:

$$Y_i = \beta_0 + \beta_1 I_i + \beta_2 Z_i + \beta_3 G_i + \gamma X_i + \epsilon_i \tag{3}$$

in which Y_i is the estimated treatment effect for household i (using estimates from Section 5.2), I_i is a categorical variable for the household's reported income, and Z_i is a categorical variable for the household's response to the Worth Time question described above. The excluded value for the reported annual income is the category $<$ \$ 50k, and for the Worth Time question the excluded categories are anyone who reported "Never".

We include several variables to control for a household's appliances and household type, denoted by X_i. These controls include an indicator variable if the household has an electric hot water heater, a categorical variable for whether a household has an electric vehicle and their corresponding charger type (No, Level 1, Level 2), a categorical variable for air conditioning in the home (No, Window Unit, Central Air), a categorical variable for electric baseboard heating (No, 1 – 3 Units, 4 or more units), and a house/duplex dummy variable (1 for house/duplex, 0 for row home).[46] G_i denotes indicator variables for the Central and Tech programs (with Manual as the baseline) to allow for differential average treatment effect levels by program. Our dependent variable is a generated regression. We report bootstrapped standard errors to address concerns that the standard errors are biased in this setting.

The regression results are reported in Table 6. Focusing first on household income, we see that all coefficients are positive, indicating reporting annual income in excess of 50k (the excluded category) is associated with larger (positive) treatment effects, representing smaller electricity consumption reductions during events. There is some evidence that these positive relationships are increasing as reported income increases. However, the estimates are noisy and imprecisely estimated.

[45] Appendix D.2 compares the characteristics of households that responded to the survey with those that did not using pre-treatment data. While the households are similar on a number of characteristics, responders consumed less electricity on average. To the extent that consumption is correlated with factors that relate to the opportunity cost of time, this could downward weight households with a higher opportunity cost of time in the end of experiment survey.

[46] In Appendix D.3, we include additional control variables using household-level pre-treatment information and a LASSO approach to select the control variables to include in the regression. The key conclusions drawn from this analysis are consistent with the baseline OLS regression, with a modest increase in precision.

Table 6. Household Treatment Effect on Worth Time and Income

	Coefficient	Std. Error	P-Value
Worth Time			
Sometimes	-3.4934	2.2404	0.1189
Half of the Time	-6.3097	2.4711	0.0107
Most of the Time	-9.0212	2.4337	0.0002
Always	-15.0505	2.9170	0.0000
Income			
50 - 99k	5.5255	4.5976	0.2294
100 - 149k	7.3086	4.6087	0.1128
150 - 200k	9.8930	4.5909	0.0312
>200 k	9.7378	4.5432	0.0321
Program Indicators			
Central	-17.2308	1.7559	0.0000
Tech	-1.3486	1.6000	0.3993

Notes. The reported results present the estimates for Worth Time, Income, and Program variables in Equation (3). The regression also includes an array of appliance and household type control variables. Standard errors are estimated via bootstrap with 5,000 iterations.

For a given level of income, demand response program, and stock of household appliances, households may have heterogeneous preferences in their willingness to allocate time to provide demand flexibility. Table 6 finds that all of the Worth Time variable categories are negative and significantly different than the excluded "Never" category, with the exception of the "Sometimes" response category. Additionally, the coefficients are more negative as we move down the spectrum of the Worth Time measure. These results suggest that the participants' perceived net benefits of allocating time to participate in events are correlated with greater reductions during events. As expected, we see a large negative and significant coefficient on the Central indicator variable and the coefficient on the Tech program is close to zero and not significantly different compared to the baseline Manual program.[47]

While these results focus on a stated-preference survey and they represent correlational relationships, they provide additional evidence that is consistent with the broader findings presented through the paper to explain why we observe large dif-

[47]In Appendix D.3, we present results for a multiple hypothesis testing robustness check for this regression analysis. While we find reduced precision and significance for the estimates for the Income coefficients, the significance thresholds for the Worth Time categories remain.

ferences in responses to peak events across our demand response programs. Factors that lead to a higher costs of allocating effort and attention are associated with lower estimated household-level treatment effects. Our findings suggest that the Central program, through the use of enabling technologies and utility-controlled/automated default responses to events, helps overcome these time allocation and effort barriers.

7 External Validity

We discuss the external validity of our study's findings using the selection, attrition, naturalness, and scaling (SANS) framework developed by List (2020).

Selection

Our study draws participants from a pool of the Utility's users that (a) are on the Utility's app and (b) meet eligibility criteria that include being located near the major metropolitan region of the province and having load controllable devices in the home. We acknowledge this results in a selected subset of potential electricity consumers, albeit ones that are likely candidates to join utility demand response programs. To assess potential concerns related to selection, and how our results might generalize beyond this study, we do a comparative analysis of observables as well as discuss potential implications of differential unobservables.

We begin by presenting empirical evidence on observables that leverages the fact that we observe the census dissemination areas (CDAs) of each household that was invited to join a demand response program (plus the never-treated households). A CDA is the smallest census block in Canada and has a population of approximately 400 to 700 individuals. To provide a sense of scale for this number, a commonly used source of data in the United States is to leverage socioeconomic and demographic information at the census tract-level. The U.S. Census Bureau defines the "optimal" census tract size to contain approximately 4,000 people, but tracts can range from 1,200 to 8,000 individuals.

In Appendix A.4, we perform a comparison of an array of socioeconomic and household demographics using the Canadian CDA data, sliced at the broader metro area, the invited sample, and finally the subset of households that joined a demand response program or a never-treated group. For the broader metro area, we summarize the population-weighted census data characteristics across all CDAs where households could have joined (absent our selection criteria), the CDA characteristics among the 1,661 households that were invited to join a demand response program, and finally the

968 households that eventually joined a demand response program (or the control).

We find that households in the Invited and Accepted samples are comparable on observables using the census data (Table A3). This coincides with our balance tables that show that the invited and accepted samples are comparable on electricity-usage and an array of household-level appliance/devices. Comparing these groups to the broader metro area, however, we find that invited and accepted households tend to come from CDAs that have higher average median income, are more likely to have single-family and owner-occupied housing, have a higher proportion of Canadian citizens, and fewer households that self-categorize themselves as immigrants and visible minorities. These observable characteristics suggest that households that made it into our program (or were invited) are from more affluent communities as compared to the broader metro population.[48]

Turning to unobservables, we expect that based on our selection criteria, households that were invited into our field experiment, and ultimately joined a demand response program, are likely to be more tech savvy and have certain observable and unobservable characteristics that differ from the broader population. These households may be more willing and able to put forth effort and attention to respond to a peak event. To the extent that a "typical/average" household is willing to allocate less effort/attention, this would reduce their response to peak events in the Manual and Tech programs. In contrast, for any given configuration of controllable devices and conditional on enrolling, our results suggest that even these typical/average households are likely to maintain high responses to peak events in the Central program because of its set-it-and-forget-it nature. This suggests that the large wedge we observe between the Central and Tech/Manual programs' responsiveness is potentially a *conservative* estimate relative to a general population, for any given set of controllable household devices.

Finally, a broader question exists related to take-up rates, which one could plausibly expect to be lower among a general population. We address this concern in the section on *scaling* below.

Attrition

We present detailed data on attrition rates in Appendix C.2. We find that overall

[48]Figure A1 presents the full distribution of these variables across all Metro CDAs and CDAs that include invited/accepted households. These results demonstrate that the invited/accepted households live in CDAs that lie across the full spectrum of values in the broader metro CDAs, but are more heavily weighted in CDAs that have higher median incomes, more Canadian citizens, and fewer immigrants/visible minorities.

attrition rates are relatively low, not different across groups, and arises primarily from households moving rather than requesting an exit from the program.

Naturalness

Peak events in our study are randomized at the household level; they are not coincident with actual peak load situations for the Utility. This presents both positives and negatives when it comes to naturalness of our experimental setting. On the one hand, the lack of coincidence with peak conditions means that other channels (e.g. news stories, word-of-mouth) that typically increase salience and thus responsiveness of events will be absent in our setting (Brewer and Crozier, 2025). On the other hand, requesting reductions outside a peak load period may lend itself to greater responsiveness due to more flexibility available from the household.

While we cannot rule out these channels, we examine how responsiveness to peak events varies with system demand conditions. In Appendix C.11, we allow household responses to peak events to vary by two measures of system conditions: quartiles of daily system-wide peak demand over the study period and similarly quartiles of heating degree days over the same period. We find no evidence of differential responsiveness across demand and temperature conditions, but for a slightly weaker response for the Central program during the lowest heating degree days.

We note that while our experimental design introduces peak events that do not necessarily correspond to the Utility's overall system-wide demand conditions, such a setting is reasonable for future demand flexibility programs. With the increase in variability from renewables on the supply side of electricity systems, demand flexibility is no longer simply needed during peak demand conditions; being able to respond to sudden changes in renewables will be increasingly valuable. Furthermore, local conditions arising from electrification that lead to distribution bottlenecks will also require greater flexibility during periods that are not coincident with overall system peak conditions. As such, the random events in our study are likely to be reflective of future demand response periods.

Finally, our experiment is set in a northern climate with a system peak in the winter months and load controllers installed on hot water heaters, baseboard heaters, and level 2 EV chargers. In Appendix A.4, we present evidence that illustrates that while our sample is relatively highly electrified, these appliances are used by a sizable proportion of households in the Utility's metro area, and across Canada more broadly. In other settings, air conditioning loads can represent a large proportion of electricity usage and the dominant flexible appliance, particularly in summer months. Prior

studies have studied the role of automation with smart thermostats and summer cooling and find that there is considerable scope for demand response with minimal opt-out rates (Harding and Lamarche, 2016; Bollinger and Hartmann, 2020; Blonz et al., 2025). While the findings in these studies suggest that our conclusions will persist, how the three demand response programs (Central, Tech, and Manual) perform in this setting is an important area for future research.

Scaling

Participants in our study are drawn from a subset of electricity customers that were already willing to download the Utility's app. While take-up among eligible households in this subset was reasonably high, it is reasonable to expect take-up amongst the broader population to be lower. Two key questions arise related to scaling from our pilot setting. The first is how the effectiveness of automated versus behavioral demand response programs will change when estimated over a broader population. As noted above when discussion selection, we expect our key findings of the benefits of automation to continue to hold, and perhaps widen relative to Tech and Manual programs, across a broader, potentially less-attentive population where the benefits of Central's automation are more pronounced. A potential countervailing force, however, is that the composition of household's potential devices to control is likely to differ. Compared to the broader population, our participants tended to be in neighborhoods with higher income, more likely to be owner-occupied, and a higher share of single detached homes (Table A3). Smaller, renter-occupied, and lower income may have fewer controllable devices from which to benefit from automation's performance.

The second question relates more specifically to the willingness of a broader population to enroll in similar demand response programs. We note the potential for scale exists by pointing to two real-world examples. Specific to our setting, and subsequent to our experiment, the Utility partner in our study deployed a widespread demand response program to its customers based in large part on this pilot. In the three years since its launch, voluntary enrollment has reached roughly 150,000 customers, or nearly 8% of the province-wide residential customer base. Further afield, we point to the success in the United Kingdom, where demand flexibility has become more commonplace, as documented in a recent large-scale experiment by Bernard et al. (2026). They recruit from a broad population of 2.7 million Octopus Energy customers into a national demand response program and end up enrolling over 50% into their treatment group, while 30% of the control group also signed up, despite no direct invitations, through general visibility to the program on social media, in-app

banners, and word-of-mouth. These examples illustrate the potential to scale demand response programs beyond the take-up rates observed in our framed field experiment.

8 Conclusion

The flexibility of electricity demand is becoming more valuable as electricity supply evolves to include a growing share of variable renewable sources, and emerging technologies such as smart thermostats and electric vehicles promise greater opportunities to provide it.

We run a large-scale field experiment to investigate the role of effort and technology as determinants of electricity demand response. On the one hand, utility-managed ("passive") demand response can take the burden of responding to price signals off consumers while still allowing them, and the grid, to capture the value of shifting consumption as system conditions change. On the other hand, "active" programs require more engagement from consumers, but in doing so hold the potential to unlock responsiveness rooted in the heterogeneous (and largely unobservable) preferences of individual consumers. Which approach delivers more demand flexibility is ultimately an empirical question.

We find that households in the automated Central program reduced consumption by 26% on average during critical "peak events", roughly five times the reductions observed in the Tech program (5%), despite the latter having the same enabling technology and information provision. Enrollment rates for the Central and Tech programs were also not appreciably different, suggesting automation was reasonably tolerated. A second important finding is that the Tech program performed no better than the Manual program (also 5%), despite Tech households being provided smart technology to remotely control their baseboard thermostats, water heaters, and EV chargers. This suggests that technology provision alone is not enough to unlock demand flexibility; what mattered was whether households had to take an active action to respond.

Leveraging the richness of our within-subject design, we explore the mechanisms underlying these results. First, decomposing the consumption response into extensive ("should I respond?") and intensive ("if so, by how much?") margins, we see the Central program's primary advantage comes from the former. Roughly 80% of Central households were consistent responders, versus about 20% in the Tech and Manual programs. The benefits on the intensive margin were less pronounced, but there

too the Central program outperformed both active-requirement programs. Overall, automation's largest effect is to facilitate a far greater likelihood of responding.

Second, data on households' engagement with the Utility's App point to effort as a key correlate of performance: those who interacted with the App more reduced consumption more, and the high performers in the Tech and Manual programs reached those reductions only through frequent engagement, interacting on more than half of event days on average. Central households, by contrast, achieved large reductions whether or not they engaged with the App. Consistent with this, raising the financial reward during periodic "high peak events" did not elicit greater reductions—at least over the range of prices we tested—reinforcing that the binding constraint is effort and attention, not necessarily the size of the incentive.

Our findings come from a framed field experiment at a single utility in a northern, winter-peaking climate, with a specific set of controllable devices—electric baseboard heaters, hot water heaters, and level 2 EV chargers—and a participant pool that is relatively affluent, highly electrified, and open to new technology. Two dimensions of generalization stand out for future research as such programs scale. The first is the breadth of participants: whether the automation advantage we document persists, or—as our selection discussion suggests—widens, as programs expand to broader and potentially less attentive populations for whom the effort barrier may loom even larger. The second is the breadth of devices. The comfort and control trade-offs that govern a household's tolerance for automation are likely device-specific, so the near-complete and well-tolerated reductions we observe for devices in our study need not carry over to loads with more salient comfort implications. This is especially relevant for the flexible loads poised to grow on an increasingly electrified grid, such as air conditioning and heat pumps, which we do not study here. How automated demand response performs across this widening set of devices, and whether opt-out behavior and effort barriers differ across them, is a critical question as demand-side flexibility becomes central to grid reliability.

More broadly, policies that relieve consumers of the cognitive and time burdens that contribute to rational inattention hold real promise for demand-side flexibility. Given our results, utility-initiated automation is one such approach—one that can both save consumers money and deliver the flexible demand that electricity markets increasingly need.

References

Abadie, Alberto and Sebastien Gay, "The impact of presumed consent legislation on cadaveric organ donation: a cross-country study," *Journal of health economics*, 2006, *25* (4), 599–620.

Anderson, Michael L, "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects," *Journal of the American statistical Association*, 2008, *103* (484), 1481–1495.

Atikol, Uğur, "A Simple Peak Shifting DSM (Demand-Side Management) Strategy for Residential Water Heaters," *Energy*, December 2013, *62*, 435–440.

Bailey, Megan, David Brown, Erica Myers, Blake Shaffer, and Frank Wolak, " Electric Vehicles and the Energy Transition: Unintended Consequences of Time-of-Use Pricing ," *American Economic Review: Insights*, 2025.

Bailey, Megan R, David P Brown, Blake C Shaffer, and Frank A Wolak, "Show Me the Money! Incentives and Nudges to Shift Electric Vehicle Charge Timing," *American Economic Journal: Economic Policy*, 2025.

Becker, Gary S, "A Theory of the Allocation of Time," *The Economic Journal*, 1965, *75* (299), 493–517.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen, "Inference on Treatment Effects after Selection among High-Dimensional Controls," *The Review of Economic Studies*, 2014, *81* (2), 608–650.

Benjamini, Yoav, Abba M Krieger, and Daniel Yekutieli, "Adaptive linear step-up procedures that control the false discovery rate," *biometrika*, 2006, pp. 491–507.

Bernard, Louise, Robert D Metcalfe, and Andrew Schein, "Critical Peak Rewards: Evidence from a Nationwide Demand Response Program," Technical Report, National Bureau of Economic Research 2026.

Bernheim, B Douglas, Andrey Fradkin, and Igor Popov, "The welfare economics of default options in 401 (k) plans," *American Economic Review*, 2015, *105* (9), 2798–2837.

Blonz, Joshua, Karen Palmer, Casey J Wichman, and Derek C Wietelman, "Smart thermostats, automation, and time-varying prices," *American Economic Journal: Applied Economics*, 2025, *17* (1), 90–125.

Boiteux, Marcel, "Peak-Load Pricing," *The Journal of Business*, 1960, *33* (2), 157–179.

Bojinov, Iavor, Ashesh Rambachan, and Neil Shephard, "Panel Experiments and Dynamic Causal Effects: A Finite Population Perspective," *Quantitative Economics*, 2021, *12* (4), 1171–1196.

Bollinger, Bryan K and Wesley R Hartmann, "Information vs. Automation and Implications for Dynamic Pricing," *Management Science*, 2020, *66* (1), 290–314.

Brandon, Alec, Christopher M Clapp, John A List, Robert D Metcalfe, and Michael Price, "The Human Perils of Scaling Smart Technologies: Evidence from Field Experiments," Technical Report, National Bureau of Economic Research 2022.

Brewer, Dylan and Jim Crozier, "Who heeds the call to conserve in an energy emergency? Evidence from smart thermostat data," *Journal of the Association of Environmental and Resource Economists*, 2025, *12* (6), 1747–1789.

Brot-Goldberg, Zarek, Timothy Layton, Boris Vabson, and Adelina Yanyue Wang, "The behavioral foundations of default effects: theory and evidence from Medicare Part D," *American Economic Review*, 2023, *113* (10), 2718–2758.

Brown, David P and Derek EH Olmstead, "Measuring Market Power and the Efficiency of Alberta's Restructured Electricity Market: An Energy-Only Market Design," *Canadian Journal of Economics*, 2017, *50* (3), 838–870.

Burkhardt, Jesse, Kenneth T Gillingham, and Praveen K Kopalle, "Field experimental evidence on the effect of pricing on residential electricity conservation," *Management Science*, 2023, *69* (12), 7784–7798.

Charness, Gary, Uri Gneezy, and Michael A Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of economic behavior & organization*, 2012, *81* (1), 1–8.

Chetty, Raj, John N Friedman, Søren Leth-Petersen, Torben Heien Nielsen, and Tore Olsen, "Active vs. passive decisions and crowd-out in retirement savings accounts: Evidence from Denmark," *The Quarterly Journal of Economics*, 2014, *129* (3), 1141–1219.

DTE Energy, "DTE Energy Smart Charge Program," 2022.

Einav, Liran, Ben Klopack, and Neale Mahoney, "Selling subscriptions," *American Economic Review*, 2025, *115* (5), 1650–1671.

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

Frangakis, Constantine E. and Donald B. Rubin, "Principal Stratification in Causal Inference," *Biometrics*, 2002, *58* (1), 21–29.

Garnache, Cloé, Øystein Hernæs, and Anders Gravir Imenes, "Demand-side management in fully electrified homes," *Journal of the Association of Environmental and Resource Economists*, 2025, *12* (2), 257–283.

Harding, Matthew and Carlos Lamarche, "Empowering Consumers Through Data and Smart Technology: Experimental Evidence on the Consequences of Time-of-Use Electricity Pricing Policies," *Journal of Policy Analysis and Management*, 2016, *35* (4), 906–931.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

IRC, "Market Design Executive Summary: 2017 IRC Markets Committee," Technical Report, ISO/RTO Council 2017.

Joskow, Paul L and Catherine D Wolfram, "Dynamic Pricing of Electricity," *American Economic Review*, 2012, *102* (3), 381–385.

Kahn, Alfred E., *The Economics of Regulation: Principles and Institutions: Volume 1: Economic Principles*, John Wiley, 1970.

List, John A, "Non est disputandum de generalizability? A glimpse into the external validity trial," Technical Report, National Bureau of Economic Research 2020.

_ , "The Experimentalist Looks Within: Toward an Understanding of Within-Subject Experimental Designs," Technical Report, National Bureau of Economic Research 2025.

_ **, Azeem M Shaikh, and Atom Vayalinkal**, "Multiple testing with covariate adjustment in experimental economics," *Journal of Applied Econometrics*, 2023, *38* (6), 920–939.

_ **,** _ **, and Yang Xu**, "Multiple hypothesis testing in experimental economics," *Experimental Economics*, 2019, *22* (4), 773–793.

_ **, Ian Muir, and Gregory Sun**, "Using machine learning for efficient flexible regression adjustment in economic experiments," *Econometric reviews*, 2024, *44* (1), 2–40.

Metcalfe, Robert D, Andrew Schein, Cohen R Simpson, and Yixin Sun, "AI in charge: Large-scale experimental evidence on electric vehicle charging demand," Technical Report, National Bureau of Economic Research 2026.

Natural Resources Canada, "Survey of Household Energy Use (SHEU), 2019: Data Tables," 2019. Accessed: 2026-06-18.

Nauze, Andrea La, Lana Friesen, Kai Li Lim, Flavio Menezes, Lionel Page, Thara Philip, and Jake Whitehead, "Can Electric Vehicles Aid the Renewable Transition? Evidence from a Field Experiment Incentivising Midday Charging," Technical Report, CESifo Working Paper 2024.

PG&E, "SmartAC Switches," 2023. https://www.pge.com/en_US/residential/save-energy-money/savings-solutions-and-rebates/smart-ac/smartac-switch.page. Accessed: 2-27-2023.

Prest, Brian C, "Peaking interest: How awareness drives the effectiveness of time-of-use electricity pricing," *Journal of the Association of Environmental and Resource Economists*, 2020, *7* (1), 103–143.

Romano, Joseph P and Michael Wolf, "Efficient computation of adjusted p-values for resampling-based stepdown multiple testing," *Statistics & Probability Letters*, 2016, *113*, 38–40.

Sallee, James M, "Rational Inattention and Energy Efficiency," *The Journal of Law and Economics*, 2014, *57* (3), 781–820.

Schneider, Ian and Cass R Sunstein, "Behavioral Considerations for Effective Time-Varying Electricity Prices," *Behavioural Public Policy*, 2017, *1* (2), 219–251.

Shepard, Mark and Myles Wagner, "Do ordeals work for selection markets? Evidence from health insurance auto-enrollment," *American Economic Review*, 2025, *115* (3), 772–822.

Smith, Jordan, "Texas Regulators Lower Electric Price Cap to $5,000 per Megawatt Hour," 2022. https://www.chooseenergy.com/news/article/texas-regulators-lower-electric-price-cap/.

Spector, Julia, "Portland General Electric Jumps Into Virtual Power Plant Business," 2020. https://www.greentechmedia.com/articles/read/portland-general-electric-launches-525-home-battery-fleet-virtual-power-plant.

Statistics Canada, "Census of Population. Data Products, 2016 Census," 2021. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/index-eng.cfm, Accessed: 2021-07-15.

_ , "Dictionary, Census of Population, 2021: Census subdivision (CSD)," https://www12.statcan.gc.ca/census-recensement/2021/ref/dict/az/definition-eng.cfm?ID=geo021 2021. Accessed June 17, 2026.

_ , "Table 38-10-0286-01: Primary Heating Systems and Type of Energy," 2022. Accessed: 2026-06-18.

Turk, Graham, Tim Schittekatte, Pablo Duenas-Martinez, Paul L Joskow, and Richard Schmalensee, "Designing Distribution Network Tariffs in the US with an Application to Increased Electric Vehicle Adoption," *The Energy Journal*, 2025, *46* (6), 143–165.

U.S. Census Bureau, "2020 Census Detailed Demographic and Housing Characteristics File A (Detailed DHC-A): Technical Documentation," Technical Report, U.S. Department of Commerce 2023.

Wattersaver, "WatterSaver Program," 2023. https://www.watter-saver.com/. Accessed: 2-27-2023.

Wolak, Frank A, "Residential customer response to real-time pricing: The anaheim critical peak pricing experiment," 2007. https://fawolak.org/pages/papers.html.

Appendix Contents

A Supplementary Experimental Framework Material **59**

A.1 Recruitment and Assignment 59

A.2 Comparison of Household Characteristics Upon Randomization . . . 61

A.3 Comparison of Household Characteristics After Acceptance 61

A.4 Comparison of Experimental Households to Broader Population . . . 64

B Treatment Details **68**

B.1 Program-Specific Event Notifications 68

B.2 Treatment Program-Specific App Functionality 70

C Extensions and Robustness **78**

C.1 Causal Transience Analysis 78

C.2 Attrition 87

C.3 Sample Selection Analysis 91

C.4 Multiple Hypothesis Testing 100

C.5 Variance Reduction 103

C.6 Fixed Effects Robustness 106

C.7 Household Treatment Effects by Event Type 110

C.8 Consistency of Demand Response 112

C.9 Principal Stratification Robustness 115

C.10 Additional Device-Specific Results 118

C.11 Heterogeneous Effects by Temperature and Daily System Load 125

C.12 Performance Quartiles by Demand Response Program 128

C.13 Pre- and Post-Event Spillover Effects 129

D End of Experiment Survey **132**

D.1 Survey Details 132

D.2 Survey Response 133

D.3 Survey Regression Robustness 135

A Supplementary Experimental Framework Material

A.1 Recruitment and Assignment

The study sample was drawn from the population of residential customers in the Utility's service territory in and near a large metropolitan city in Canada. We employed a two-step recruitment strategy. In Phase 1, starting in August 2021, the Utility invited households to join an App operated by a third-party company in partnership with the Utility. The App provides households with household-level hourly consumption posted at a one-day lag. The App can be coupled with other devices to provide more detailed information on real-time usage and device control. Households were recruited to the App using several marketing strategies, including advertisements on the Utility's website, social media posts, the Utility's newsletter, website notifications when users logged into their Utility accounts, and emails sent to approximately 306,000 residential households.

The recruitment onto the Utility's App provided us with a pool of 9,020 households to draw from. When households signed up to join the App, they were required to answer a six-question survey. The survey asked households about their motivation for joining the App and whether the household rents or owns their home. It asked about devices eligible for load control in our experiment, including whether the household has an electric hot water tank, an electric vehicle (EV), and electric baseboard heaters as the primary heat source. Households with EVs were asked what type of charger (level 1 or 2) they use. It also asked whether households have air conditioning, a major source of demand flexibility.

We applied several selection criteria to this pool of households. Customers had to be in or near a large metropolitan city in the province for which it was feasible for Utility-partnered electricians to install load control equipment, as needed. Only homeowners were permitted to participate. Condos and apartments were removed, leaving primarily single-family homes, duplexes, and row homes as eligible. Households must have at least one month of historical consumption data as of September 2021, and the customers must have at least one controllable electric device. Recall, the set of controllable electric devices includes a level 2 electric vehicle charger, electric baseboard heaters used as the primary heat source, and an electric hot water heater tank. This left us with a sample of 1,661 potential households that we used for our randomized assignment to experimental programs.

In Phase 2 of recruitment, we randomized the eligible households into our treat-

ment programs and never-treated groups.[49] Starting in October 2021, we sent program-specific emails to households inviting them to join a new "Trial" program. These emails provided details about the specific experiences households would face in the program to which they were being invited, including a summary of the expected rewards they could earn over the course of the Trial, equipment they would receive and its estimated value, and future peak events. Households were also randomly offered a small sign-on incentive of the amounts $10 or $20, or no incentive. All households faced a yes/no decision regarding accepting our program-specific offer. The never-treated Control group that received no equipment, price incentives, or real-time usage information (recall Table 1) received no further communication beyond joining the App in the first phase of recruitment.

Households had to accept the invitation to join the relevant experimental program actively. After selecting to join, households were mailed a device called the "Hub" that facilitates monitoring real-time energy usage via the App. Installers contacted households in the Central and Tech programs to install the load controller equipment.

This two-phase recruitment process occurred over the months of August 2021 - February 2022. The second phase of recruitment occurred in five waves starting in October 2021. As additional households joined the App, we collected the survey responses, identified eligible households, randomized households into programs, and sent the second-phase recruitment emails. This process was used to facilitate the time required to schedule and install the load controllers, as well as to achieve the targeted sample size.

Finally, during the invitations to join each program, we randomized the upfront incentive. While we find a higher rate of initial acceptance with higher upfront incentive payments, the differences are small and not significantly different.[50]

[49] Specifically, we used a randomization procedure designed to balance important observable characteristics over programs and groups. We first used the machine learning algorithm "kmeans" to group households based on observable characteristics. These included cumulative household electricity consumption (in kWh) and load factor by season (Fall, Spring, Winter, and Summer), variables that indicate if a household has an electric vehicle, electric baseboard heating, or air conditioning, and census data on median household income. Load factor is the average electricity consumption divided by maximum consumption over a specific time period; it is a way to capture the relative utilization rate of consumption at the household level. We then randomized program assignments so households within a cluster were balanced across programs.

[50] Households that received a $0, $10, and $20 upfront incentive accepted the initial invitation with a 63%, 67%, and 68% probability, respectively.

A.2 Comparison of Household Characteristics Upon Randomization

We evaluate if there are differences in pre-treatment characteristics across our various programs to assess the quality of our randomization. Table A1 provides summary statistics by program for a number of variables, including those used in the clustering procedure during randomization (recall the discussion in Footnote 49). The sample presented in this Table represents all 1,661 households invited to participate in the experiment. For all variables, we report the p-values from a one-way ANOVA test to evaluate if there are statistical differences in means across the programs.[51]

Table A1 shows that we do not find significant differences in key characteristics pre-treatment across our programs. These results indicate that our randomization approach effectively achieved balance on observables pre-treatment. In addition, Table A1 demonstrates that the majority of households in our sample have electric hot water heating and use baseboard heating as the primary heat source. In contrast, electric vehicles are less common, representing approximately 30% of households. The majority of households are single-family homes or duplexes, with the remainder being primarily row homes. The households consume considerably more electricity during the winter, with the lowest consumption arising in summer. This demonstrates the potential for larger opportunities for load shifting during these months.

A.3 Comparison of Household Characteristics After Acceptance

We compare the pre-treatment means in observable characteristics by program, including only the households that accepted our invitation to join each program. Large differences in observable characteristics would raise questions about the comparability of our estimated treatment effects from the main specifications.

Table A2 shows observables across programs for the final set of households included in each program. We observe limited differences in these characteristics across programs. The exceptions are that we find a statistically significant difference in the proportion of households that live in single-family homes/duplexes. There is a larger proportion of households in this building type in the Manual program than in other programs, in particular. We also observe a difference across programs in the proportion of households that have EVs, but this difference is only marginally statistically

[51] The seasonal cumulative consumption and load factor data only contain households with a full year's worth of historical consumption. We computed analogous statistics for the entire sample of households using only data from September 2021, the month in which all households have complete pre-treatment consumption data. We find no evidence of statistically significant differences in means across the programs using this data.

Table A1. Comparison of Means by Programs - Initial Randomization

	Central	Tech	Manual	Info	Control	ANOVA (p-value)
Cumul. kWh						
Winter	5,279	5,268	5,442	4,859	5,265	0.27
	(2,694)	(3,032)	(3,076)	(2,748)	(2,950)	
Spring	3,760	3,773	3,818	3,503	3,712	0.48
	(1,924)	(2,112)	(1,911)	(2,116)	(1,974)	
Summer	2,845	2,836	2,708	2,614	2,729	0.54
	(1,742)	(1,872)	(1,539)	(1,861)	(1,710)	
Fall	3,633	3,670	3,700	3,458	3,623	0.66
	(1,663)	(1,945)	(1,974)	(1,796)	(1,860)	
Load Factor						
Winter	24.66	24.98	25.41	24.73	24.67	0.81
	(8.20)	(8.15)	(8.80)	(8.29)	(8.63)	
Spring	19.52	20.12	20.01	19.28	19.91	0.65
	(7.25)	(6.97)	(6.70)	(7.73)	(7.41)	
Summer	16.82	16.55	16.73	16.12	16.32	0.82
	(7.89)	(6.30)	(5.93)	(8.11)	(8.29)	
Fall	18.56	18.90	19.34	18.42	19.06	0.42
	(5.89)	(6.23)	(6.00)	(6.48)	(6.50)	
Electric Vehicle	0.27	0.27	0.27	0.33	0.27	0.41
	(0.44)	(0.45)	(0.45)	(0.47)	(0.45)	
Baseboard Heating	0.61	0.64	0.61	0.63	0.63	0.95
	(0.49)	(0.48)	(0.49)	(0.48)	(0.48)	
Air Conditioning	0.52	0.51	0.50	0.51	0.54	0.95
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	
Electric Hot Water	0.70	0.66	0.70	0.66	0.72	0.38
	(0.46)	(0.47)	(0.46)	(0.47)	(0.45)	
House/Duplex	0.77	0.76	0.81	0.78	0.84	0.17
	(0.42)	(0.43)	(0.39)	(0.42)	(0.37)	
Median Income	86,376	88,291	85,931	87,470	85,948	0.48
	(19,503)	(22,227)	(19,255)	(21,574)	(21,541)	
Households	423	382	409	259	188	

Notes. This table compares pre-treatment average values across the five different programs. Parentheses contain the standard deviations. Cumul. kWh and Load Factor represent the cumulative household-level consumption and load factor by season. The seasonal cumulative consumption and load factor data only contain households with a full year's worth of historical consumption. Electric Vehicle, Baseboard Heating, Air Conditioning, and Electric Hot Water are indicator variables denoting the presence of each device. House/Duplex is an indicator variable that equals one if the home type is a single-family home or duplex and zero otherwise. Median Income reports the median household-level income of the Census Dissemination Area where the household is located. ANOVA reports the p-value from one-way ANOVA tests for differences in means across programs. Statistical significance: $^{*}p < 0.10$, $^{**}p < 0.05$, and $^{***}p < 0.01$.

significant. Overall, these results suggest that the balance on observables that arose due to the initial randomization largely remains in the final sample.

Table A2. Comparison of Means by Program - Final Accepted Households

	Central	Tech	Manual	Info	Control	ANOVA (p-value)
Cumul. kWh						
Winter	5,507	5,302	5,422	5,037	5,265	0.71
	(2,706)	(2,737)	(3,240)	(2,768)	(2,950)	
Spring	3,900	3,739	3,797	3,642	3,712	0.85
	(1,934)	(1,791)	(1,939)	(2,159)	(1,974)	
Summer	2,851	2,672	2,766	2,702	2,729	0.93
	(1,869)	(1,759)	(1,547)	(1,849)	(1,710)	
Fall	3,754	3,550	3,677	3,547	3,623	0.86
	(1,733)	(1,659)	(1,992)	(1,788)	(1,860)	
Load Factor						
Winter	24.62	25.56	24.93	24.93	24.67	0.90
	(8.68)	(8.21)	(9.04)	(7.76)	(8.63)	
Spring	19.33	20.48	19.80	19.59	19.91	0.72
	(7.43)	(6.30)	(6.45)	(7.05)	(7.41)	
Summer	16.33	16.87	16.95	16.61	16.32	0.91
	(8.54)	(6.02)	(5.95)	(7.80)	(8.29)	
Fall	18.17	18.97	19.11	18.53	19.06	0.65
	(6.27)	(5.78)	(6.29)	(6.11)	(6.50)	
Electric Vehicle	0.25	0.21	0.30	0.34	0.27	0.07*
	(0.43)	(0.41)	(0.46)	(0.47)	(0.45)	
Baseboard Heating	0.68	0.70	0.60	0.59	0.63	0.12
	(0.47)	(0.46)	(0.49)	(0.49)	(0.48)	
Air Conditioning	0.46	0.46	0.51	0.51	0.54	0.41
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)	
Electric Hot Water	0.75	0.74	0.68	0.65	0.72	0.16
	(0.43)	(0.44)	(0.47)	(0.48)	(0.45)	
House/Duplex	0.82	0.77	0.89	0.84	0.84	0.02**
	(0.39)	(0.42)	(0.32)	(0.37)	(0.37)	
Median Income	84,978	88,274	86,718	89,504	85,948	0.23
	(19,647)	(20,432)	(19,494)	(21,079)	(21,541)	
Households	177	184	242	177	188	

Notes. This table compares pre-treatment average values across the five different programs for households that were in our final programs. Parentheses contain the standard deviations. Cumul. kWh and Load Factor represents the cumulative household-level consumption and load factor by season. Electric Vehicle, Baseboard Heating, Air Conditioning, and Electric Hot Water are indicator variables denoting the presence of each device. House/Duplex is an indicator variable if the home type is a single-family home or duplex. Median Income reports the median household-level income of the Census Dissemination Area where the household is located. ANOVA reports the p-value from one-way ANOVA tests for differences in means across programs. Statistical Significance * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

A.4 Comparison of Experimental Households to Broader Population

In this section, we present summary statistics to compare our experimental sample with the broader population of households in the Utility's largest metro area. As outlined in Section A.1, our recruitment strategy involved a range of marketing strategies where we invited households to join the Utility's App targeting households within and around the Utility's largest metro region. Once enrolled on the App, we selected a subset of households based on several selection criteria, including home ownership and the ownership of at least one controllable device. This led to a subset of 1,661 households that were randomly offered to enroll in a demand response program (or assigned to the Control). We observe the census dissemination area (CDA) for these 1,661 households. A CDA is the smallest census block in Canada and has a population of approximately 400 – 700 individuals (Statistics Canada, 2021b).[52]

Table A3 presents socio-economic and demographic census characteristics of the CDAs where the 1,661 households offered to join a demand response program (or assigned to the Control) were located, the CDAs of the households that accepted the offers, and the broader characteristics of households in the Utility's largest metro area where the households were recruited from.[53] For the metro area statistics, the summary statistics are constructed using sampling weights, meaning they should be interpreted as population-representative estimates.

Table A3 illustrates that the households in the Invited and Accepted samples live in CDAs with comparable characteristics. This coincides with the balance tables in Tables A1 and A2 that show that the invited and accepted samples are comparable on their mean individual electricity-usage and an array of household-level appliance/devices. Comparing the CDA characteristics of our experimental population to the broader metro area, we find that the households invited and those that accepted the program tend to come from CDAs that have higher average median income, are more likely to have single-family and owner-occupied housing, have a higher proportion of Canadian citizens, and fewer households that self-categorize themselves as immigrants and visible minorities. In Section 8, we provide a discussion of the potential implications this has on the external relevance of our findings.

[52] For comparison, the US Census defines their "optimal" census tract size to contain approximately 4,000 individuals, but tracts can range from 1,200 - 8,000 individuals (U.S. Census Bureau, 2023).

[53] We selected the CDAs to include in the Metro Area definition based on the regions that our Utility partner targeted in their marketing materials. This targeting was constrained to a region where the Utility could send out electricians to install load control devices near the province's largest metro area.

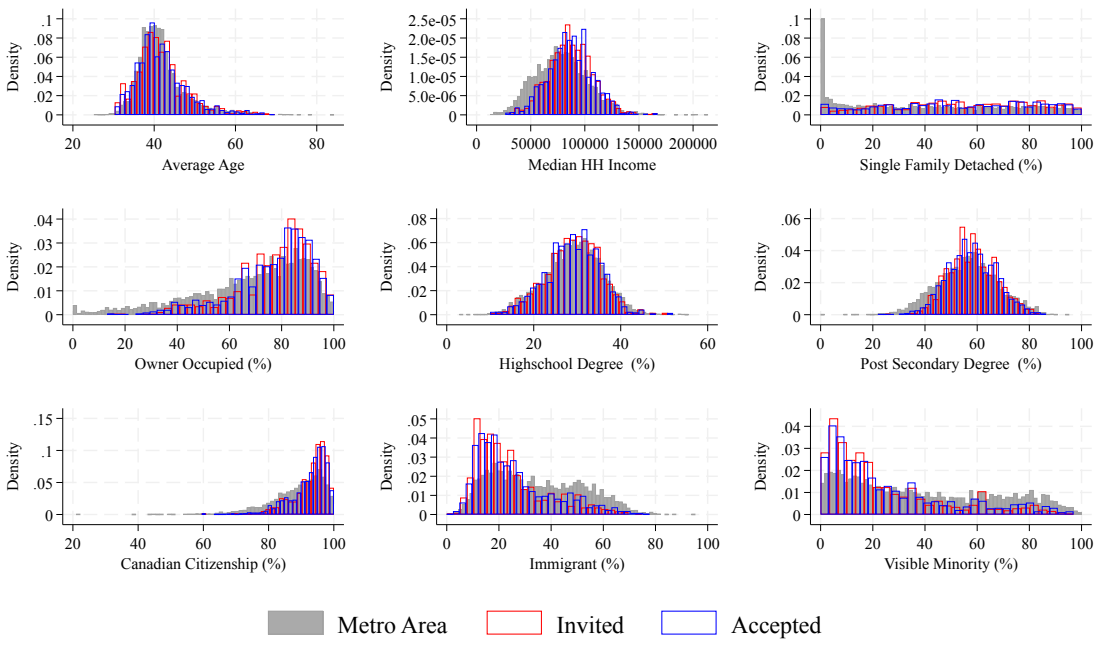
Table A3. Comparison of Census Dissemination Area Characteristics

	Metro Area	Invited	Accepted
Mean Age	41.8	41.8	41.9
	(6.1)	(6.3)	(6.3)
Median Household Income	78,518.2	86,828.9	87,055.4
	(24,875.0)	(20,657.2)	(20,420.7)
Single Detached (%)	39.7	52.8	54.2
	(31.8)	(27.8)	(26.8)
Owner Occupied (%)	68.1	76.8	77.3
	(21.8)	(15.2)	(14.5)
High School Education (%)	29.1	29.1	29.0
	(6.7)	(6.2)	(6.2)
Post Secondary Education (%)	56.3	58.0	58.1
	(11.7)	(9.3)	(9.0)
Canadian Citizenship (%)	88.7	92.4	92.8
	(8.4)	(5.9)	(5.6)
Immigrant (%)	34.8	25.7	24.2
	(17.3)	(14.6)	(13.6)
Visible Minority (%)	39.3	25.7	23.5
	(27.9)	(23.8)	(22.0)
Observations	3,382,162	1,661	968

Notes. This Table summarizes data from the 2016 Canadian Census by census dissemination area (CDA). The data presents mean and standard deviations (in parentheses) for households in CDAs in the metro area where participants were recruited, the households invited into a demand response program or never-treated groups, and the households that made it into our experiment (i.e., those that Accepted). The variables Canadian Citizenship (%), Immigrant (%), and Visible Minority (%) come from the long-form census that is sent to approximately 25% of households in Canada. For the Metro Area numbers, published averages are constructed using sampling weights, meaning they should be interpreted as population-representative estimates for that geographic area rather than raw sample averages.

Figure A1 presents the distribution of these variables, going one step further beyond the simple averages in Table A3. These results demonstrate that it is not just the mean values, but the full distributions of the Invited and Accepted samples are comparable. Comparing these two samples to the broader metro area, we can get a more nuanced understanding of what is driving the differences in means. For example, for the median household income, our sample of invited/accepted households has fewer households at the bottom end of the income distribution and more in the middle of the income distribution. The differences in single family housing percentages is driven by the large mass of near-zero in the metro sample, capturing CDAs in the dense city center that is largely absent from our sample. While our distribution lies broadly across the full spectrum of variables that relate to citizenship, % immigrants, and % visible minority, it is more heavily weighted to citizens and lower immigrant CDAs.

Figure A1. Distributional Comparison of Census Dissemination Area Characteristics



Notes. This Figure plots data from the 2016 Canadian Census by census dissemination area (CDA). The data presents the distribution of CDA-level values for several key variables, including Canadian Citizenship (%), Immigrant (%), and Visible Minority (%) come from the long-form census that is sent to approximately 25% of households in Canada. For the Metro Area numbers, published averages are constructed using sampling weights, meaning they should be interpreted as population-representative estimates for that geographic area rather than raw sample averages.

Finally, households enrolled in our experiment were required to have at least one of the following appliances: an electric hot water heater, electric baseboard heating, and/or a level 2 electric vehicle charger. As a result, the household devices installed play an important role in our analysis. We leverage publicly available data on appliance/device adoptions within our metro region, the broader province, and Canada-wide to better understand the external relevance of our analysis. We use data as close to the start date of our experiment as possible.

Using 2021 data from Statistics Canada, within the Utility's metro area and the broader province, electric baseboard heating served as the primary source of heating for approximately 30% of households. This compares to 25% Canada-wide (Statistics Canada, 2022). In a 2019 nationally representative survey on household's energy usage from Natural Resources Canada (NRCan), 33% of households in the Utility's province reported having an electric hot water heater. This compares to 41% across Canada (Natural Resources Canada, 2019). While broad statistics on the presence

of a level 2 electric vehicle (EV) charger are not publicly available to the best of our knowledge, the 2019 NRCan survey reports that 9% of households in the Utility's province have EV and 4.4% of households Canada-wide reported having an EV.

Tables A1 and A2 illustrate that the households that were invited and enrolled in our program are more likely to have each of the devices compared to the province-wide and Canada-wide averages. While these statistics illustrate that our sample includes relatively highly electrified set of households compared to households in the Utility's largest metro area and Canada more broadly, the broader statistics on appliance/device adoption rates show that these devices are used by a sizable proportion of households across the province and country.

B Treatment Details

B.1 Program-Specific Event Notifications

Each treatment program experienced event notifications tailored to their treatment. Each program received a notification 21 and 2 hours before an event. All participants were shown a short notification according to their device and in-app notification settings. If participants touched and pressed the notification, they were shown the long notification specific to their program, featured below, with event incentive details.

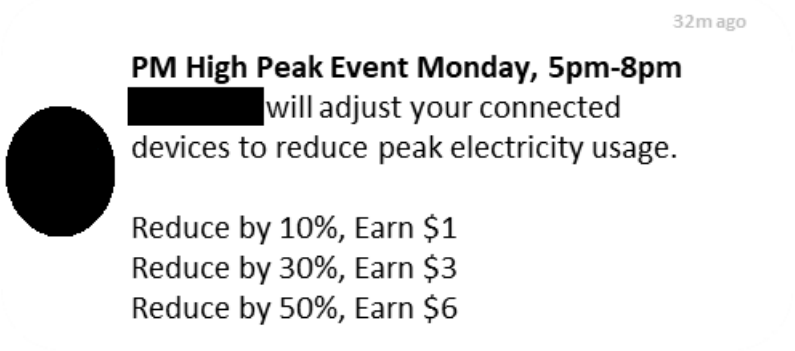


Figure B1. Long Notification for Central program

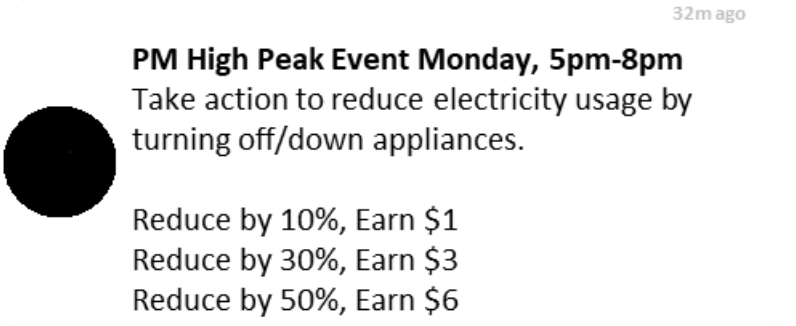


Figure B2. Long Notification for Tech program

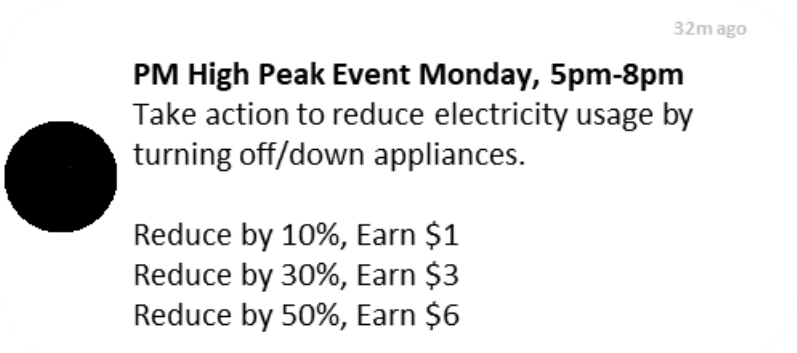


Figure B3. Long Notification for Manual program

Note that all program participants in the three programs were able to locate event details in the "Advisor" tab of the App, a centralized location for information from the App company, once they received an event notification. The "Learn More" button at the bottom right of this information card took participants to the "FAQs" section of the program-specific experiment website.

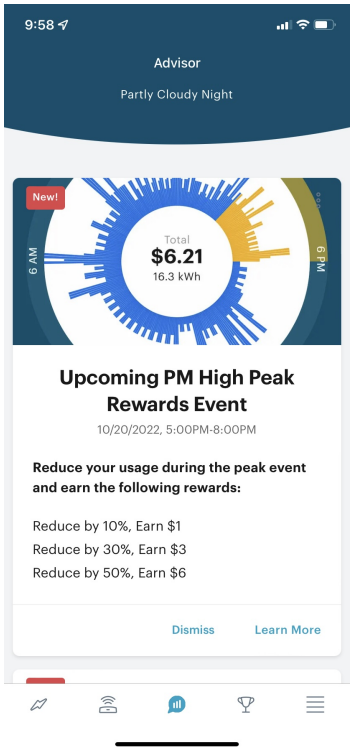


Figure B4. Event info in App

B.2 Treatment Program-Specific App Functionality

Each program in our experiment had an App experience and functionality that differed according to their program assignment. We detail that here and walk through how participants in each program could have responded to peak events, given the options in the App.

B.2.1 Central Program

The Central program participants receive 21-hr and 2-hr notifications regarding upcoming events, as described in Appendix B.1. These notifications allow them to see the timing of the event and the magnitude of rewards for electricity consumption reductions. They also remind participants that their devices with load controllers would be altered by the Utility to reduce consumption, unless they opted-out of the event.

There are several ways that Central program participants can opt-out of events. Before an event starts, they can push an "Opt-out" button in the "My Devices" tab of the App (Figure B5). (This tab is a central App location that allows App users to remotely control devices that have load controllers and see the individual electricity consumption of those devices.) This button removes the participant from the event globally by removing all of their load-controlled devices from the event.

If they do not opt-out in this way, they see a series of screens in the "My Devices" tab. These indicate the progression of the event to the participant and signal when their devices' electricity consumption is being controlled by the utility, via the icons above the text "You are opted in", "Event", and "Complete" (Figure B6).

During an event, participants can cancel Utility device control in a device-specific way. For EV chargers and hot water heaters, they can remotely opt-out their device from being controlled, or they can physically turn off the load controller at the device itself. For thermostats, participants can opt-out of load control by adjusting them physically or remotely through the App, during an event.

Note that the Central program has remote and manual control of all devices with load controllers, just like the Tech program. Central program households can also change anything else in the house to alter their electricity consumption during events.

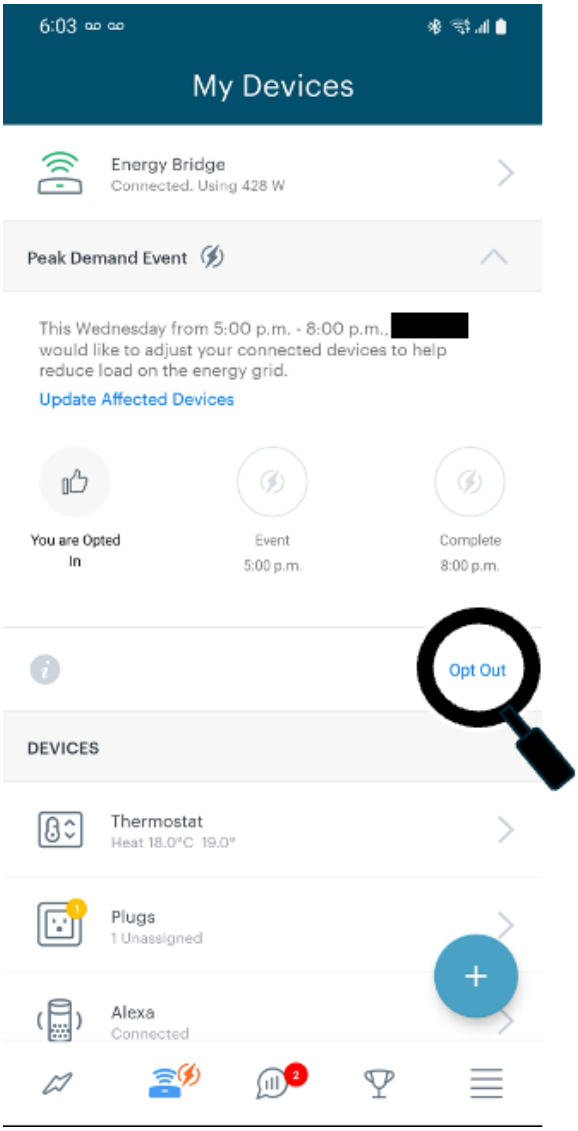


Figure B5. Central Program Opt-Out Functionality

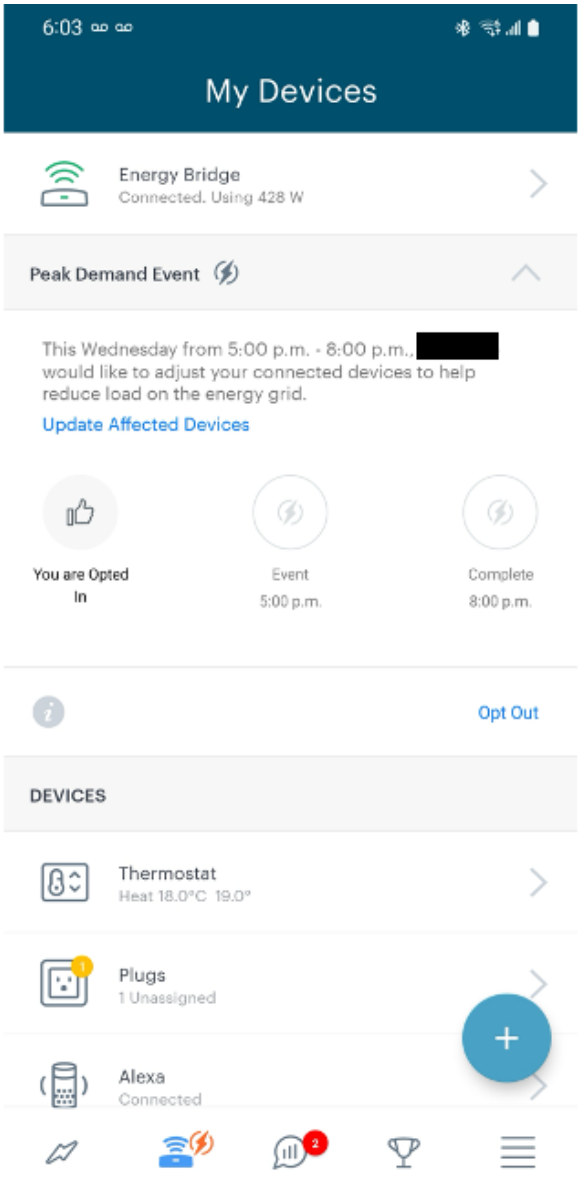


Figure B6. Central Program Event Experience

B.2.2 Tech Program

The Tech program participants receive 21-hr and 2-hr notifications regarding upcoming events, as described in Appendix B.1. These notifications allow them to see the timing of the event and the magnitude of rewards for electricity consumption reductions. They also remind participants that they need to "take action" to make consumption changes to receive the rewards offered.

The Tech program can remotely control any device that has an installed load controller through the App. For EV chargers and hot water heaters, they can turn them off via two clicks from the My Devices section of the App. (See Figure B7 below for the instructions sent to participants that explain these actions.) Tech program participants cannot make a schedule to turn off these devices before events start and must turn them off before or during events to reduce consumption this way. (They must also remember to turn them on unless they set up a turn-on schedule.)

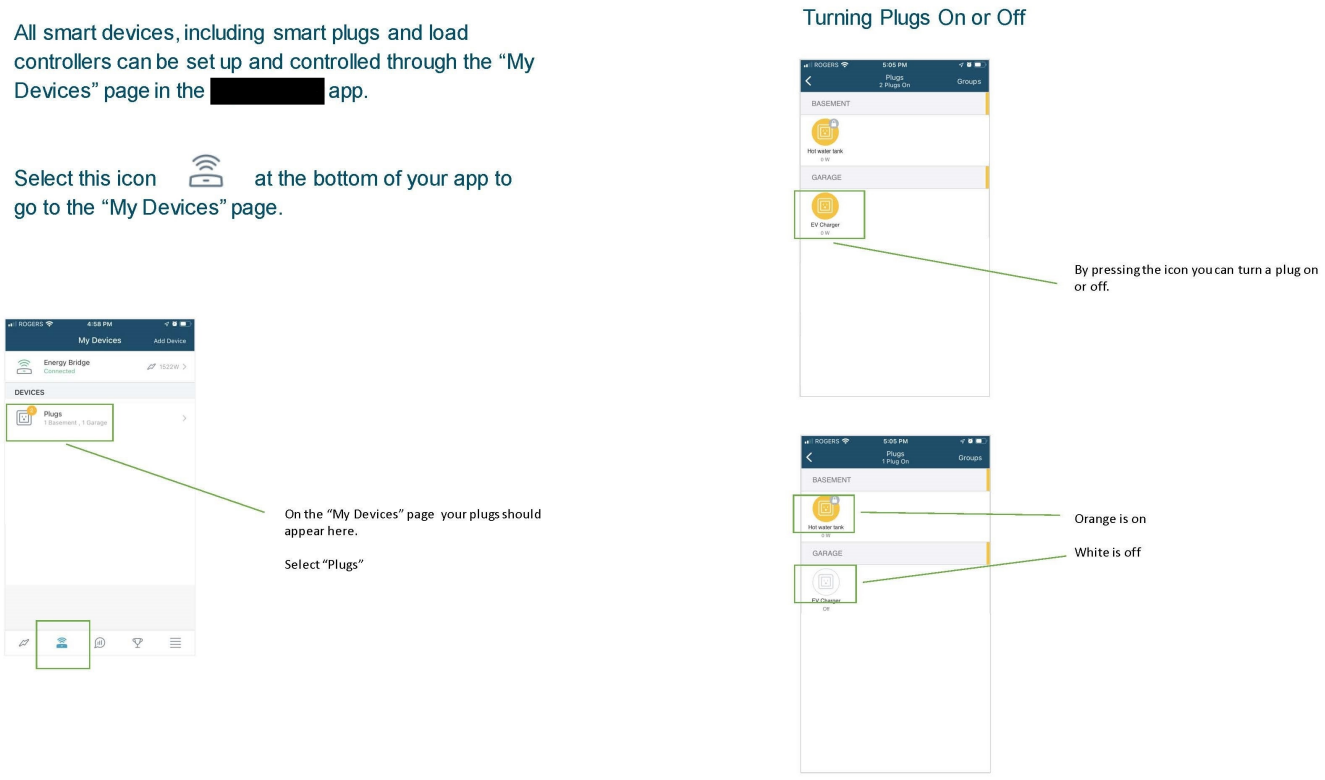


Figure B7. Controller Guide for Tech Program

For thermostats, the Tech program can set up schedule for their thermostat set-

point before events, using the App. They can also adjust their thermostats remotely during events with the App.

B.2.3 Manual Program

The Manual program participants receive 21-hr and 2-hr notifications regarding upcoming events, as described in Appendix B.1. These notifications allow them to see the timing of the event and the magnitude of rewards for electricity consumption reductions. They also remind participants that they need to "take action" to make consumption changes to receive the rewards offered.

Manual program participants do not load controllers given to them as part of this experiment or Utility control of any devices. They therefore only observe these notifications as well their aggregate, real-time household consumption through the App. If Manual program participants install their own smart home devices, they may be able to link them to the smart electricity consumption technology ecosystem used in this experiment. If so, they may have the capabilities of the Tech program to observe the real-time consumption of those devices/devices individually and adjust them remotely through the App. (Only three households in the Manual program installed their own smart thermostats over our sample period.)

B.2.4 Central, Tech, and Manual Programs

After each event, all three of the Central, Tech, and Manual programs receive a result on their performance, as depicted below. This appears in the "Advisor" tab of the App, a central location for information from the App company. This result card reminds participants of the event type (reward magnitudes being "high" or not) and the day and time of the event. It shows the incremental reward the participant earned from the event as well as their cumulative rewards throughout the entire experiment, including the reward from the prior event. The text below the reward for the last event is variable and depends on whether a participant met one of the reward tiers. The rewards screen with one of these text options is shown below in Figure B8.

From this rewards screen, participants can select "Event History" and see their recent history of event rewards, as shown in Figure B9.[54]

[54]Figure B9 was created for illustrative purposes using a series of simulated events. As a result, the event times differ from the event times considered in our study (i.e., 7:00 AM - 10:00 AM and 5:00 PM - 8:00 PM).

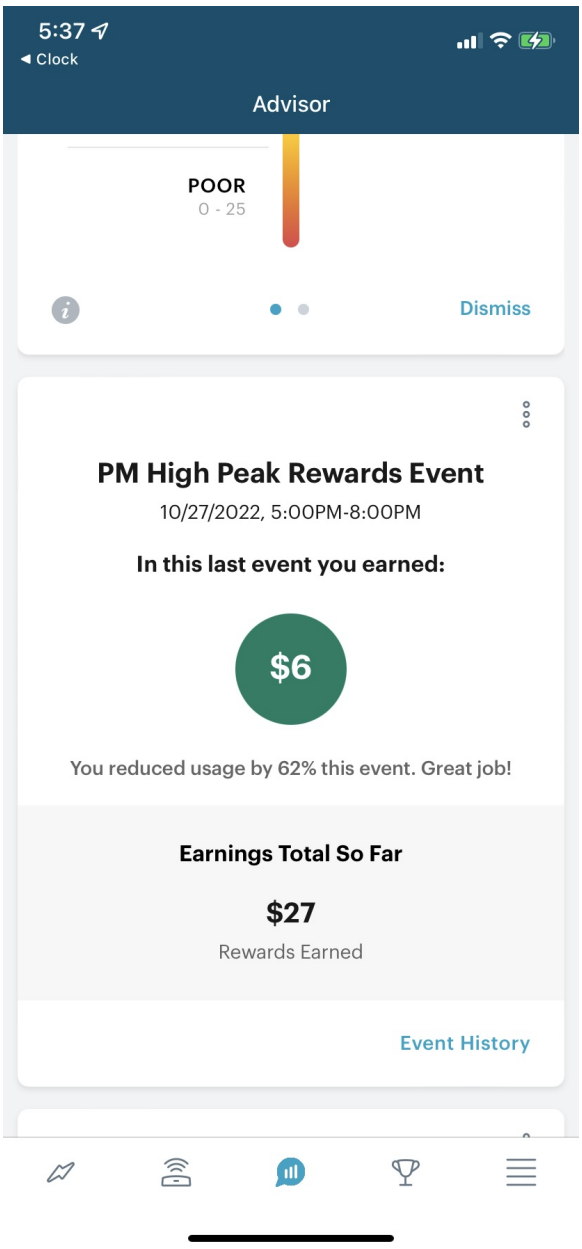


Figure B8. Rewards Screen

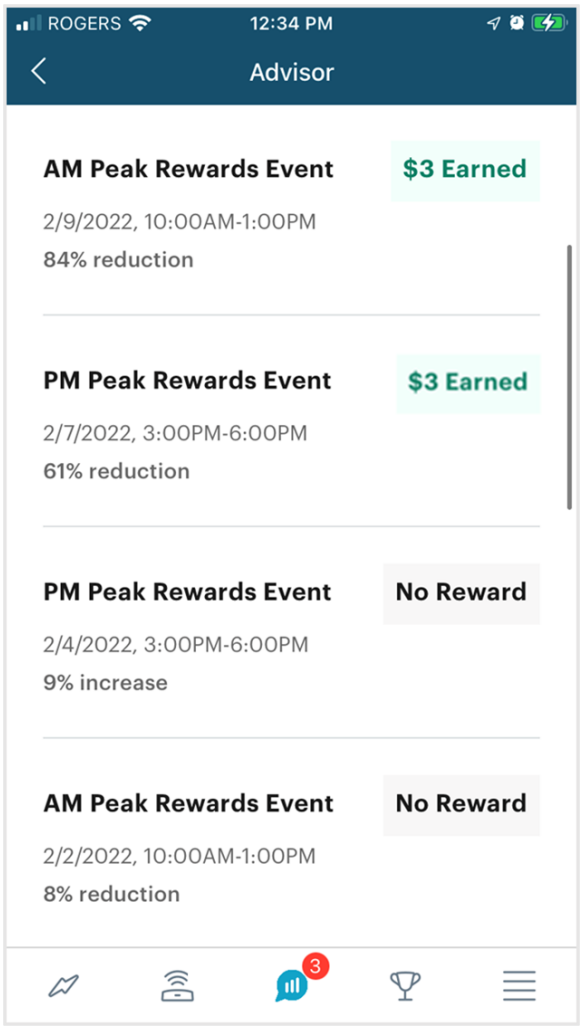


Figure B9. Event History

C Extensions and Robustness

C.1 Causal Transience Analysis

In this section, we carry out robustness checks to address the causal transience assumption. First, we carry out a test to evaluate if households adjusted their consumption in event hours on non-event days once peak events began. Second, we evaluate if there is any evidence that households responded differently depending on the number of days between peak events or depending on the event type of the prior event. Third, we present additional results related to heterogeneous average treatment effects of peak events and investigate Central Program's opt-out behavior over the duration of our 17-month experiment.

C.1.1 No Carryover to Non-Event Days Post-Treatment

As discussed in Section 4.3, the causal transience assumption requires that our randomized peak events do not lead to persistent changes in consumption in the event hours on non-event days. We test the validity of this assumption by estimating a DID regression. Separately for each demand response program, we run the following regression that excludes event days in the post-treatment period and includes the never-treated households:

$$ln(c_{it}) \;=\; \beta\, D_i \cdot \text{EventWindow}_{it} + \alpha_i + \tau_t + \delta X_{it} + \varepsilon_{it} \tag{4}$$

where EventWindow_{it} equals 1 in the post-treatment period for hours where morning or evening events occur and 0 otherwise. D_i is a categorical variable that captures if the household is in a demand-response program (i.e., C, T, or M) and 0 otherwise. All other features of the regression analysis are identical to those in Equation (1).

Because peak events started on February 22, 2022, EventWindow_{it} effectively "turns on" in the relevant peak event hours after this date on non-event days. This analysis evaluates whether households in each demand response program adjusted their consumption during the event windows on non-event days in the post-treatment period, relative to the never-treated households. If households treated with events did not systematically alter their behavior on non-event days in response to being exposed to events on other days (e.g., reflecting some type of systematic adaptive behavior/habit formation after receiving events), β should be statistically indistinguishable from zero. In addition, we consider a specification that estimates separate

effects for the morning and evening event windows on non-event days.

Table C1 presents the results of our no carryover assumption DID test across all events, while Table C2 presents the results when we allow for differential effects across the morning and evening event windows on non-event days. In both specifications, we find no evidence of changes to non-event day consumption during the event windows.

Table C1. Carry Over DID Estimates by Program

	Central	Tech	Manual
Event Window	0.0241	0.0302	-0.0014
	(0.0169)	(0.0181)	(0.0155)

Notes. The reported results are the program-specific Event Window coefficients from equation (4). For each demand response program, the sample includes households from their own treatment program and the never-treated groups. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

Table C2. Carry Over DID Estimates by Program and Event Type

	Central	Tech	Manual
Morning Event Window	0.0188	0.0353	-0.0226
	(0.0206)	(0.0235)	(0.0201)
Evening Event Window	0.0295	0.0251	0.0196
	(0.0212)	(0.0210)	(0.0185)

Notes. The reported results are the program-specific Event Window coefficients from equation (4), allowing for differential effects during the morning and evening event windows. For each demand response program, the sample includes households from their own treatment program and the never-treated groups. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

C.1.2 Gaps between Event Days and Prior Event Type

Additional sequence-dependent channels could arise via households responding differentially depending on the number of days between peak events or if the prior peak event occurred in the morning or evening hours. We run two additional specifications to address each concern. First, there is an average of 9.5 days between peak events, with an inner quartile range of 6 to 12 days. We create a categorical variable that

represents the quartile of days between peak events and interact this with the peak event indicator variable in equation (1). This allows us to test if there are heterogeneous treatment effects of peak events by how long it has been since a household was exposed to a peak event. Second, we extend our regression in equation (1) by not only allowing for differential responses by event type (morning, evening, and high evening), but we also allow for heterogeneous effects by whether or not the prior event occurred in the morning or evening event windows.[55]

Table C3 presents results of the average estimated treatment effect of peak events by quartile of days since the last event. Across all three demand response programs, we do not find significantly different effects across the quartiles.

Table C3. Treatment Effects of Participants by Program and Days Since Last Event Quartile

	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Central	-0.3115	-0.3011	-0.3095	-0.2841
	(0.0232)	(0.0232)	(0.0226)	(0.0223)
Tech	-0.037	-0.068	-0.0527	-0.0414
	(0.0189)	(0.0177)	(0.0201)	(0.0190)
Manual	-0.0515	-0.0575	-0.0592	-0.0497
	(0.0148)	(0.0149)	(0.0153)	(0.0155)

Notes. The reported results are the program-specific Event coefficients from equation (1), allowing for differential effects by days since last event quartile. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

Table C4 presents results that evaluate if there are differences in the estimated treatment effects by prior event timing (i.e., morning or evening hours). There is only one significant difference across all programs and event types; the Tech program's response to evening peak events are marginally significantly higher if the prior event was a morning event. The effect reverses in direction for the high evening event treatment effect for the Tech program with the response being larger if the last event occurred during evening hours (although it is not statistically significant). Consequently, we intepret these results to indicate that there is limited evidence of any systematic differences in responses depending on the last event type.

[55] We considered a further granular specification that permitted differential effects if the prior event was an evening or high evening event. We continue to find no evidence of heterogeneous effects on this margin.

Table C4. Treatment Effects of Participants by Program, Event-Type, and Last Event Type

	Prior Event - Morning	Prior Event - Evening
Central		
Morning	-0.3187	-0.2806
	(0.0235)	(0.0340)
Evening	-0.2988	-0.2978
	(0.0244)	(0.0269)
High Evening	-0.3237	-0.3257
	(0.0275)	(0.0290)
Tech		
Morning	-0.0183	-0.0058
	(0.0195)	(0.0338)
Evening	-0.0837	-0.0464
	(0.0220)	(0.0222)
High Evening	-0.0666	-0.0843
	(0.0227)	(0.0292)
Manual		
Morning	-0.0846	-0.1159
	(0.0158)	(0.0290)
Evening	-0.0456	-0.0587
	(0.01720	(0.0182)
High Evening	-0.0567	-0.0569
	(0.0186)	(0.0230)

Notes. The reported results are the program-specific Event coefficients from equation (1), allowing for differential effects by Event Type and if the past event was a morning or evening/high evening event. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

C.1.3 Dynamic Treatment Effects and Opt-Out Behavior

Figure C1 presents the distribution of household-level estimated treated effects, permitting differential effects by Phase. Similar to the program-level estimates reported in Figure 5, we can observe a reduction in the estimated treatment effect for all three demand response programs. The Central program, with the documented technical disconnections occurring during Phase 1, has the largest change between Phase 1 and Phases 2 and 3.[56] For all three programs, we continue to see a stabilization in the

[56]For the Central Program, the average estimated household-level treatment effect represents a 27.1%, 22.2%, and 24.1% reduction in Phases 1, 2, and 3, respectively. For Tech (Manual), we

distribution of the estimated effects in Phases 2 and 3. While the Central program exhibits the most distinct change, the estimated effects in the Central program remain systematically above the Tech and Manual programs for all three phases of our experiment.

Figure C2 expands on the phase-specific analysis in Figure 5 by presenting heterogeneous program-level treatment effects by month. This analysis provides further insight into the dynamics over our 17-month program, but comes with the cost of precision as we are estimating a larger number of coefficients.

For the Central program, we continue to observe a reduction in the estimated program-level treatment effect in the first couple of months, with the smallest effects arising in the summer of 2022. This summer-specific feature likely reflects reduced controllable loads with limited heating in the summer. The estimated effects then level off and remain above the Tech and Manual program estimates. For the Tech program, we see an elevated estimated peak event response early on in our sample. It is possible that this reflects usage of the connected technologies initially, with a leveling off after the first 5 months of our program. For the Manual program, we see a slight reduction in the estimated peak event treatment effect over time. These results are consistent with our phase-specific analysis.

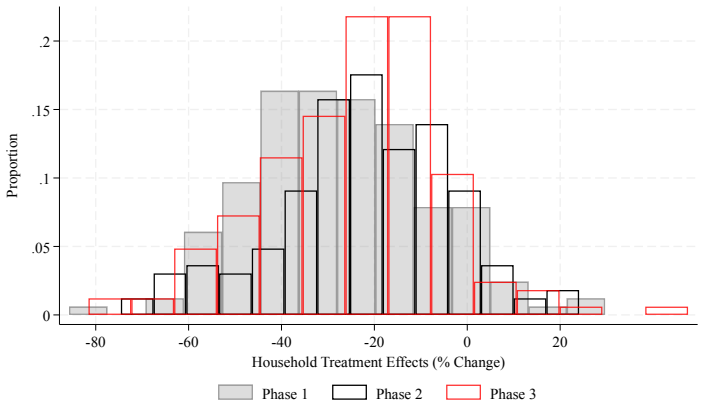
One avenue through which households in the Central program could adjust their behavior running up to or during peak events was via actively opting out of events. We observe an approximate 4% opt-out rate across all peak events for the Central program, with 90% arising via thermostat adjustments. We investigate if event opt-out behavior changed over the duration of our experiment.

Figure C3 presents the percentage of event days with opt-outs at the device-level installed on Central program households, along with the average monthly temperatures. We see a distinct seasonal pattern, with higher opt-out rates negatively associated with average temperatures. This is consistent with the observation that when opt-outs occur, they primarily arise via baseboard thermostat overrides/adjustments. Visual inspection suggests that there is limited evidence of a systematic changes in opt-out rates outside of clear seasonal weather patterns.

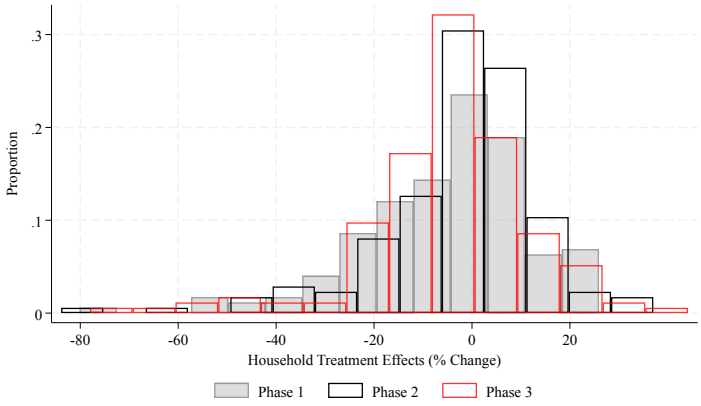
To more formally test this, we carry out a regression analysis at the device-event day-level. For our dependent variable, we use an opt-out indicator variable that equals one if an opt-out occurs and zero otherwise. Our primary objective is to estimate

observe an estimated demand reduction of 5.5%, 2.3%, and 4.4% (4.5%, 2.9%, and 3.6%) during Phases 1, 2, and 3, respectively.

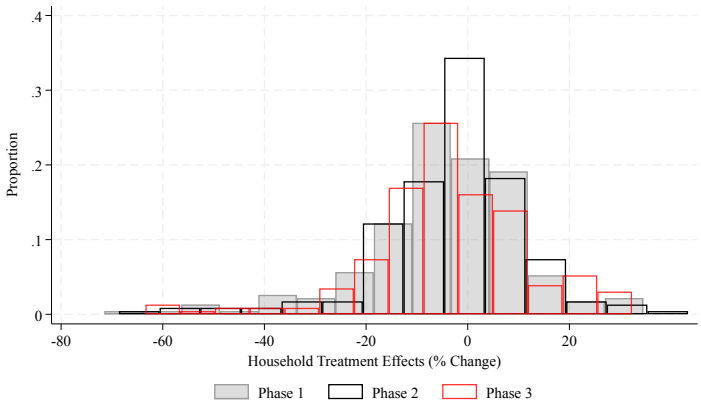
Figure C1. Household-Level Estimated Treatment Effects by Program and Phase



(a) Central Program



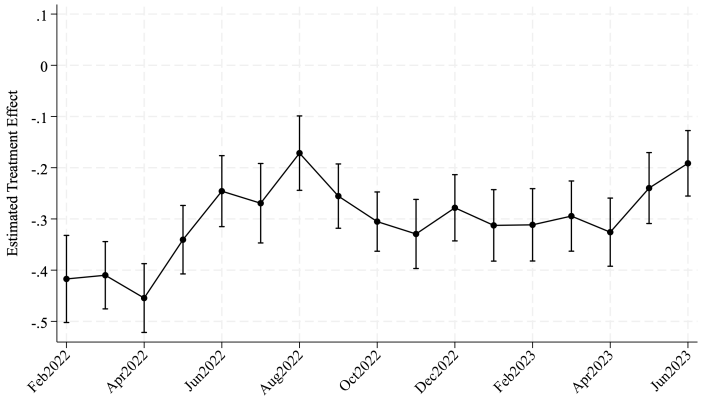
(b) Tech Program



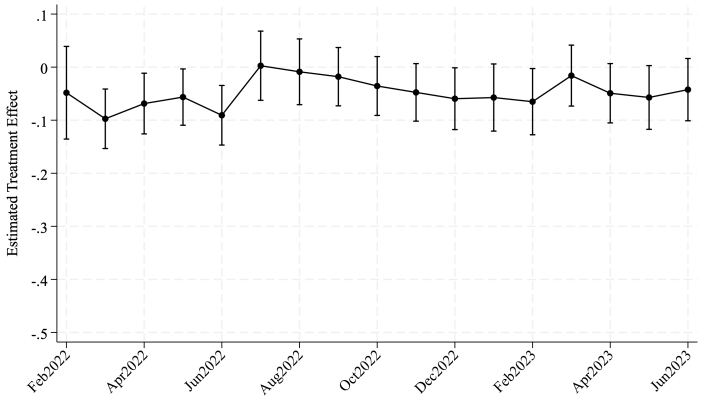
(c) Manual Program

Notes: The reported results summarize the distribution of estimated household-level treatment effects obtained from estimating specification (2), permitting heterogeneous effects by Phase. We consider three phases: (1) February 2022 - July 2022, (2) August 2022 - January 2023, and (3) February 2023 - June 2023. The histogram presents the proportion of observations that fall within a specific bin's range. We present the marginal effects on the x-axis to be a percentage change in consumption using the transformation $100 \times (\exp(\widehat{\beta}_i) - 1)$.

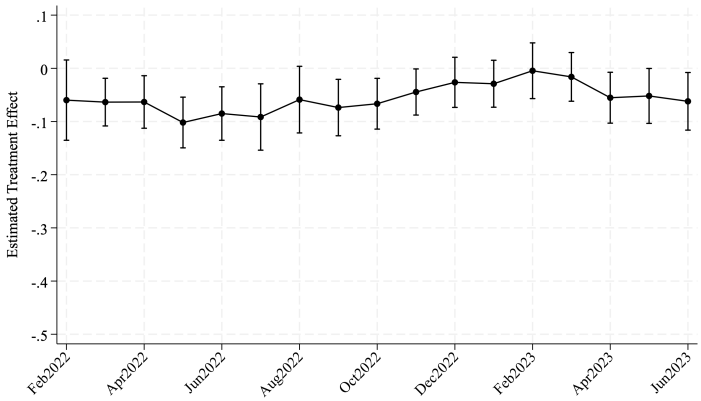
Figure C2. Program-Level Estimated Treatment Effects by Program and Month



(a) Central Program



(b) Tech Program



(c) Manual Program

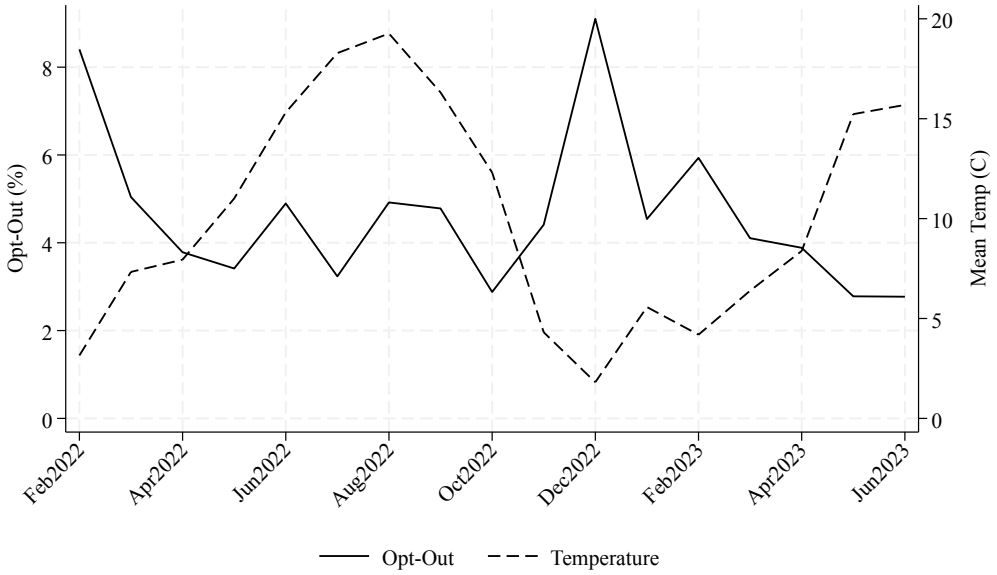
Notes: The reported results are program-specific marginal effects calculated from estimating $\widehat{\beta}_j$ in (1) for $j \in \{C, M, T\}$, adjusted to allow for monthly interactions with the program-event indicator variables $Program_{ji} \cdot E_{it}$. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

month-specific coefficients to quantify opt-out rates over time. We include cubic controls for average daily cooling and heating degrees and humidity and fixed effects for each device, day-of-week, and event-type. These controls aim to remove weather, calendar and event, and device-specific time-invariant factors that could confound potential time trends. Standard errors are clustered at the device-level.

Figure C4 reports the estimated month-specific coefficient estimates. The estimates indicate a relatively higher percentage of opt-outs during the first 5 weekdays of peak events that occurred in February 2022 (i.e., events started on Feb 22, 2022). While we are not able to test if there were year-over-year changes for all months, we can compare Feb - Jun 2022 to their corresponding months in 2023. Only February had a significantly higher opt out rate at the 5% level than February 2023.

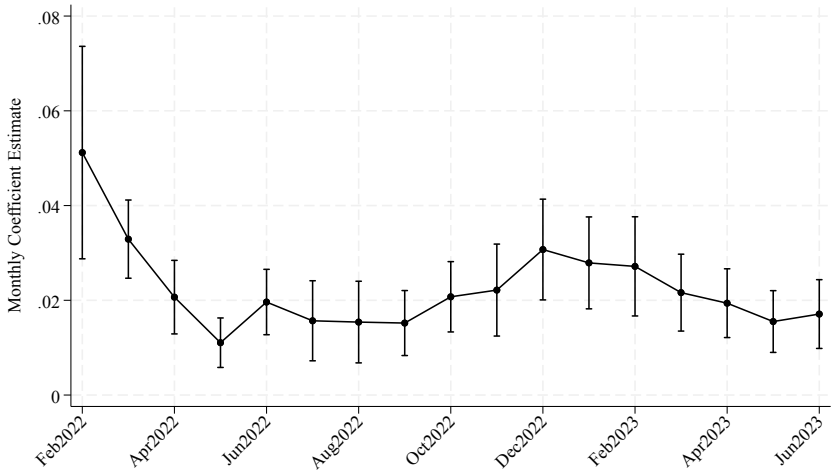
We continue to see an elevated opt-out rate in the winter months even after including our daily temperature controls. This corresponds with the fact that opt-out rates are driven by the usage and subsequent utility-control of baseboard thermostats more broadly. Once we remove February 2022, we do not find significantly higher opt-out rates in March 2022 from the winter/spring months in 2023, higher opt-out rates in the spring 2022 months from the spring 2023 months, or higher opt-out rates in the summer 2022 months and June 2023. Taken together, we view these descriptive results to suggest that there is limited evidence of systematic changes in opt-out rates for the Central program over time, outside of the first few days of the experiment.

Figure C3. Central's Opt-Out Percentage and Average Temperature by Month



Notes: This Figure presents the percentage of event days with opt-outs at the device-level, alongside the average monthly temperature (in Celsius).

Figure C4. Regression Estimates of Month-Specific Opt-Out Percentages



Notes: This Figure presents the monthly coefficient estimates of a regression analysis at the device-event day level for utility-controlled devices installed at Central program households. The dependent variable is an indicator measuring event opt-outs. Control variables include indicator variables for each month of our experiment, cubic controls for average daily cooling and heating degrees and humidity, and day-of-week, event-type, and device-level fixed effects. Standard errors are clustered at the device-level.

C.2 Attrition

In this section, we address the extent and nature of attrition. There are two types of attrition that could arise in our setting and lead to households no longer participating in our experiment and their data no longer being available: (i) attrition where households request to leave the experiment or (ii) attrition where the household's close their Utility's account because they move and presumably sell their homes.

We observed few households requesting to unenroll from of the program altogether, only 3 households explicitly asked to be unenrolled (with 2 in the Manual and 1 in Central program). We had a small subset of users unenroll from the experiment because of Utility account closures. Discussions with our Utility partner indicate that these represent households that sell their homes and move.

Table C5. Enrolled Participants by Program and Month

Year	Month	Central	Tech	Manual	Info	Control
2022	1	177	184	242	175	188
2022	2	177	184	242	175	188
2022	3	177	184	242	175	188
2022	4	177	184	242	173	188
2022	5	176	184	241	172	188
2022	6	175	184	241	171	187
2022	7	173	184	238	170	185
2022	8	172	182	237	168	184
2022	9	171	182	237	168	180
2022	10	170	181	237	168	178
2022	11	169	180	237	168	177
2022	12	169	179	236	167	177
2023	1	169	179	235	167	177
2023	2	168	178	234	167	177
2023	3	166	177	233	166	176
2023	4	166	177	231	166	176
2023	5	166	177	231	165	175
2023	6	166	175	231	165	175
Final Attrition (%)		6.21	4.89	4.55	6.78	6.91

Notes. This table reports the count of households that were enrolled by program. Final attrition (%) represents the percentage of initially enrolled users that unenrolled by the end of our experiment (June 2023). 2 Info users attritted from the program between the initial enrollment starting in October 2021 and January 2022, leading to 175/177 enrolled in January 2022.

Table C5 demonstrates that the attrition rates over the 17-month demand response program are modest and fall in the range of 4.55% – 6.91%. The Control group provides us with a beneficial baseline as they did not receive any intervention. We passively received their data. Their disconnections represent natural account closures, which we take to mean that they are moving households and closing their Utility accounts. The attrition rate in this group reaches 6.9% by the end of our sample. We observe comparable attrition rates across the Central, Tech, Manual, and Info programs that received varying degrees of intervention intensity. Attrition was gradual across each program, consistent with natural turnover from movers. We use a two-sample test of proportions for each pair-wise program comparison and find no evidence of a statistically differential attrition rates across programs by June 2023 (end of our experiment). This mitigates concerns that attrition was differential by demand response program.

Our within-subject design, as outlined by List (2025), comes with the assumption of a balanced panel. Our main specifications include all households, hereon the Full sample. We carry out a robustness check that re-estimates our key results, but focuses only on households that remained in the experiment for the full 17-months (Balanced sample). We consider a version of this analysis focusing on the entire 17-month experiment, and permitting differential effects by Phase 1, 2, and 3 using the approach to estimate dynamic treatment effects outlined in Appendix C.1.3. This allows us to assess if our heterogeneous treatment effect estimates over time are impacted by the presence of attrition.

Table C6 reports the coefficient estimates using Equation (1) in our program-level regressions across Phases 1, 2, and 3 for the Full (in columns (2) - (4)) and Balanced samples (in columns (6) - (9)). The results for the Full and Balanced samples for the entire experimental period are reported for reference in columns (1) and (5), respectively. We can see that our results and key qualitative conclusions are highly robust across each analysis when comparing the Full and Balanced samples.

Figure C5 reports the distribution of our household-level estimates for the Full and Balanced samples across the Central, Tech, and Manual programs, using the specification in Equation (2). We continue to see that our results are robust to removing the households that attrit over our experiment.[57] The distribution of the Manual program is adjusted slightly when focusing on the Balanced sample as there is

[57] We carried out an analysis that looks at household-level estimates over Phases 1, 2, and 3, split by the Full and Balanced sample. Similar to our program-level results reported in Table C6, we find that our results are robust to excluding the attrited households.

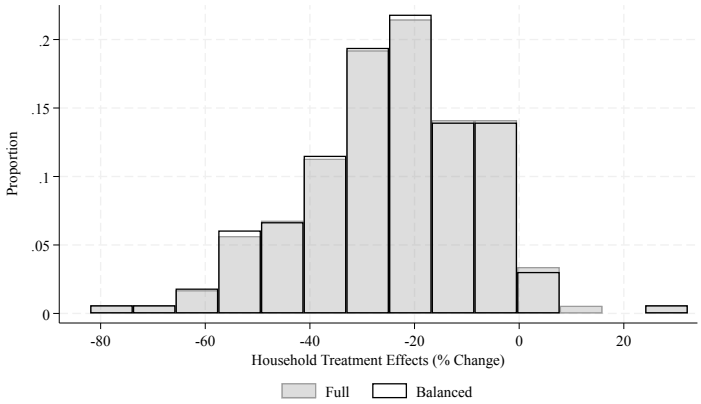
a reduced mass at zero. However, the average treatment effect remains approximately -4% with a similar distribution.

Table C6. Program-Level Estimates by Program and Phase with Full and Balanced Samples

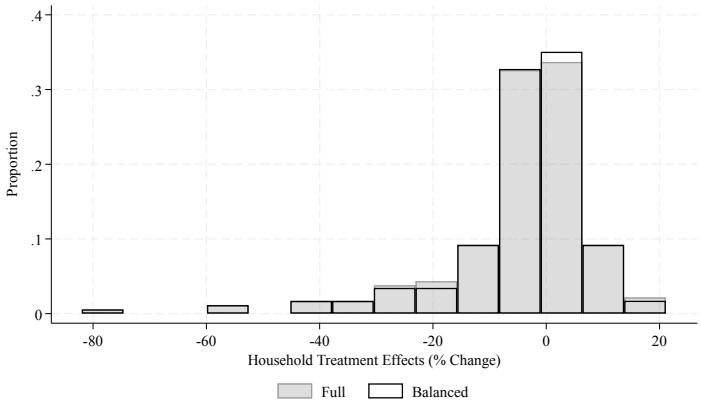
	(1) Full	(2) Phase 1	(3) Phase 2	(4) Phase 3	(5) Balanced	(6) Phase 1	(7) Phase 2	(8) Phase 3
Central	-0.3036	-0.3608	-0.2824	-0.2717	-0.3042	-0.3646	-0.2823	-0.2730
	(0.0204)	(0.0242)	(0.0212)	(0.0247)	(0.0208)	(0.0250)	(0.0216)	(0.0248)
Tech	-0.0496	-0.0656	-0.0398	-0.0453	-0.0536	-0.0731	-0.0432	-0.047
	(0.016)	(0.0184)	(0.019)	(0.0194)	(0.0161)	(0.019)	(0.0191)	(0.0194)
Manual	-0.0542	-0.0783	-0.0484	-0.0377	-0.0538	-0.0765	-0.0494	-0.0376
	(0.0123)	(0.0154)	(0.0139)	(0.0157)	(0.0124)	(0.0158)	(0.014)	(0.0158)

Notes. The reported results are program-specific treatment effect coefficients using the specification in Equation (1). Standard errors are reported in the parentheses and clustered at the household level. Columns (1) and (5) report the regression results for the Full and Balanced samples for all months. Columns (2) - (4) and (6) - (8) report the regression results including the Full and Balanced samples across Phases 1, 2, and 3, respectively. All specifications include fixed effects at the household and hour-of-sample levels. Standard errors are clustered at the household-level.

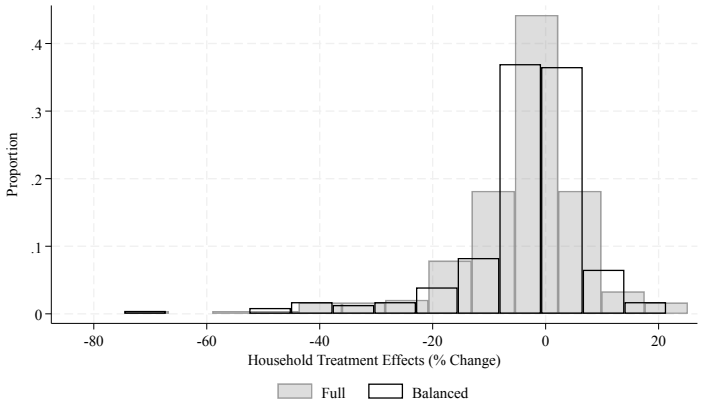
Figure C5. Household-Level Estimated Treatment Effects by Program with Full and Balanced Samples



(a) Central Program



(b) Tech Program



(c) Manual Program

Notes: The reported results summarize the distribution of estimated household-level treatment effects obtained from estimating specification (2), separated by the Full and Balanced Samples. The histogram presents the proportion of observations that fall within a specific bin's range. We present the marginal effects on the x-axis to be a percentage change in consumption using the transformation $100 \times (\exp(\widehat{\beta}_i) - 1)$.

C.3 Sample Selection Analysis

In this section, we present several analyses to address potential concerns over sample selection across our demand response programs. First, we present estimates along the lines of an intention-to-treat (ITT), including households that did and did not accept the program offers. Second, we perform a bounds analysis to serve as a lower-bound on our Central and Tech program's estimated event treatment effect, accounting for differences in program offer acceptance rates presented in Section 2.4. Third, we present a robustness to evaluate how our program-level estimates are affected by the non-event comparison groups included in our regression analysis.

C.3.1 Intention-to-Treat

In our main analysis, we estimate the average treatment effects to peak events by program for households that accepted the demand response program offers. In this section, we estimate an estimator similar to an intention-to-treat (ITT) estimate.

Our treatment (events) are randomized within demand response programs for households that participated in our experiment. Therefore, unlike standard ITT specifications, we cannot have a binary treatment indicator that turns on for all households assigned to a specific demand response program, regardless of whether they accept that program. Because we have periodic events that are randomly assigned to those households that chose to enroll in a demand response program, we create an analogous environment in our setting for all households invited to each program. To do this, we assign households that were randomized to receive the Central, Tech, and Manual program offers but did not accept the offer a distribution of randomized "synthetic" events that is the same as the households that participated. For these households, this creates a new variable $\widehat{E}_{it}$ that includes the synthetic events for households that were invited to the Central, Tech, or Manual programs but did not accept our offer, and the observed events for households in our experiment (i.e., E_{it} in Equation (1)).

Using the full sample of households (including those that did and did not accept demand response program offers), we estimate the following equation separately for each demand response program and including the never-treated households that were not subject to events:[58]

[58] We run two additional specifications of this analysis where we only include households within each demand response program, and all demand response programs and the never-treated simultaneously. Our ITT results are robust to these alternative specifications.

$$ln(c_{it}) \;=\; \beta_{ITT}\,\text{Assigned}_i \cdot \widehat{E}_{it} + \gamma_i\,\text{Assigned}_i \cdot \widehat{Spillover}_{it} + \alpha_i + \tau_t + \delta X_t + \varepsilon_{it} \quad (5)$$

where Assigned_i is an indicator variable that equals one if the household was assigned to receive a program offer (i.e., the Central, Tech, or Manual offer) and zero otherwise. $\widehat{E}_{it}$ is the household-specific event indicator that equals one if household i experiences an event in hour t (or is assigned a synthetic event, for households that did not accept the program offer) and zero otherwise. $\widehat{Spillover}_{it}$ is a vector that includes pre- and post-event indicator variables that equal one if hour t is two hours prior to and after a peak event (or synthetic events), and zero otherwise. Similar to our main analysis, we include fixed effects at the household and hour-of-sample levels; X_t includes the same vector of hourly weather controls. We cluster standard errors at the household level.

We also consider a modified version of equation (5) that permits event type-specific treatment effects. More specifically, $\widehat{E}_{it}$ is now a household-specific set of indicator variables that denote whether the household is was assigned a morning, evening, or high evening event. For households that did not enter into our demand response programs, the (synthetic) allocation of the three event types is randomized with the same frequency distribution as for households that did accept the program offers and receive the event treatments. In this extension, we have separate pre- and post-event spillover control variables for the hours surrounding morning and evening event (or synthetic event) windows. All other details of the regression specification remain the same.

Table C7 provides the ITT estimates. The coefficients imply that offering households the opportunity to join the Central, Tech, and Manual demand response programs results in an average reduction in consumption of -13.7%, -3.7%, and -2.8% during events, respectively, from each program of invited households.[59] As expected, the results are attenuated downward compared to our main estimates as households that did not join the demand response programs were unaware of the peak events. These percent changes are what an electric utility company can expect "at the end of the day" in terms of event-time consumption reductions after both making offers such as these and conducting event-time alerts. That is, these estimates include both the extensive margin of program offer acceptance and intensive margin of household

[59]These calculations reflect the relevant coefficient estimate $\widehat{\beta}_{ITT}$, transformed to percentage change in consumption as follows: $100 * (\,exp(\widehat{\beta}_{ITT}) - 1\,)$.

electricity consumption reductions from those who accepted the offer.

Table C7. ITT Estimates by Program

	Central	Tech	Manual
Assigned × Event	-0.1475	-0.0376	-0.0286
	(0.0139)	(0.0102)	(0.0089)

Notes. The reported results are program-specific ITT coefficients from equation (5). For each demand response program, the sample includes households from their own treatment program and the never-treated groups. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

These ITT results, like our main event-level treatment effect results, indicate centralized electricity consumption reductions result in an expected average reduction in demand during events that is several orders of magnitude larger than non-centralized programs, without or without the same technology. The ITT estimates here show uniquely that the relative advantage of centralized demand is still realized, even when one considers the differential take-up rates across programs. In addition, the Tech and Manual programs have ITT estimates that are similar in magnitude, consistent with the findings in our main treatment effect estimates.

Table C8 reports the ITT estimates by event type. The findings parallel the qualitative conclusions from our main average treatment effects on the households that participated in our experiment, with smaller estimated effects due to the ITT including the extensive acceptance margin, as described above. The estimated ITT coefficients represent an average reductions in consumption during events ranging from approximately 12% to 16% for the Central program, 3% to 4% for the Tech program, and 2% to 4% for the Manual program, respectively. We continue to find modest differences across event types for each demand response program. The precision of estimates and significance declines for the Manual program; the high peak event estimate is no longer significant. Further, for each program, we do not observe any evidence of a significantly higher response to increased financial incentives during High Evening events.

Table C8. ITT Estimates by Program and Event Type

	Central	Tech	Manual
Assigned × Morning Event	-0.1710	-0.0293	-0.0432
	(0.0155)	(0.0127)	(0.0116)
Assigned × Evening Event	-0.1327	-0.0390	-0.0233
	(0.0152)	(0.0125)	(0.0110)
Assigned × High Evening Event	-0.1450	-0.0457	-0.0198
	(0.0167)	(0.0136)	(0.0124)

Notes. The reported results are program-specific ITT coefficients from equation (5), allowing for event type-specific estimates and event-type specific spillover controls. For each demand response program, the sample includes households from their own treatment program and the never-treated. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

C.3.2 Bounds Analysis

As discussed in Section 4.4, we carry out a bounds analysis to serve as a lower-bound estimate, particularly to investigate whether the higher response for the Central program can be explained by concerns over differences in program offer acceptance rates.

In this analysis, we assume that households are (weakly) more likely to accept a program offer that involves less intervention by the utility. This is consistent with the monotonic ranking observed in Table 2, where the percentage of acceptance into a program that receives peak events is highest in Manual, followed by Tech, and then Central. We take this ranking to imply that if a household accepted the Central program offer, they would have been willing to accept the Tech and Manual offer, and similarly if the household accepted the Tech offer they would have accepted a Manual program offer.

To illustrate, focusing on the comparison between Central and Manual and using the acceptance rates in Table 2, if we offered 100 households to join the Manual or Central programs, 59 and 42 would have accepted, respectively. Under our assumptions, this example indicates that 17 more households would have joined Manual than Central. As a bounding exercise, we compute the average response to peak events for the Central program, assuming the additional 17 households that would have joined the Manual programs but not Central joined the Central program, but their response to peak events was a 0% reduction in demand. This approach takes the conservative assumption that the households that did not join the Central program, but would have joined the other program(s), are different based on unobservable factors that leads them to not respond to peak events. As we see in Figure 3, we find an average demand reduction to peak events among the households in the Tech and Manual programs. As a result, it is reasonable to assume that a null response assumption will serve as a lower bound for these additional households.

We carry out a similar bounding analysis comparing Central to Tech and Tech to Manual to account for the difference in acceptance rates across programs, and bound our estimated treatment effects of peak events among those that joined the experiment. More formally, for each program $j \in \{C, T, M\}$, define N_j to be the number of offers made, A_j to be the number of offers accepted, and $\widehat{\beta}_j$ to be the estimated program-level response to peak events reported in Figure 3. The lower bound estimate for program j relative to program k is as follows:

$$\underline{DR}_{j,k} = \frac{\widehat{\beta}_j \overbrace{N_j \left(\frac{A_j}{N_j}\right)}^{\text{HH in } j} + 0 \quad \overbrace{N_j \left(\frac{A_k}{N_k} - \frac{A_j}{N_j}\right)}^{\text{Additional HH in } j \text{ with } k\text{'s Higher Acceptance Rate}}}{\underbrace{N_j \left(\frac{A_j}{N_j}\right)}_{\text{HH in j}} + \underbrace{N_j \left(\frac{A_k}{N_k} - \frac{A_j}{N_j}\right)}_{\text{Additional HH in } j \text{ with } k\text{'s Higher Acceptance Rate}}} \tag{6}$$

where we consider three cases: (i) $j = C$ and $k = T$, (ii) $j = C$ and $k = M$, and (iii) $j = T$ and $k = M$.

Table C9 provides lower-bound estimates on the percentage reduction in peak events by demand response program for the Central program, adjusting by the higher acceptance rates for Tech and Manual, and the Tech program, adjusting by the higher acceptance rate for Manual. As expected, the lower bound calculation reduces the estimated treated effects. However, our key conclusions persist. We continue to observe an average estimated response for the Central program that is considerably larger than that in the Tech and Manual programs. Similarly, the estimated response of the Tech program continues to be in the same range as the Manual program, even adjusting our estimate downward by the proportion of additional households that would have accepted the Manual program offer.

Table C9. Lower Bound Program-Level Average Treatment Effects for Central and Tech (Percentage Change in Peak Event Demand)

	Central	Tech
Tech-Adjusted	-23.18	–
Manual-Adjusted	-19.32	-3.96

Notes. The reported results reflect our average estimated treatment effects of program participants in Equation (1), adjusted downward using our lower-bounding exercise detailed in Equation (6). A lower bound exercise is implemented for Central, relative to the higher Tech and Manual acceptance rates, while a lower bound calculation is only done for Tech relative to Manual. The lower bound estimates are translated into percentage change as follows: $100 \times (\exp(\underline{DR}_{j,k}) - 1)$.

C.3.3 Comparison Group in Program-Level Regression

In our main specification in Equation (1), our analysis includes all three demand response programs and never-treated households. As discussed in Section 4.3, this approach compares event time consumption to non-event time consumption of households in the same demand response program, other demand response programs, and never-treated households. One may be concerned that the programs are differentially selected, and this impacts our ability to compare event-time to non-event-time consumption decisions across programs, even after including our various control variables.

Tables C10 and C11 present our estimated treatment effects of participants by program to events and separated by event type, respectively, allowing for regressions only including households in the same demand response program (Column (1)), same demand response program and only the Control housholds (Column (2)), same demand response program and only the Info housholds (Column (3)), same demand response program and all never-treated households (Column (4)), and the results from our main specification with all groups in Column (5) for comparison purposes. Our results are consistent across all specifications indicating that the comparison group does not change our key conclusions.

Table C10. Treatment Effects of Participants by Program

Program	(1)	(2)	(3)	(4)	(5)
Central	-0.3151	-0.297	-0.3100	-0.2997	-0.3036
	(0.0206)	(0.0202)	(0.0206)	(0.0204)	(0.0204)
Tech	-0.0661	-0.0457	-0.0584	-0.0474	-0.0496
	(0.0132)	(0.0147)	(0.0152)	(0.0155)	(0.0160)
Manual	-0.0507	-0.0418	-0.0528	-0.0459	-0.0542
	(0.0092)	(0.0108)	(0.0113)	(0.0117)	(0.0123)
Comparisons					
Own Program	Y	Y	Y	Y	Y
Other Treated					Y
Never Treated-Control		Y		Y	Y
Never Treated-Info			Y	Y	Y

Notes. The reported results are program-specific treatment effect coefficients. Standard errors are reported in the parentheses and clustered at the household level. Column (1) reports the regression results including only within demand response program comparisons, column (2) includes both within demand response program and the Control households, column (3) includes both within demand response program and the Info households, column (4) includes both within demand response program and the all never-treated households, and column (5) reports results when all programs/groups are included. All specifications include fixed effects at the household and hour-of-sample levels.

Table C11. Treatment Effects of Participants by Program and Event-Type

Program	(1)	(2)	(3)	(4)	(5)
Central					
Morning	-0.3198	-0.3138	-0.3186	-0.315	-0.3122
	(0.0221)	(0.0223)	(0.0227)	(0.0224)	(0.0223)
Evening	-0.3040	-0.2806	-0.2971	-0.2840	-0.2905
	(0.0218)	(0.0219)	(0.0229)	(0.0226)	(0.0228)
High Evening	-0.3291	-0.3057	-0.3228	-0.3095	-0.3166
	(0.0237)	(0.0237)	(0.0244)	(0.0241)	(0.0243)
Tech					
Morning	-0.0495	-0.0299	-0.0357	-0.0272	-0.0159
	(0.0128)	(0.0173)	(0.0176)	(0.0188)	(0.0199)
Evening	-0.0713	-0.0506	-0.0665	-0.0544	-0.0625
	(0.0155)	(0.0170)	(0.0181)	(0.0181)	(0.0185)
High Evening	-0.0786	-0.0569	-0.0727	-0.0603	-0.0677
	(0.0166)	(0.0188)	(0.0197)	(0.0200)	(0.0206)
Manual					
Morning	-0.0488	-0.0614	-0.0651	-0.0689	-0.0815
	(0.0097)	(0.0136)	(0.0140)	(0.0150)	(0.0159)
Evening	-0.0488	-0.0304	-0.0446	-0.0331	-0.0399
	(0.0108)	(0.0129)	(0.0139)	(0.0143)	(0.0149)
High Evening	-0.0567	-0.0377	-0.0519	-0.0402	-0.0470
	(0.0126)	(0.0147)	(0.0155)	(0.0159)	(0.0166)
Comparisons					
Own Program	Y	Y	Y	Y	Y
Other Treated					Y
Never Treated-Control		Y		Y	Y
Never Treated-Info			Y	Y	Y

Notes. The reported results are program-specific treatment effect coefficients by event type. Standard errors are reported in the parentheses and clustered at the household level. Column (1) reports the regression results including only within demand response program comparisons, column (2) includes both within demand response program and the Control households, column (3) includes both within demand response program and the Info households, column (4) includes both within demand response program and the all never-treated households, and column (5) reports results when all programs/groups are included. All specifications include fixed effects at the household and hour-of-sample levels.

C.4 Multiple Hypothesis Testing

Our empirical analysis presents an array of null hypotheses that are tested simultaneously, hereon referred to as Multiple Hypothesis Testing (MHT). There is a literature that raises concerns that MHT can lead to a higher probability of a false rejection of a null because of the large number of tests being performed (Benjamini et al., 2006; Anderson, 2008; List et al., 2019, 2023). In this Appendix, we carry out several robustness checks to evaluate if our key results and conclusions are robust to address concerns over MHT. In particular, we implement a combination of the False Discovery Rate (FDR) approach used to compute Sharpened Q-values (Benjamini et al., 2006; Anderson, 2008) and the more conservative Familywise Error Rate (FWER) using the Romano–Wolf procedure in (List et al., 2019, 2023).

We follow the framework outlined in List et al. (2023) and segment our core empirical analysis into "families" of null hypotheses. For our program-level regression analysis, we consider the most complete family that includes 24 null hypotheses that simultaneously evaluates: (i) whether there is evidence of a significant response to peak events by program relative to non-event days [3], (ii) whether these responses are different across our three programs: Central, Tech, and Manual [3 pair-wise comparison], (iii) whether there is evidence of a significant response for each of the three different peak event types by program relative to non-event days [9], and (iv) if there are differences across event-type responses within a given program [9]. We implement both the FDR and FWER adjustments for this analysis.[60]

For our household-level regression analysis, we define our family of null hypotheses to be the set that evaluates if there is a significant response to peak events compared to non-event days, simultaneously considering all households in the Central, Tech, and Manual programs. Due to limitations associated with the computational complexity of the FWER, we are unable to carry out a joint MHT adjustment using the FWER for the over 600 household-level regressions and their corresponding null hypotheses tests. We leverage the FDR approach to address concerns of MHT in this setting. While this is less conservative than the FWER approach, we view this as a reasonable trade-off for the household-level regressions as policy decisions are unlikely to be made based on the share of households with a significant response. Rather, they are more likely to be informed by the headline program-level average treatment effects.

[60] We use code provided by Anderson's (2008) analysis to compute sharpened False Discovery Rate (FDR) q-values. For the FWER MHT adjustment, we the STATA command rwolf2 that computes Romano-Wolf stepdown adjusted p-values and permits clustering and high dimensional fixed effects.

We extend this analysis by considering a second family of null hypotheses where we allow households to have a differential response based on whether or not they interacted with the App (recall Section 6.2). We consider this broader set of null hypotheses to evaluate if there is a response to peak events on days the household did or did not interact with the App.

Table C12. Program-Level Multiple Hypothesis Testing Adjustments

Comparison	Model P-Value	FDR (Sharpened Q)	FWER
Program-Event			
Central Event vs. No Event	0.000	0.001	0.001
Tech Event versus No Event	0.002	0.003	0.001
Manual Event versus No Event	0.000	0.001	0.001
Central Event vs. Manual Event	0.000	0.001	0.001
Tech Event vs. Manual Event	0.827	0.267	0.791
Central Event vs. Tech Event	0.000	0.001	0.001
Program-Event-Type			
Central Morning vs. No Event	0.000	0.001	0.001
Central Evening vs. No Event	0.000	0.001	0.001
Central High Evening vs. No Event	0.000	0.001	0.001
Tech Morning vs. No Event	0.424	0.146	0.479
Tech Evening vs. No Event	0.001	0.002	0.001
Tech High Evening vs. No Event	0.001	0.002	0.001
Manual Morning vs. No Event	0.000	0.001	0.001
Manual Evening vs. No Event	0.008	0.008	0.001
Manual High Evening vs. No Event	0.005	0.006	0.001
Central Morning vs. Central Evening	0.299	0.124	0.252
Central Morning vs. Central High Evening	0.841	0.267	0.791
Central Evening vs. Central High Evening	0.032	0.024	0.005
Tech Morning vs. Tech Evening	0.037	0.025	0.005
Tech Morning vs. Tech High Evening	0.026	0.021	0.002
Tech Evening vs. Tech High Evening	0.660	0.220	0.720
Manual Morning vs. Manual Evening	0.034	0.024	0.005
Manual Morning vs. Manual High Evening	0.084	0.049	0.022
Manual Evening vs. Manual High Evening	0.503	0.168	0.618

Notes. Comparison represents the null hypotheses being tested across demand response programs and events/non-event hours. The Model p-value reports the p-values on the relevant comparisons using the regression model specified in Equation (1) that are not adjusted for MHT. FDR represents the Sharpened Q-value False Discovery Rate using the procedure outlined in Benjamini et al. (2006). The FWER is implemented using the rwolf2 command in STATA with 1,000 bootstraps. Standard errors are clustered at the household-level.

Table C12 presents the results of our MHT-adjustments for the program-level

hypotheses. There are 24 hypothesis tests separated by two subfamilies: (i) Program-Event tests and (ii) Program-Event-Type. The Model p-value represents our non-MHT-adjusted p-values, the FDR represents Sharpened Q-value FDR, and is the FWER p-values. These results demonstrate that our results are robust to both the FDR and FWER MHT adjustments. The Central program continues to be highly significant and different from Tech and Manual, while Tech and Manual are not statistically significantly different from each other (but they have significant non-zero responses). The event-type specific results yield the same broad general conclusions outlined in Section 5.1.

Table C13 presents the percentage of the estimated household-level treatment effects that are significant and negative at the 5% level using the unadjusted model p-values and the adjusted FDR sharpened q-values. The results demonstrate that our key conclusions are robust to the FDR MHT adjustment. Focusing first on Family 1, we find that a considerably larger proportion of households show a significant response to peak events compared to Tech and Manual. Under the FDR-adjustment, we only observe a modest reduction in the percentage of household-specific estimates that are significant.

Moving to Family 2, we observe fewer household-specific coefficients that are significant in the Central program. Further, there is little difference in the proportion of coefficients that are significant with or without App interactions. Taken together, this suggests that App interactions play a modest role in the likelihood of whether or not a household responds to peak events in the Central program. Rather, the inclusion of the App Interaction increases the noise in our household-specific peak event estimates. The FDR MHT adjustment lowers the proportion of household coefficients that are significant, but only by a small margin.

Moving to Tech and Manual, we continue to observe a smaller proportion of households with significant responses both with and without App interactions. However, unlike Central, we find a larger increase in the proportion of households with significant responses by whether or not the household interacted with the App on a peak event day. Despite the higher proportion of significant results when a household interacts with the App, we continue to find a sizable difference in the proportion of significant responses when comparing the Central program to the Tech and Manual programs. The FDR MHT adjustment lowers the proportion of significant responses by a few percentage points, but the key qualitative conclusions are robust.

Table C13. Percentage of Significant Household-Level Estimated Treatment Effects - Multiple Hypothesis Testing FDR Adjustment

Comparison	Model	FDR (Sharpened Q)
Family 1: Event vs. Non-Event		
Central	0.80	0.79
Tech	0.20	0.17
Manual	0.18	0.16
Family 2: Event vs. Non-Event by App Interact		
Central		
No App Interact	0.68	0.66
App Interact	0.68	0.64
Tech		
No App Interact	0.11	0.10
App Interact	0.23	0.20
Manual		
No App Interact	0.12	0.10
App Interact	0.26	0.21

Notes. Comparison represents the null hypotheses being tested across demand response programs and events/non-event hours or across event types. Family 1 includes the null hypothesis for event versus non-event comparisons. Family 2 considers the same comparisons, but permits different effects by whether or not the household interacted with the App. Model reports the proportion of households that had a significant response to peak events (with a p-value ≤ 0.05) using the regression model specified in Equation (2) that are not adjusted for MHTs. FDR represents the Sharpened Q-value False Discovery Rate using the procedure outlined in Benjamini et al. (2006). The FDR is computed for each family of tests separately. Standard errors are clustered at the household-level.

C.5 Variance Reduction

In this Appendix, we use the approach developed by List et al. (2024) to evaluate the impact of increasing precision (i.e., reducing the variance) on our estimated peak event treatment effects at the program-level. To perform the regression adjustment, we use pre-treatment information on hourly household-level consumption for five months prior to treatment and household-level responses to a survey at the invitation phase of our experiment that reported appliance-level information. Using this information, we construct measures of the average daily load factor, maximum hourly consumption, and total consumption pre-treatment, as well as variables denoting if a household had an electric vehicle, electric baseboard heating, electric water heater, and air conditioning.

Tables C14 and C15 report the estimated effect of peak events across all event types and separately by event type, respectively, using these variance-reduction techniques. Column (1) reports a difference-in-sample means (SM) focusing only on post-treatment data and including no covariate adjustments, and Columns (2) and (3) include the regression adjustments. LRA denotes a linear regression adjustment and FRA denotes a flexible regression adjustment where we use a random forest algorithm to estimate the conditional expectation function.

Consistent with findings in List et al. (2024), the LRA and FRA reduce the standard error compared to the SM specification, while the LRA and FRA yield similar standard error levels. In all three approaches, our results yield analogous conclusions and levels of significance when compared to our main analysis. For the event type-specific results, the increased precision leads to significant effects for each program and event type. Compared to findings in Figure 4, the morning event effect is now significant for the Tech program. While we have increased precision in our estimates, the conclusions from our analysis are consistent with our main specifications.

Table C14. Variance Reduction by Program

	SM	LRA	FRA
Central	-0.3097	-0.3087	-0.308
	(0.0057)	(0.0011)	(0.0011)
Tech	-0.0552	-0.055	-0.055
	(0.0054)	(0.0011)	(0.0011)
Manual	-0.0465	-0.0468	-0.0468
	(0.0048)	(0.001)	(0.001)

Notes. The reported results are program-specific treatment effect using the variance reduction techniques in List et al. (2024). The analyses focus on post-treatment data and include pre-treatment covariates constructed using the 5 months of data between September 2021 - January 2022: average daily load factor, maximum hourly consumption, total hourly consumption, and variables denoting if the household reported having an electric vehicle, electric baseboard heating, air conditioning, or an electric hot water heater. Column (1) reports simple difference in sample means, column (2) denotes the linear regression adjustment (LRA), and column (3) denotes a flexible regression adjustment (FRA) where we use a random forest algorithm to estimate the conditional expectation function.

Table C15. Variance Reduction by Program and Event Type

	SM	LRA	FRA
Central			
Morning	-0.3054	-0.3061	-0.3045
	(0.0101)	(0.0015)	(0.0015)
Evening	-0.2965	-0.2949	-0.2952
	(0.0086)	(0.0017)	(0.0016)
High Evening	-0.3258	-0.3237	-0.3223
	(0.0115)	(0.0017)	(0.0016)
Tech			
Morning	-0.0204	-0.0203	-0.0198
	(0.0097)	(0.0015)	(0.0014)
Evening	-0.059	-0.0594	-0.0612
	(0.008)	(0.0016)	(0.0015)
High Evening	-0.0712	-0.0701	-0.0733
	(0.0105)	(0.0015)	(0.0015)
Manual			
Morning	-0.0367	-0.0365	-0.0365
	(0.0087)	(0.0013)	(0.0012)
Evening	-0.0371	-0.038	-0.0382
	(0.0071)	(0.0014)	(0.0013)
High Evening	-0.0477	-0.0484	-0.0499
	(0.0094)	(0.0014)	(0.0013)

Notes. The reported results are program-specific treatment effects by program type using the variance reduction techniques in List et al. (2024). The analyses focus on post-treatment data and include pre-treatment covariates constructed using the 5 months of data between September 2021 - January 2022: average daily load factor, maximum hourly consumption, total hourly consumption, and variables denoting if the household reported having an electric vehicle, electric baseboard heating, air conditioning, or an electric hot water heater. Column (1) reports simple difference in sample means, column (2) denotes the linear regression adjustment (LRA), and column (3) denotes a flexible regression adjustment (FRA) where we use a random forest algorithm to estimate the conditional expectation function.

C.6 Fixed Effects Robustness

In this Section, we present our program-level and household-level regression results with alternative fixed effects. First, we begin by considering a range of fixed effects for our program-level regression in Equation (1). We begin with only including fixed effects at the year-month, hour-of-day, and household-level in column (1), followed by our main specification in column (2) that considers more granular calendar fixed effects at the hour-of-sample. Columns (3) - (5) add on more granular fixed effects at the household-hour-of-day, household-day-of-week, or household-year-month. These latter specifications help absorb household-specific schedules at various time scales.

Tables C16 and C17 presents results using our peak event indicator variable and event-type-specific effects, respectively. While the precise estimates can differ across specifications, our conclusions are robust across these fixed effect combinations.

Second, our household-level analysis in Equation (2) includes fixed effects at the day-of-week, hour-of-day, and year-month to absorb household-specific factors that impact consumption decisions. While we are unable to include hour-of-sample fixed effects in this specification, we consider a specification with more granular calendar fixed effects at the year-month-hour-of-day. This absorbs household-specific schedules that can vary over the day.

Figure C6 presents the distribution of the household treatment effects as a percentage change in demand during peak events. These results closely reflect the distributions presented in our main specification in Figure 6. We continue to observe Central having a much higher (negative) response, while Tech and Manual are closely concentrated around zero.

We observe a slight reduction in the average estimated household-level treatment effects for Central (-24% to -23.3%), Tech (-4.6% to -3.9%), and Manual (-4% to -3.1%). However, even after absorbing more variation in the data with these calendar fixed effects, our key conclusions persist.

Table C16. Treatment Effects of Participants by Program - Fixed Effects Robustness

Program	(1)	(2)	(3)	(4)	(5)
Central	-0.2980	-0.3036	-0.3189	-0.3024	-0.3050
	(0.0202)	(0.0204)	(0.0204)	(0.02060	(0.0196)
Tech	-0.0442	-0.0496	-0.0636	-0.0515	-0.0479
	(0.016)	(0.016)	(0.0146)	(0.0160)	(0.0155)
Manual	-0.0490	-0.0542	-0.0517	-0.0541	-0.0560
	(0.0123)	(0.0123)	(0.0098)	(0.0123)	(0.0120)
Fixed Effects					
Household	Y	Y			
Year-Month	Y				
Day-of-Week	Y				
Hour-of-Day	Y				
Hour-of-Sample		Y	Y	Y	Y
Household-Hour-of-Day			Y		
Household-Day-of-Week				Y	
Household-Year-Month					Y

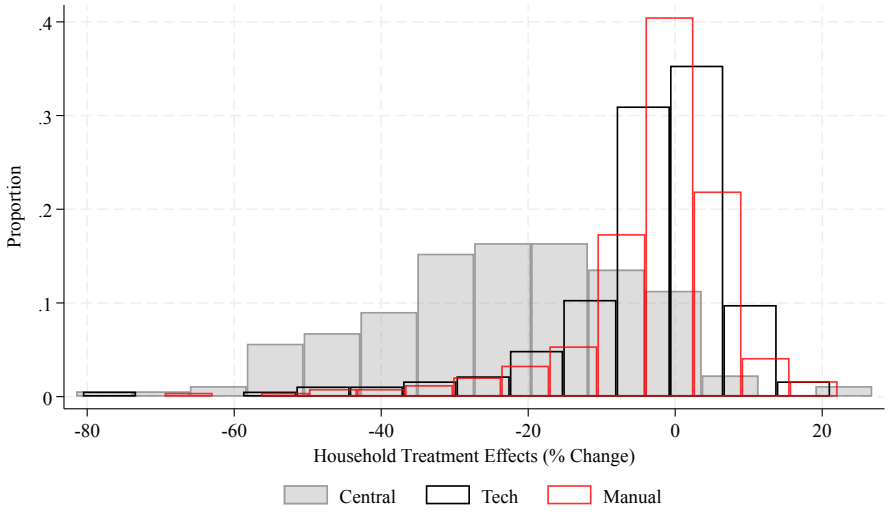
Notes. The reported results are program-specific treatment effect coefficients. Standard errors are reported in the parentheses and clustered at the household level. Columns (1) - (5) vary the fixed effects included in the analysis indicated in the bottom panel of the Table.

Table C17. Treatment Effects of Participants by Program and Event-Type - Fixed Effects Robustness

Program	(1)	(2)	(3)	(4)	(5)
Central					
Morning	-0.2974	-0.3122	-0.3253	-0.3119	-0.3131
	(0.0220)	(0.0223)	(0.0220)	(0.0222)	(0.0221)
Evening	-0.2896	-0.2905	-0.3069	-0.2893	-0.2944
	(0.0225)	(0.0228)	(0.0215)	(0.0231)	(0.0221)
High Evening	-0.3145	-0.3166	-0.3329	-0.3144	-0.3144
	(0.0242)	(0.0243)	(0.0236)	(0.0245)	(0.0233)
Tech					
Morning	0.0030	-0.0159	-0.0457	-0.0176	-0.0130
	(0.0201)	(0.0199)	(0.0151)	(0.0197)	(0.0200)
Evening	-0.0653	-0.0625	-0.0697	-0.0641	-0.0628
	(0.0187)	(0.0185)	(0.0167)	(0.0187)	(0.0181)
High Evening	-0.0640	-0.0677	-0.0754	-0.0702	-0.0636
	(0.0203)	(0.0206)	(0.0177)	(0.0207)	(0.0200)
Manual					
Morning	-0.0668	-0.0815	-0.0559	-0.0806	-0.0809
	(0.0158)	(0.0159)	(0.0114)	(0.0158)	(0.0160)
Evening	-0.0409	-0.0399	-0.0472	-0.0403	-0.0426
	(0.0150)	(0.0149)	(0.0112)	(0.0150)	(0.0146)
High Evening	-0.0424	-0.0470	-0.0546	-0.0467	-0.0497
	(0.0170)	(0.0166)	(0.0132)	(0.0166)	(0.0166)
Fixed Effects					
Household	Y	Y			
Year-Month	Y				
Day-of-Week	Y				
Hour-of-Day	Y				
Hour-of-Sample		Y	Y	Y	Y
Household-Hour-of-Day			Y		
Household-Day-of-Week				Y	
Household-Year-Month					Y

Notes. The reported results are program-specific treatment effect coefficients by event type. Standard errors are reported in the parentheses and clustered at the household level. Columns (1) - (5) vary the fixed effects included in the analysis indicated in the bottom panel of the Table.

Figure C6. Household-Level Estimated Treatment Effect Histograms by Program with Year-Month-Hour-of-Day Fixed Effects



Notes. The reported results summarize the distribution of estimated household-level treatment effects obtained from estimating specification (2), but including more granular calendar fixed effects at the day-of-week and year-by-month-by-hour-of-day. The histogram presents the proportion of observations that fall within a specific bin's range. We present the marginal effects on the x-axis to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$.

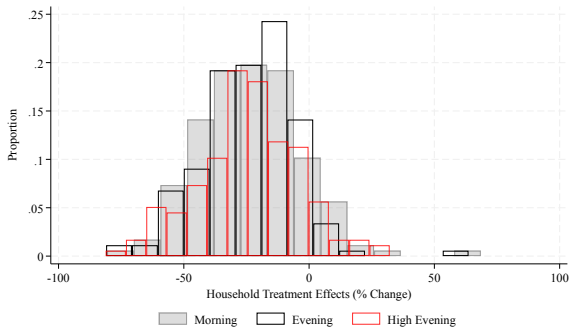
C.7 Household Treatment Effects by Event Type

Figure C7 presents the distributions of our household-level treatment effects by program and event type. More specifically, we estimate the specification detailed in (2) but permit different treatment effects by event type: morning, evening, and high evening.

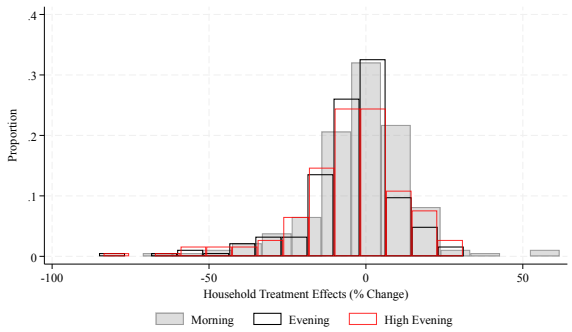
For the Central program, looking across all households, the average percentage reduction in consumption ranges from 23% during morning events to 24% in evening and high evening events. For Tech, this ranges from 1% during morning events to 5% during evening and high evening events. Finally, for Manual, the average reduction is 3% during morning events and 4% during evening and high evenings. While the precise values vary slightly, the average household-level treatment effects are in a similar range to those found in the program-level regressions summarized in Section 5.1. Importantly, we do not find any evidence of a systematically larger average reduction in consumption during high peak evening events compared to evening events.

A benefit of estimating household-level treatments effects is that we can observe the full distributions for evening and high evening events. Figure C7 shows that the Central program's distributions are quite similar by event type. For the Tech and Manual programs, there is some evidence to suggest that the response to evening events are more tightly distributed around zero compared to the high evening events. However, this does not translate into differences in average treatment effects when looking across all households, as documented above.

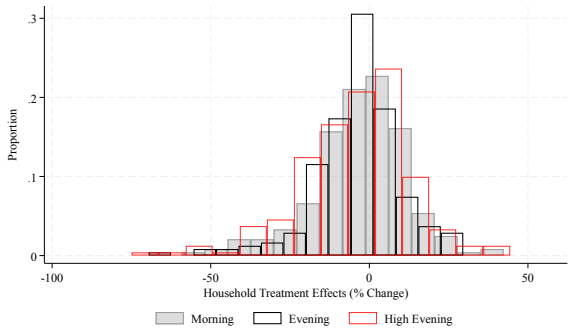
Figure C7. Household-Level Estimated Treatment Effect Distributions by Program and Event-Type



(a) Central Program



(b) Tech Program



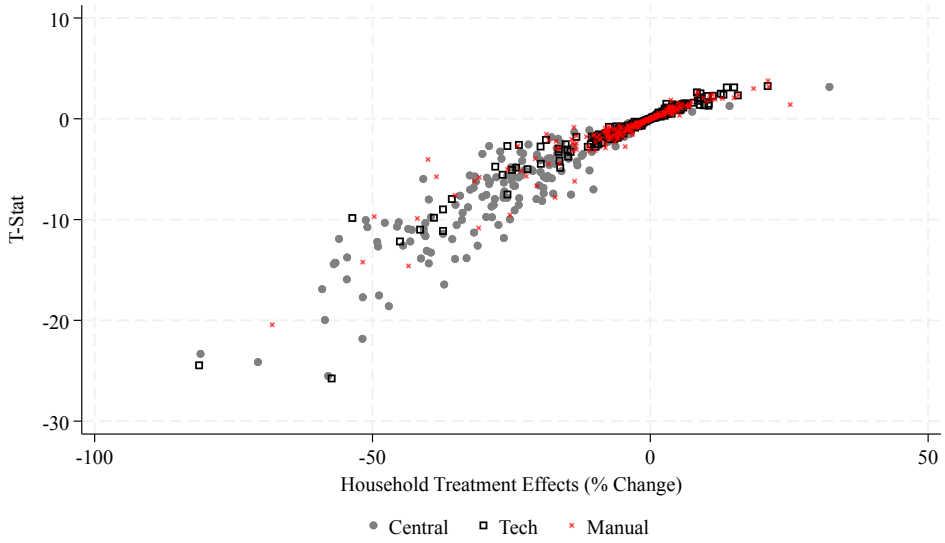
(c) Manual Program

Notes: The reported results summarize the distribution of estimated household-level treatment effects obtained from estimating specification (2), permitting different treatment effects by event type: morning, evening, and high evening. The histogram presents the proportion of observations that fall within a specific bin's range. We present the marginal effects on the x-axis to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$.

C.8 Consistency of Demand Response

In this section, we expand upon the household-level results to investigate the consistency of responses to events by program. This builds off of the discussion in Section 5.2 and Figure 7 in particular.

Figure C8. Household-Level Estimated Treatment Effects and T-Statistics by Program



Notes. The reported results plot the estimated household-level treatment effects and corresponding t-statistic obtained from estimating specification (2). We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$.

Figure C8 plots the estimated household event responses and the corresponding t-statistics to better understand the significance of the estimates. The majority of the Manual and Tech values are closely distributed near zero in both the estimated event response and t-statistics. Only 20% and 18% of the household-level estimated treatment effects are negative and significant at the 5% level for the Tech and Manual programs, respectively.[61] This suggests that only a small subset of high-performing households in these programs were consistent responders to peak events over our 17-month treatment period. Figure C8 reinforces our finding that the Central program has large and precisely estimated responses to peak events. 80% of the Central household-level estimated treatment effects are negative and significant, indicating

[61] There is a small subset of households with positive estimated treatment effects. These estimated effects are systematically statistically insignificant, with only 3% being positive and statistically significant.

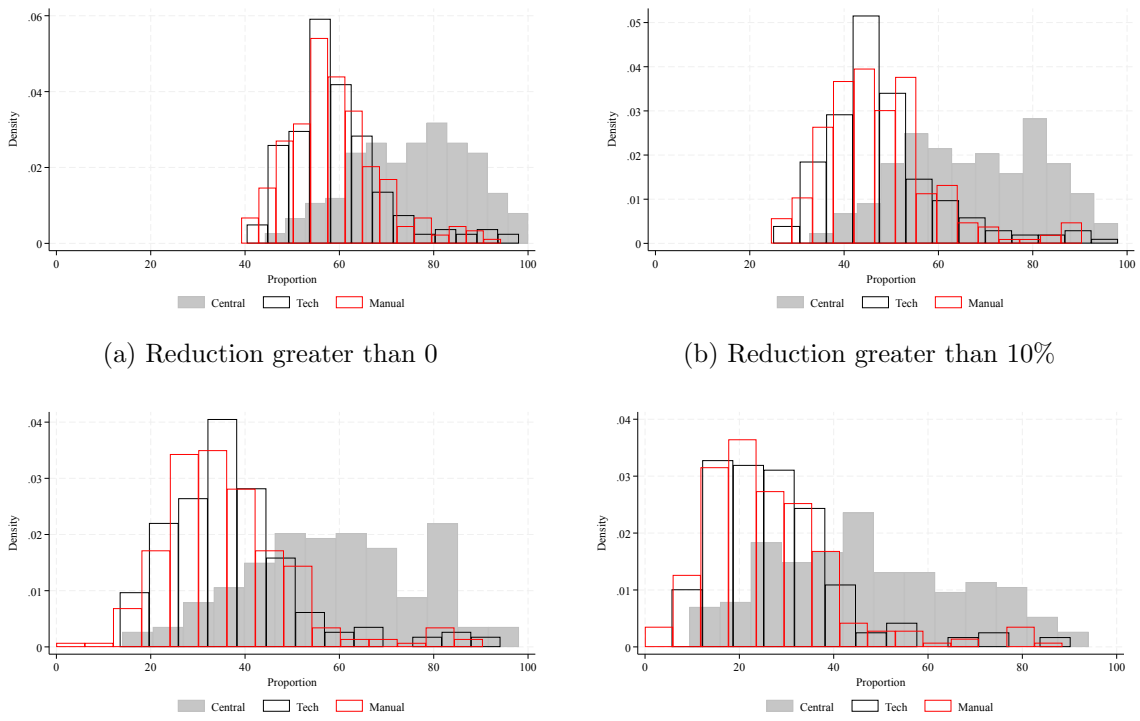
that the majority of Central households provide large and consistent reductions in consumption during peak events.

In addition to having estimates for household-level treatment effects, we observe each household's consumption relative to their baseline for *each* event. As described in Section 2.2, the baselines were calculated based on a household's average consumption during the relevant event time window over the last five weekdays prior to the event, excluding event days. For each household, we calculate the proportion of events where their observed percentage reduction in consumption relative to their baseline was greater than 0%, 10%, 20%, and 30%. It is important to recognize that this measure will have some noise. This is particularly relevant for the comparison of consumption relative to the baseline being greater than 0%. Even if a household did not adjust its behavior to respond to the event, natural variation would induce consumption to fall below the baseline for a certain proportion of events. If this randomness is normally distributed around the baseline, the proportion would be 50%. While this makes interpreting the results for this measure more nuanced, as we will describe below, it is still valuable because we can compare the relative performance across programs.

Figure C9 presents, at the household-level, the proportion of events where the observed percentage reduction in consumption relative to the baseline was greater than 0%, 10%, 20%, and 30% by program. Focusing initially on the results for greater than 0% threshold, the Tech and Manual programs have similar distributions and are slightly rightward shifted above 50%. As noted above, noise in this measure could result in values near 50% being viewed as having a minimal reduction relative to the baseline for the greater than 0% threshold. The fact that the Tech and Manual programs have distributions situated just to the right of 50%, with a small subset of households exceeding 70%, is consistent with our household-level regression results. In contrast, the majority of Central households have proportions in excess of 70% of event days being greater than 0%. This continues to suggest that Central households consistently reduce their consumption during events.

Figure C9 shows that as the threshold increases to greater than 10%, 20%, and 30% reductions relative to the baseline, the proportion of events where households achieve reductions above this threshold declines across all programs. However, the proportion of events where Tech and Manual achieve reductions above this threshold declines at a much faster rate than Central. For example, the majority of Central households achieve reductions greater than 20% in at least 50% of events. In contrast, this number is in the range of 20% - 30% of events for the majority of Tech and Manual

Figure C9. Household-Level Proportion of Responses with Reductions Relative to Baseline Greater Than 0%, 10%, 20%, and 30%



(a) Reduction greater than 0

(b) Reduction greater than 10%

(c) Reduction greater than 20%

(d) Reduction greater than 30%

Notes: Figure plots at the household-level the proportion of events where their observed percentage reduction in consumption relative to their baseline was greater than 0%, 10%, 20%, and 30% by program over the period February 1, 2022 - June 30, 2023.

households, with a small subset of higher performers who consistently reduce their consumption above this threshold in the majority of events.

Taken together, these results suggest that the Central program was able to achieve a high and consistent demand reductions for the majority of events over the 17-months they were exposed to peak events. This differs from the Tech and Manual programs that achieved more modest demand reductions for a smaller proportion of events.

C.9 Principal Stratification Robustness

In this Appendix, we begin by presenting descriptive information on the households that are deemed to be responders in Section 6.1. Responders in our baseline results are those with significant reductions in consumption during peak events at the 5% level. We then present results that consider alternative approaches to categorizing responders.

With Table C18, we see that, consistent with results in Section 6.2 Table 5, for the Tech and Manual programs, households that are deemed to be responders interact with the App at a significantly higher rate. In contrast, for the Central program, responder households interact with the App at a rate that is close to the broader program-level average. As would be expected, households that are defined as responders almost exclusively fall within the top two performance quartiles, and there is a considerably higher number of households in the top quartile among the Central program.

Table C18. App Interaction and Household-Level Estimated Treatment Effects of Responders

	Sample	App Interactions	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Central	Full Sample	0.27	111	44	15	7
	Responders	0.29	111	31	0	0
Tech	Full Sample	0.24	20	43	53	68
	Responders	0.44	20	16	0	0
Manual	Full Sample	0.19	20	64	83	75
	Responders	0.37	19	24	1	0

Notes. The reported results present the daily frequency of App Interactions by program separately for the Full Sample and for those households that are deemed to be responders. The quartiles represent the number of households that fall into each performance quartile, where performance is based on the estimated household-specific treatment effects obtained from Equation (2).

We consider alternative thresholds to define responders to evaluate the robustness of our conclusions to the classification used in the principal stratification framework. Our analysis in Section 6.1 categorized households as responders as a household whose estimated treatment effect during peak events is negative and significant at the 5% level. We present results that vary the significance threshold in the range $\{0.01, 0.025, 0.05, 0.075, 0.10\}$. As an alternative, we define a responder based on the magnitude of their household-level estimated treatment effect. We consider a range of thresholds where a household is classified as a responder if their estimated percentage peak event demand reduction is at least 2.5%, 5%, 7.5%, 10%, or 15%.

Table C19. Responder Shares and Average Peak Demand Reduction Among Responders - P-Value Robustness

	Central	Tech	Manual
p = 0.01			
Share of Responders	0.76	0.15	0.13
Household Treatment Effect (%)	-32.13	-31.18	-26.94
p = 0.025			
Share of Responders	0.79	0.17	0.15
Household Treatment Effect (%)	-31.67	-28.94	-24.9
p = 0.05			
Share of Responders	0.80	0.20	0.18
Household Treatment Effect (%)	-31.29	-27.11	-22.93
p = 0.075			
Share of Responders	0.83	0.22	0.21
Household Treatment Effect (%)	-30.65	-25.46	-21.19
p = 0.10			
Share of Responders	0.84	0.26	0.24
Household Treatment Effect (%)	-30.36	-22.83	-19.53

Notes. The reported results present the average household-level estimated treatment effects from estimating specification (2), focusing only on households deemed to be responders. We vary the definition of responders whose coefficients are negative and significant at the p-value $0.01, 0.025, 0.05, 0.075$, and 0.10 levels. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$. Share of responders represents the percentage of households within a demand response program that are classified as responders.

Table C19 summarizes the results using the range of significance thresholds. As expected, a less (more) stringent threshold increases (decreases) the proportion of responders across all three programs. We continue to see a weakly monotonic relationship in the share of responders as we move from Manual to Tech and Tech to Central, with only a small increase in the share for Manual to Tech. There is a considerable increase in the share as we move from Tech to Central. Further, the average household level treatment effect (% reduction in demand) among households that are categorized as Responders is largest for Central followed by Tech and then Manual. The effects remain significantly different for Central compared to Tech and Manual for all cases except for the most stringent definition with $p = 0.01$ where Central and Tech are not significantly different. One possible explanation for this result could be that with this high threshold of significance, we are isolating the small number of highest performers among the Tech program who are willing to put a high degree of effort, or possibly utilizing the installed load controllers.

Table C20. Responder Shares and Average Peak Demand Reduction Among Responders - % Reduction Threshold

	Central	Tech	Manual
Reduction 2.5%			
Share of Responders	0.93	0.47	0.45
Household Treatment Effect (%)	-28.4	-15.35	-13.46
Reduction 5%			
Share of Responders	0.87	0.34	0.34
Household Treatment Effect (%)	-29.75	-19.2	-16.21
Reduction 7.5%			
Share of Responders	0.82	0.24	0.24
Household Treatment Effect (%)	-30.87	-24.27	-19.77
Reduction 10%			
Share of Responders	0.78	0.18	0.17
Household Treatment Effect (%)	-31.95	-28.79	-24.94
Reduction 15%			
Share of Responders	0.71	0.14	0.11
Household Treatment Effect (%)	-33.73	-33.33	-30.83

Notes. The reported results present the average household-level estimated treatment effects from estimating specification (2), focusing only on households deemed to be responders. We vary the definition of responders whose coefficients are negative and achieve percentage reductions at or above thresholds ranging from 2.5%, 5%, 7.5%, 10%, to 15%. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_i) - 1)$. Share of responders represents the percentage of households within a demand response program that are classified as responders.

Table C20 summarizes the results using the percentage reduction thresholds to define responder households. Despite differences in the scales, these results continue to reflect the conclusions reported in Section 6.1. The difference in the share of responders remains considerably higher for Central compared to Tech and Manual, and the average household treatment effect is the highest for Central program households among those deemed responders. The latter difference is significantly different for all cases except for the most stringent threshold of a 15% reduction.

Taken together, both robustness tables continue to suggest that the Central outperforms the Tech and Manual programs along both margins: Substantially more households exhibit a response to peak events, and the households that do respond under Central reduce demand by more on average. These results suggest that the former extensive-response-margin effect is the key driver.

C.10 Additional Device-Specific Results

Table C21 reports the number of households that had load-controllers installed on thermostats (binary), hot water heaters, and electric vehicle level 2 chargers for our Central and Tech programs. The table also presents the count of thermostats to capture the fact that an average household in our sample with an electric baseboard unit had 4 thermostat devices. We can see that a large portion of households in our experiment had load controllers installed on thermostats and/or their electric hot water heater. Only a small share of participants had load-controllers installed on level 2 EV charges, likely driven in part by the relatively low penetration of EVs in 2021-2022. The majority of households in both programs had only one device, and when a household had multiple load-controllable devices, the majority of these households had an electric hot water heater and electric baseboard thermostats.

Table C21. Installed Load Controllable Device Counts - Central and Tech Programs

	Households (Count)	Thermostats (Binary)	Thermostat (Count)	Water Heaters	EV
Central	177	112	469	127	23
Tech	184	116	468	132	16

Notes. This table reports the number of installed load-controllable devices by technology. Thermostats (Binary) report the number of households that have a load-controllable thermostat, while Thermostats (Count) is the count of controllable thermostats as households often have multiple thermostats.

Figures C10 - C12 present the hourly electricity usage data for each of the three device types for Central and Tech households, separated by days where there was no event, a morning event, an evening event, and a high peak evening event. The shaded regions in the graph denote the hours with peak events in the morning or evening windows.

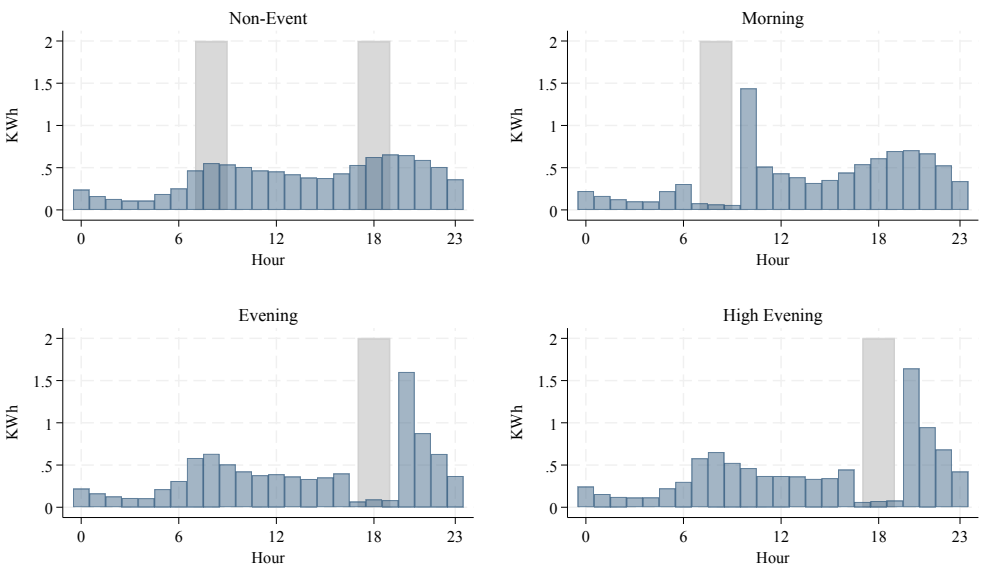
We can see a near zero electricity usage during peak events across all types for the Central program for the hot water heater and EV level 2 charger. This is followed by a rapid snap-back effect as the load control ends at the end of an event. For the thermostats, we observe a reduction during events, but the effect is smaller compared to the other devices, and the magnitude of reduction tends to decrease over the duration of the event. This likely reflects the fact that the utility-managed thermostats responded to a peak event by reducing the set point by up to three degrees (C), and once the new set point was reached, the devices turned back on to maintain this new

set point. In contrast, for the households in the Tech program, we see limited evidence of a change in device usage for the hot water heaters or smart thermostats. For the small number of households with load controllers on EVs, there is some suggestive evidence of a response, but it is much smaller in magnitude than that observed for the Central households. These results are consistent with the estimated device-specific average peak event responses reported in Section 6.3.

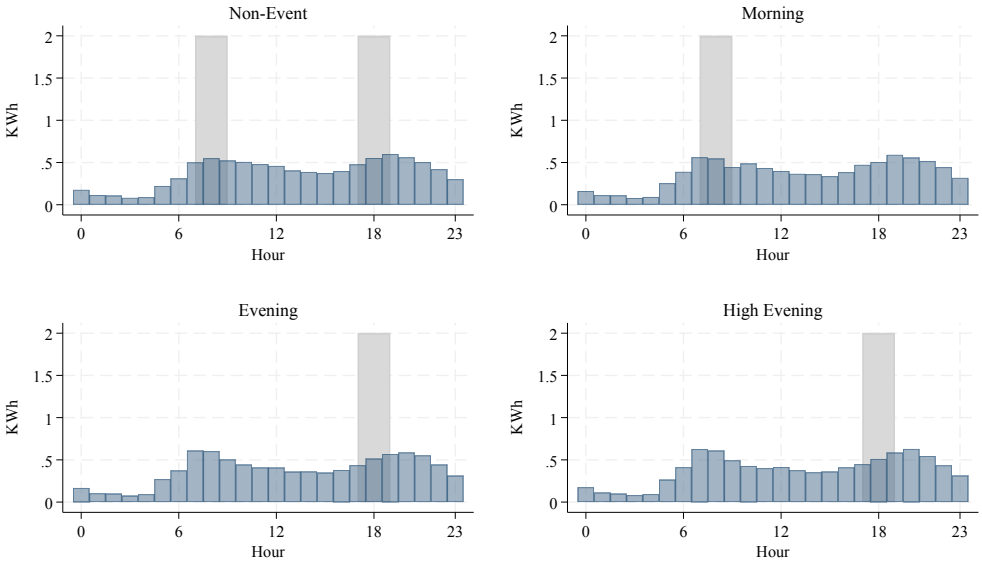
Figure C13 reports the linear specification to compare with the results in Figure 9. We continue to find limited evidence of a large response for Tech households. There is now a small average reduction in hot water usage during peak events, and the reduction in EV usage is significant at the 5% level. For Central, we continue to observe large and significant effects for all device types. The hot water heater has the largest estimated effect, followed by the EV and thermostats.

Figure C14 extends the analysis reported in Figure 9 by allowing the response to peak events to vary by event type. Across both the Central and Tech programs, we do not find any evidence of a differential response by event type.

Figure C10. Average Hot Water Usage by Day-Type and Program



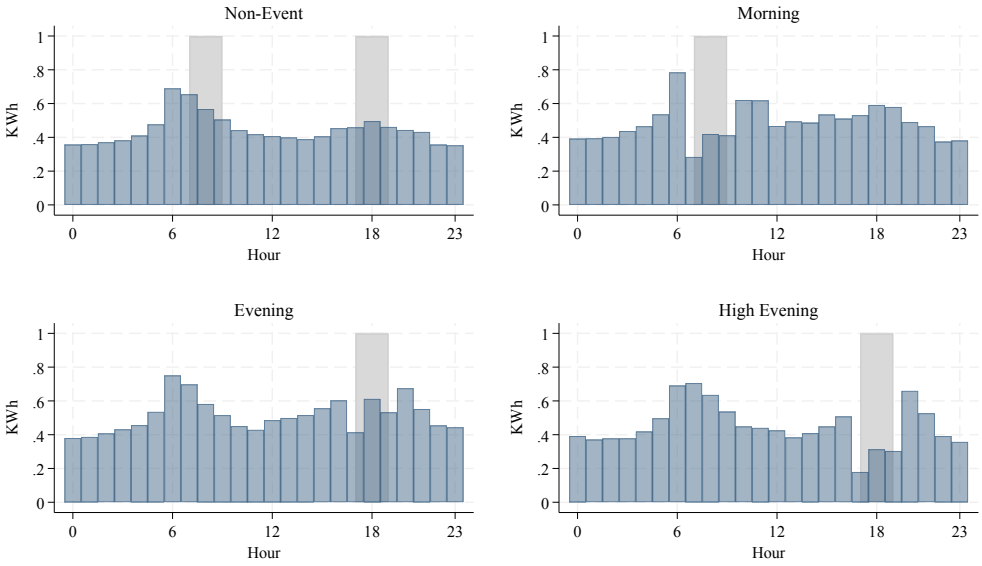
(a) Central



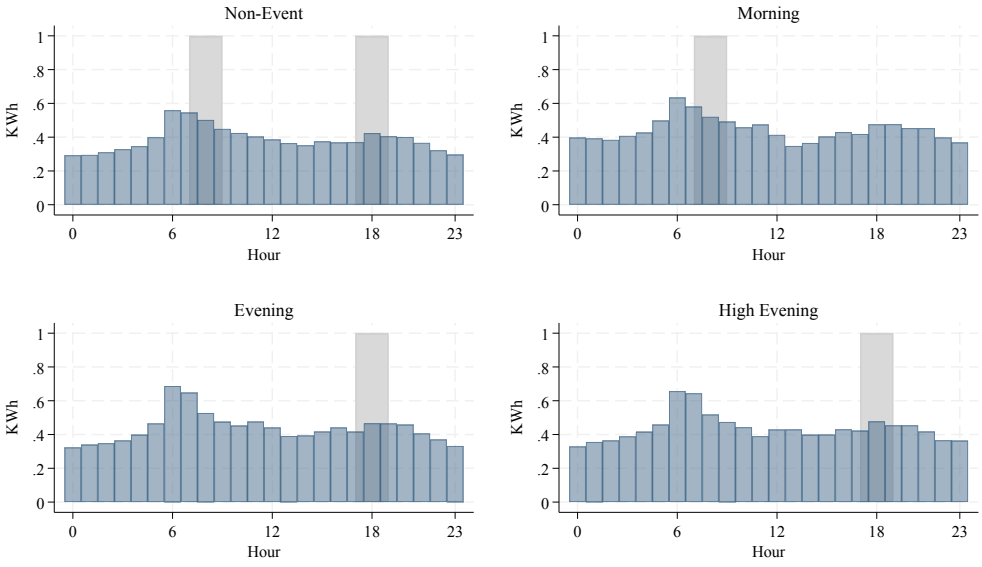
(b) Tech

Notes: Figure plots mean electric hot water heater consumption by hour and demand response program on non-event days, on event days, and non-event days over the period February 1, 2022 - June 30, 2023. Event days are separated by type: Morning, Evening, and High-Evening. The shaded area represents the relevant 3-hour event period.

Figure C11. Average Thermostat Usage by Day-Type and Program



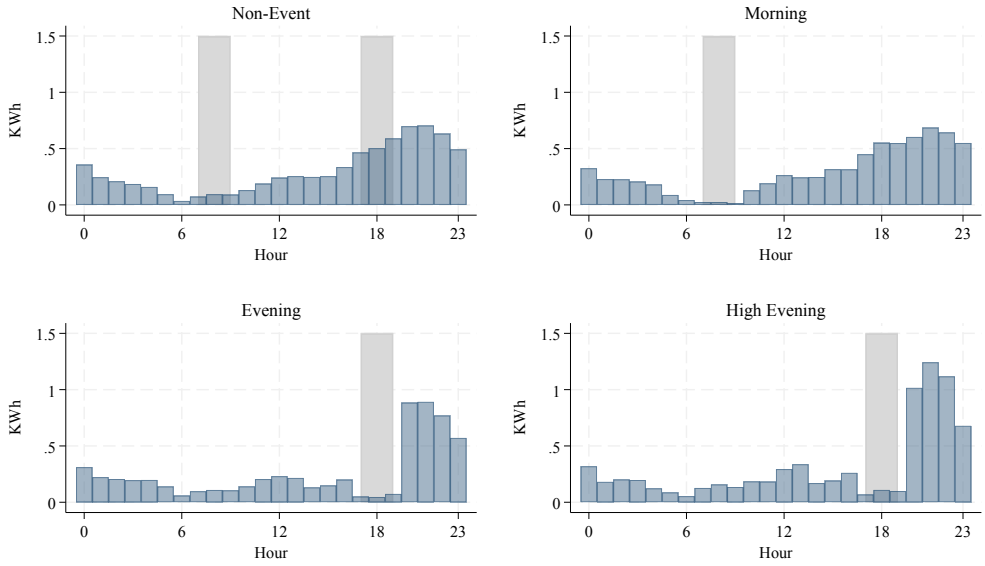
(a) Central



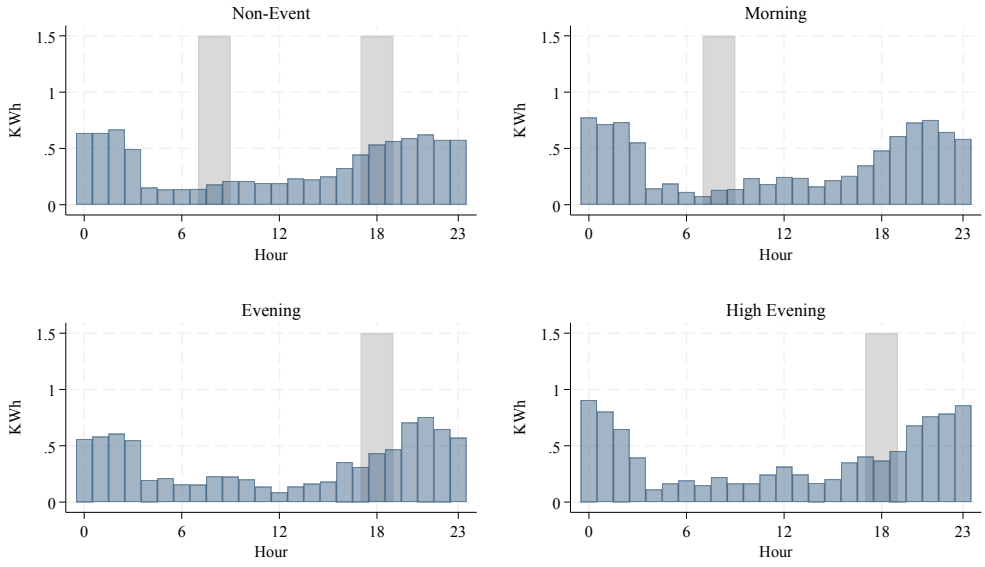
(b) Tech

Notes: Figure plots mean electric baseboard thermostat consumption by hour and demand response program on non-event days, on event days, and non-event days over the period February 1, 2022 - June 30, 2023. Event days are separated by type: Morning, Evening, and High-Evening. The shaded area represents the relevant 3-hour event period.

Figure C12. Average Electric Vehicle Usage by Day-Type and Program



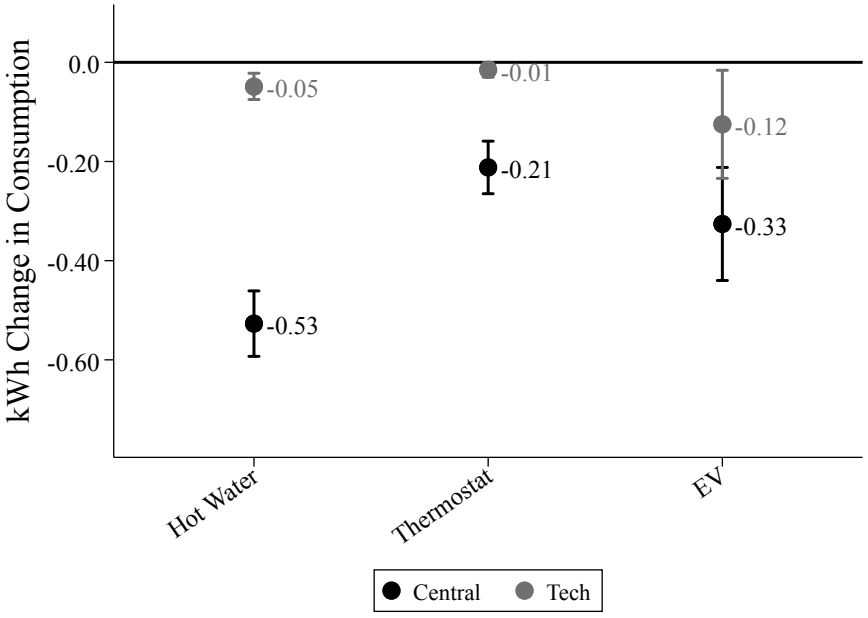
(a) Central



(b) Tech

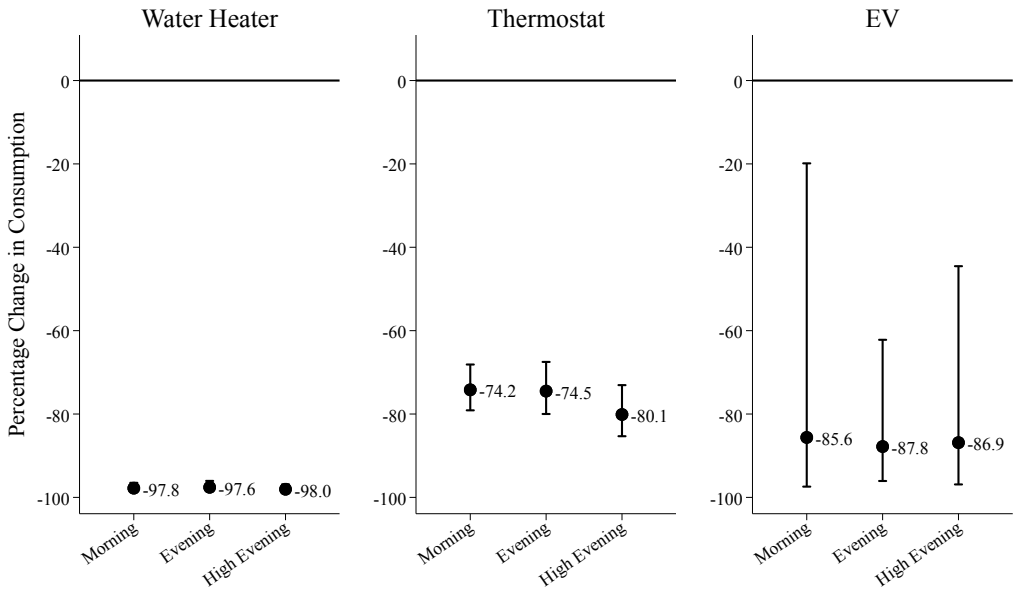
Notes: Figure plots mean electric vehicle consumption by hour and demand response program on non-event days, on event days, and non-event days over the period February 1, 2022 - June 30, 2023. Event days are separated by type: Morning, Evening, and High-Evening. The shaded area represents the relevant 3-hour event period.

Figure C13. Device-Specific Estimated Treatment Effects, Linear Specification - Central and Tech Programs

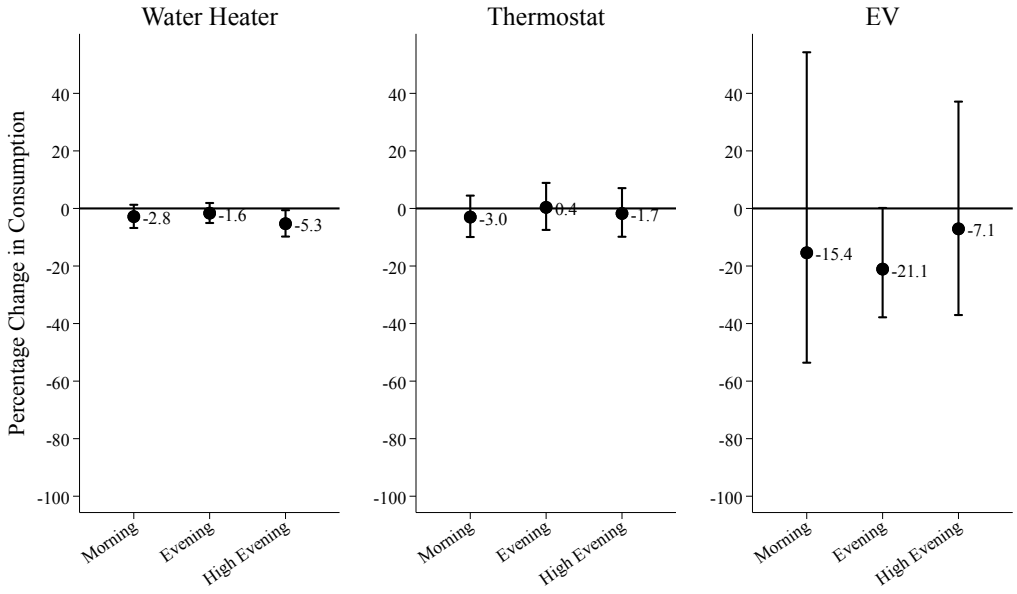


Notes: The reported results are device-specific marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$, replacing the household-level consumption with device-specific consumption and considering a linear-specification instead of a log dependent variable. We present the marginal effects to be a kWh change in electricity during peak events. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

Figure C14. Device-Specific Estimated Treatment Effects by Event Type - Central and Tech Programs



(a) Central



(b) Tech

Notes: The reported results are device-specific marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$ permitting differential peak event responses by event type and replacing the household-level consumption with device-specific consumption. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

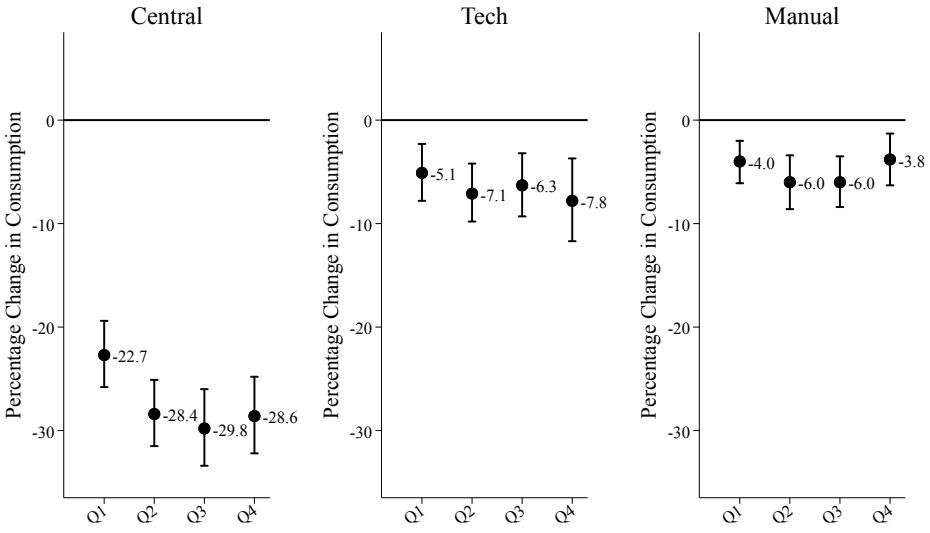
C.11 Heterogeneous Effects by Temperature and Daily System Load

Peak events are randomized across households and days. In this Appendix, we evaluate whether households responses vary with outdoor temperature and system peak load in the Utility's territory. First, we run our program-level regressions in Equation (1) separately for each demand response program, but interact the peak event indicator with heating degree day (HDD) quartile. This permits differential effects by outdoor temperature. We focus on heating degrees because cooling degrees are infrequent in our setting. Second, we adjust this specification by instead permitting the peak event indicator to vary by aggregate system-wide daily peak demand quartile in the Utility's territory.[62]

Figures C15 and C16 demonstrate that there is limited evidence of any differential effects across these dimensions for the Tech and Manual programs. There are no significant differences across quartiles at the 5% level. For the Central program, the peak event response is lower in the first quartile of HDD and daily system peak demand. However, the estimated effect is only statistically significantly different than the Q2 - Q4 quartiles for HDD. These results continue to demonstrate that the Central program produces a consistently higher reduction in demand during peak events. More broadly, for all three programs, the demand reductions are broadly consistent across a range of system-wide demand and temperature conditions.

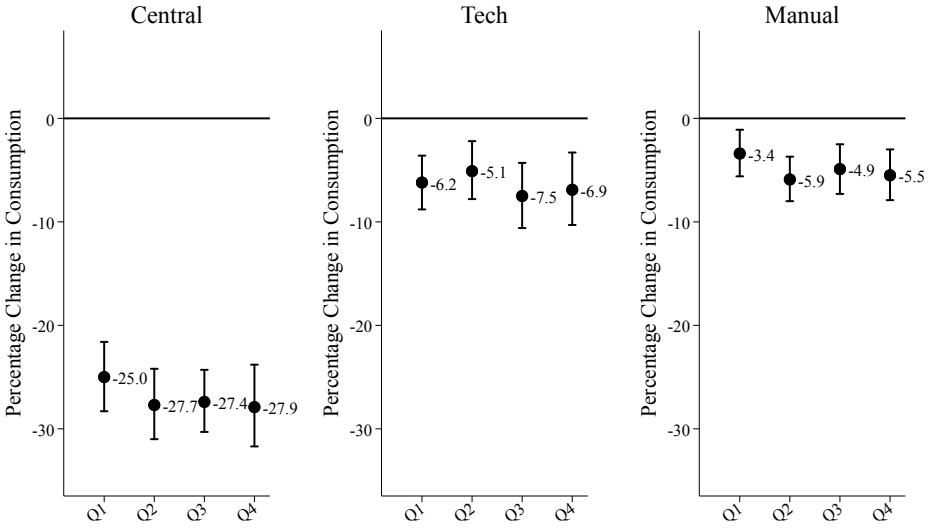
[62] We also ran a version of this analysis that permitted differential effects by quintile. The qualitative conclusions are robust to this more granular specification.

Figure C15. Average Treatment Effect of Participants by Program and HDD Quartile



Notes: The reported results are program-specific marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$, permitting differential effects by HDD quartile. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

Figure C16. Average Treatment Effect of Participants by Program and Daily System Peak Demand Quartile

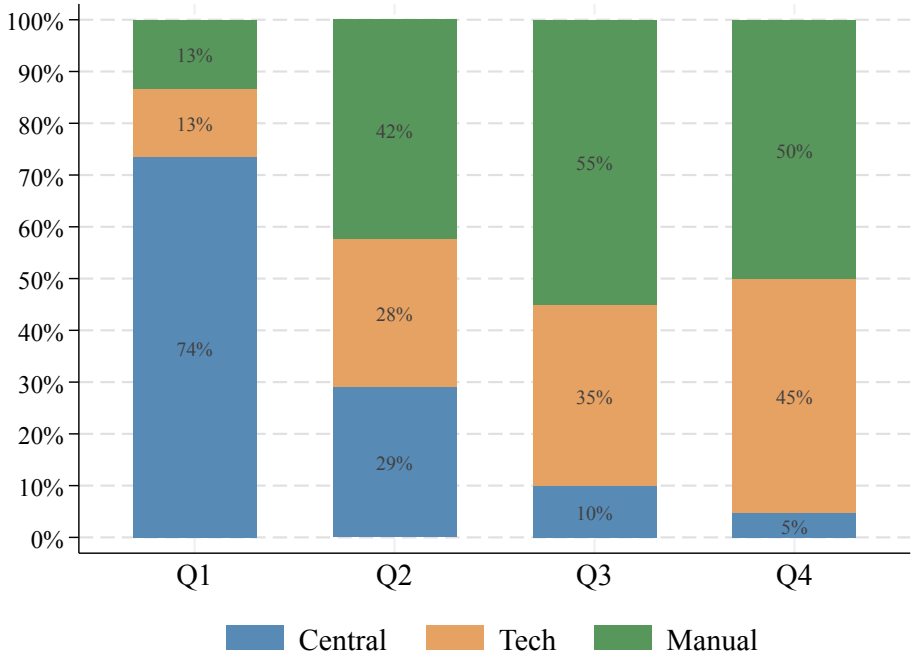


Notes: The reported results are program-specific marginal effects calculated from estimating $\widehat{\beta}_j$ using the specification in (1) separately for each $j \in \{C, T\}$, permitting differential effects by daily system-wide peak demand quartile. We present the marginal effects to be a percentage change in consumption using the transformation $100 \times (exp(\widehat{\beta}_j) - 1)$. Vertical lines indicate 95% confidence intervals. Standard errors are clustered at the household level.

C.12 Performance Quartiles by Demand Response Program

We use the household-level treatment effect estimates from Equation (2) to define performance quartiles, looking across all demand response programs. Figure C17 presents the proportion of each performance quartile from households in the Central, Tech, and Manual programs. We can see that the top quartile consists largely of Central program households, with 13% from the Tech and Manual programs. As we move down the performance quartile, a smaller share of households are from the Central program, while an increasing share are in the Tech and Manul programs.

Figure C17. Share of Participants by Performance Quartile



Notes: The reported results use the household-level estimated treatment effects in Equation (2) to define performance quartiles and compute the share of households in each quartile that come from the Central, Tech, and Manual programs.

C.13 Pre- and Post-Event Spillover Effects

Tables C22 and C23 report results of the pre- and post-event spillover effect coefficients in Equation (1), along with the peak event treatment effects. For the Tech and Manual, there is limited evidence of pre- or post-event effects. In contrast, for Central we observe a distinct positive and significant "snapback" effect post-event. These results persist when we break out the peak event analysis by event type.

Table C24 presents analogous results using our device-specific data detailed in Section 6.3 for the Central and Tech programs. We continue to see limited significance for the Tech program pre-, post-, and during peak events. In contrast, as detailed in Section 6.3, we see large reductions during peak events for all three devices for the Central program. The largest post-event snap-back effects arise from the hot water heater, followed by the thermostat load-controllers. For the EV, there is no evidence of a post-event snap-back effect. However, there is a general reduction in EV charging both prior to and during the event, with the event effect being an approximately 5 times larger. This result indicates that at least a subset of Central households reduced charging running up to events after receiving the event notification (e.g., by simply not plugging in their EV or delaying charging via an App).

Footnote 14 provides a detailed discussion of this snap-back effect and how it should be mitigated with a more sophisticated algorithm upon a full-scale rollout of an automation-driven program.

Table C22. Treatment Effects of Participants by Program - Spillover Effects

Program	Pre-Event	Event	Post-Event
Central	0.0056	-0.3036	0.2063
	(0.0123)	(0.0204)	(0.0195)
Tech	0.0163	-0.0496	0.0034
	(0.0118)	(0.016)	(0.0121)
Manual	0.0030	-0.0542	-0.0031
	(0.0102)	(0.0123)	(0.0101)

Notes. The reported results are the program-specific pre-event, event, and post-event coefficients from equation (1). Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

Table C23. Treatment Effects of Participants by Program - Spillover Effects

Program	Pre-Event	Event	Post-Event
Central			
Morning	0.0166	-0.3122	0.2104
	(0.0196)	(0.0223)	(0.0238)
Evening	0.0024	-0.2905	0.2061
	(0.0173)	(0.0228)	(0.0238)
High Evening	-0.0047	-0.3166	0.2157
	(0.0139)	(0.0243)	(0.0336)
Tech			
Morning	0.0284	-0.0159	0.0237
	(0.0219)	(0.0199)	(0.0188)
Evening	0.0084	-0.0625	-0.0034
	(0.0153)	(0.0185)	(0.0177)
High Evening	0.007	-0.0677	-0.0052
	(0.0147)	(0.0206)	(0.0116)
Manual			
Morning	-0.0155	-0.0815	-0.0433
	(0.0165)	(0.0159)	(0.0162)
Evening	0.0161	-0.0399	0.0078
	(0.0151)	(0.0149)	(0.0145)
High Evening	-0.0137	-0.0470	0.0174
	(0.0125)	(0.0166)	(0.0117)

Notes. The reported results are the program-specific pre-event, event, and post-event coefficients from equation (1), allowing for differential effects by event type. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

Table C24. Device-Specific Treatment Effects of Participants by Program - Spillover Effects

Program	Pre-Event	Event	Post-Event
Central			
Hot Water	-0.0014	-3.8001	0.5624
	(0.0117)	(0.2344)	(0.0231)
Thermostat	-0.0175	-1.4137	0.2532
	(0.0249)	(0.0951)	(0.0310)
Electric Vehicle	-0.4232	-2.0528	-0.1985
	(0.1656)	(0.4866)	(0.3089)
Tech			
Hot Water	0.0046	-0.0298	0.0286
	(0.0122)	(0.0140)	(0.0168)
Thermostat	-0.0146	-0.0131	0.0062
	(0.0246)	(0.0246)	(0.0231)
Electric Vehicle	0.0439	-0.1757	0.0907
	(0.1505)	(0.1272)	(0.1242)

Notes. The reported results are the pre-event, event, and post-event coefficients from estimating equation (1) separately for the Central and Tech using device-level data. Standard errors are reported in the parentheses and clustered at the household level. All specifications include fixed effects at the household and hour-of-sample levels.

D End of Experiment Survey

D.1 Survey Details

The following is text from a voluntary, end-of-experiment survey sent out to participants in the Central, Tech, and Manual programs via email, in mid-June, 2023.

Survey instructions:
"This short survey is designed to hear about your experience in the Peak Rewards Trial through [APP NAME]. All homes had a different experience, and we want to hear about yours.

We appreciate the time and thought you put into this survey.

Properly completed surveys will be rewarded with $20 on bill credit as a token of our gratitude."

Survey questions used in our analysis:

"What is your approximate household income?

- Less than $50k per year
- $50-99k per year
- $100-149k per year
- $150-200k per year
- Over $200k per year
- Don't know/Rather not say"

"For the events you noticed, how often was it worth your time to participate by attempting to reduce your electricity consumption?"

- Never
- Sometimes
- About half the time
- Most of the time
- Always
- Don't know/Not Applicable"

D.2 Survey Response

In this section, we compare the observable characteristics of participants who filled out the end-of-experiment survey to those who did not. The Table below recreates our balance Table A2 (using pre-treatment data), but separates the results by whether or not the household responded to the exit survey. The p-value corresponds to a difference in means test.

Table D1 demonstrates that the non-respondents had larger cumulative consumption during the pre-treatment period. Non-respondents also were more likely to have an electric vehicle. All other characteristics are similar across the two groups.

Table D1. Balance by Exit Survey Response (Pre-Treatment Data)

	Yes	No	p-value
Cumul. kWh			
Winter	5,229	5,892	0.04**
	(2,810)	(3,199)	
Spring	3,675	4,167	0.02**
	(1,788)	(2,104)	
Summer	2,614	3,155	0.01***
	(1,492)	(2,142)	
Fall	3,528	4,012	0.02**
	(1,721)	(2,016)	
Load Factor			
Winter	24.96	25.21	0.78
	(8.61)	(8.89)	
Spring	19.82	19.98	0.83
	(6.64)	(6.93)	
Summer	16.56	17.21	0.37
	(6.87)	(6.76)	
Fall	18.66	19.12	0.51
	(5.91)	(6.72)	
Electric Vehicle	0.24	0.32	0.05**
	(0.42)	(0.47)	
BaseBoard Heating	0.66	0.64	0.71
	(0.47)	(0.48)	
Air Conditioning	0.47	0.51	0.32
	(0.50)	(0.50)	
Electric Hot Water	0.72	0.70	0.56
	(0.45)	(0.46)	
House Duplex	0.83	0.82	0.71
	(0.37)	(0.38)	
Median Income	86,377	87,434	0.55
	(19,853)	(19,835)	
Observations	429	174	

Notes. This table compares pre-treatment average values by whether or not the household participated in the exit survey. Parentheses contain the standard deviations. Cumul. kWh and Load Factor represents the cumulative household-level consumption and load factor by season. Electric Vehicle, Baseboard Heating, Air Conditioning, and Electric Hot Water are indicator variables denoting the presence of each device. House/Duplex is a indicator variable if the home type is a single-family home or duplex. Median Income reports the median household-level income of the Census Dissemination Area where the household is located. p-value from a difference in means test across the two groups. Statistical Significance $^{*}\, p < 0.10$, $^{**}\, p < 0.05$, and $^{***}\, p < 0.01$.

D.3 Survey Regression Robustness

We extend the analysis reported in Table 6 by considering a more flexible set of control variables. In addition to the covariates reported in Section 6.4, we include household-level pre-treatment covariates constructed using the 5 months of data between September 2021 - January 2022: average daily load factor, maximum hourly consumption, and total hourly consumption. We use the post-double-selection (PDS) methodology established by Belloni et al. (2014). The PDS approach uses the LASSO estimator to select control variables to include in our regression. This allows us to include a broader set of control variables in our survey analysis that is limited by a relatively small number of observations. The results of this analysis are reported in Table D2. The conclusions drawn from this regression analysis are consistent with those found in Table 6, with moderate increases in precision on the income-based survey responses in particular.

Table D2. Household Treatment Effect on Worth Time and Income - PDS-LASSO Robustness

	Coefficient	Std. Error	P-Value
Worth Time			
Sometimes	-2.3561	2.1419	0.2713
Half of the Time	-4.6151	2.3247	0.0471
Most of the Time	-7.2430	2.3268	0.0019
Always	-12.5793	2.7561	0.0000
Income			
50 - 99k	5.9511	4.1659	0.1531
100 - 149k	7.8401	4.1132	0.0566
150 - 200k	10.0919	4.1621	0.0153
>200 k	9.7884	4.0547	0.0158
Program Indicators			
Central	-16.7716	1.6714	0.0000
Tech	-1.1261	1.4763	0.4456

Notes. The reported results present the estimates for Worth Time, Income, and Program variables in Equation (3), but including additional household-level pre-treatment covariates constructed using the 5 months of data between September 2021 - January 2022: average daily load factor, maximum hourly consumption, and total hourly consumption. The results are estimated using the PDS-LASSO algorithm to select the control variables. Heteroskedastic-robust standard errors are reported in parentheses. Statistical Significance * $p < 0.10$, ** $p < 0.05$, and *** $p < 0.01$.

The regression in Equation (3) carries out several tests on the Worth Time and Income categories simultaneously. As detailed in Appendix C.4, in this setting it is important to account for multiple hypothesis testing. We follow the same approach in the Survey regression and implement both the FDR Sharpened Q and FWER MHT corrections. The results are reported in Table D3. Focusing on the most conservative FWER MHT-adjustment, we see a reduction in the p-values as expected. The Worth Time coefficients maintain the same thresholds of significance, but the Income coefficients are no longer significant. [63]

Table D3. Household Treatment Effect on Worth Time and Household Income - Multiple Hypothesis Testing Adjustments

Comparison	Model P-Value	FDR (Sharpened Q)	FWER
Worth Time			
Sometimes	0.1025	0.0530	0.2392
Half of the Time	0.0088	0.0180	0.0502
Most of the Time	0.0001	0.0010	0.0054
Always	0.0000	0.0010	0.0008
Income			
50 - 99k	0.2254	0.0930	0.2392
100 - 149k	0.1090	0.0530	0.2392
150 - 200k	0.0302	0.0320	0.1042
$>$ 200 k	0.0310	0.0320	0.1042

Notes. Comparison represents the null hypotheses being tested across demand response programs. The Model p-value reports the p-values on the relevant comparisons using the regression model specified in Equation (3) that are not adjusted for MHT. FDR represents the Sharpened Q-value False Discovery Rate using the procedure outlined in Benjamini et al. (2006). The FWER is implemented using the rwolf2 command in STATA with 5,000 bootstraps. In all three estimates, the standard errors are heteroskedastic-robust.

[63] As can be seen in Table D3, the p-values of the FDR and FWER can be the same for a subset of categories. In these MHT corrections, adjusted p-values from the stepdown procedures are constructed to be monotone in the ordering of the underlying test statistics, with each hypothesis's adjusted value taken as the running maximum over all hypotheses ranked ahead of it (Romano and Wolf, 2016). Where the raw ordering would otherwise violate this constraint, a block of adjacent hypotheses is assigned the same binding value, so identical adjusted p-values across several outcomes reflect the step-function form of the correction.