UC Berkeley
UC Berkeley Electronic Theses and Dissertations

Title
Essays in Energy and Environmental Economics

Permalink
https://escholarship.org/uc/item/8rs2g05g

Author
Gillan, James Michael

Publication Date
2018

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

Essays in Energy and Environmental Economics

by

James Michael Gillan

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Agricultural and Resource Economics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Associate Professor Meredith Fowlie, Co-chair
Professor Catherine Wolfram, Co-chair
Professor Stefano DellaVigna
Assistant Professor James Sallee

Spring 2018

Essays in Energy and Environmental Economics

Copyright 2018
by
James Michael Gillan

Abstract

Essays in Energy and Environmental Economics

by

James Michael Gillan

Doctor of Philosophy in Agricultural and Resource Economics

University of California, Berkeley

Associate Professor Meredith Fowlie, Co-chair

Professor Catherine Wolfram, Co-chair

The coming century will bring numerous environmental challenges and understanding the strategic decisions involved in energy production and consumption will be central to addressing them effectively. In this dissertation, I use methods from applied econometrics, behavioral economics, and industrial organization to investigate various lines of inquiry around this broader motivation. In Chapter 1, I study how residential electricity consumers respond to increasingly complicated incentives that are meant to improve allocative efficiency and test whether their behavior is consistent with standard models. In Chapter 2, I estimate the impact of temperature on high school students' standardized test performance in order to understand how environmental factors affect educational outcomes. In Chapter 3, I evaluate a targeting strategy meant to improve the efficiency of an electricity pricing program and develop a theoretical framework to ground the findings.

The first chapter studies whether consumers are attentive to time-varying incentives to reduce electricity consumption. Dynamic pricing models typically assume that consumers respond to marginal incentives. I use a field experiment to assess the impact of dynamic pricing on residential electricity consumption and find strong evidence of inattention. I propose a model to interpret the results which suggests that the benefits of dynamic pricing may be substantively undermined by inattention. I also explore the role of automation in dynamic pricing, which holds the promise of reducing the cognitive choice frictions that cause inattention and lowering the effort cost of responding to price changes. I report three primary findings. First, households—both with and without automation—significantly respond to a short term price increase by reducing consumption. Second, responses are very insensitive to the size of the price change. A price increase of 31 percent causes consumption

to fall by 12 percent on average, whereas a price increase of 1,875 percent causes an average reduction of 14 percent. Third, automation causes responses that are more than three times larger than the average effect, but are still insensitive to the price level. The results suggest that households use simplifying heuristics when facing dynamic prices and that automation reduces effort costs, but does not resolve inattention. I apply the model to recover bounds on the price elasticity of demand and shed light on the potential attention costs of dynamic pricing.

The second chapter, coauthored with Maximilian Auffhammer and Catherine Wolfram, studies the impacts of extreme temperature on over 5 million students standardized test performance. We exploit plausibly exogenous year-to-year within-school daily weather variation in order to measure the contemporaneous effect of maximum outdoor temperature on aggregate student performance. The exam studied is the California High School Exit Exam, a state-wide standardized test that evaluates high school students' mathematics and English-language arts aptitude and was a requirement for receiving a diploma from 2006-2015. We document a nonlinear relationship between temperature and performance. Temperatures above 27.5°C show statistically significant negative impact on pass rates in both subjects and scores in the math assessment. We also document heterogeneity in the effect by income in the area surrounding the school and find more pronounced effects for schools in the lowest income quartile.

The third chapter, coauthored with Maximilian Balandat and Datong Zhou, evaluates the effect of targeting based on heterogeneous treatment effects using an experiment. We provide a theoretical framework for how various factors undermining external validity affect targeting and the how experimental evaluation of targeting can be used to parse competing mechanisms. Our theoretical framework distinguishes between group-level heterogeneity as defined by covariates and subject-level effects we call individual treatment effects (ITEs). ITEs can only be gleaned through observing program participation using panel data, but capture additional effect heterogeneity within the group-level effects. We partnered with a energy technology company in order to examine the impact using ITEs to target in the field. We find our targeting strategy reduces the costs of the partner by 52 percent and the results are highly significant. The strategy also reduces revenue by 24 percent, indicating an overall increase in profit on the order of 28 percent. We also examine the persistence of the effects and find the cost savings begin to diminish only 60 days after deployment of the targeting strategy. These findings suggest significant potential for reducing the cost of the program, but only in the short-term. Importantly, the experimental evaluation allows us to understand its performance without having to rely on the common practice of conducting ex-post simulations.

Contents

Contents **i**

List of Figures **iii**

List of Tables **iv**

Acknowledgements **vii**

1 Are Households Attentive to Time-varying Marginal Incentives? Evidence from a Field Experiment **1**
1.1 Introduction 1
1.2 Theoretical Model 7
1.3 Design and Data 12
1.4 Empirical Strategy and Results 23
1.5 Model Application 42
1.6 Discussion 49
1.7 Conclusion 53

2 The Effect of Temperature on Student Test Performance: Evidence from High School Exit Exams **54**
2.1 Introduction 54
2.2 Data and Summary Statistics 57
2.3 Estimation Strategy 66
2.4 Results 68
2.5 Conclusion 76

3 The Economics of Targeting: Theory and Evidence from a Field Experiment **78**
3.1 Introduction 78
3.2 Theory 81

3.3 Research Design and Setting 85
3.4 Empirical Results 89
3.5 Conclusion 94

References **96**

A Are Households Attentive to Time-varying Marginal Incentives? Evidence from a Field Experiment – Appendix **103**
A.1 Proof of Proposition 1 103
A.2 Event Feedback 103
A.3 Supplementary Summary Statistics 104
A.4 Attention Dynamics 107
A.5 Sample Attrition 108
A.6 Raw Variation 109
A.7 Results with kWh Consumed as Dependent Variable 112
A.8 Examples of Household Consumption Distributions 119
A.9 Point Balances Over Time 119
A.10 Variation by Point Balance 119
A.11 Results for Candidate Heuristics 124
A.12 Event Calling 126
A.13 Survey 127

B The Effect of Temperature on Student Test Performance: Evidence from High School Exit Exams – Appendix **137**
B.1 Main Results with Alternative Control Specifications 137
B.2 Main Results for Balanced Panel 140

C The Economics of Targeting: Theory and Evidence from a Field Experiment – Appendix **145**
C.1 Proof of Proposition 1 145
C.2 ITE Ranking 145

C.3 Supplemental Empirical Results 148

List of Figures

1.2.1 Model Intuition 10
1.3.2 Example of SMS and Email Event Messaging 15
1.3.3 Enrollment and Assignment by ZIP Code 22
1.4.4 Pricing Hour Event Studies 29
1.4.5 Effect of Marginal Price 31
1.4.6 Effect Heterogeneity by Temperature 39
1.4.7 Persistence of Price Response 41
1.5.8 Demand Estimation 47

2.2.1 Performance Metric Distributions in 2004 and 2015 64
2.2.2 Temperature Variation 65
2.4.3 Effect of Temperature 69
2.4.4 Effect of Temperature on Days Prior to and After Exam 72

3.3.1 Experimental Design 88
3.4.2 Persistence of Targeting Effect 95

A.2.1 Web-Portal Homepage with Event Results 104
A.6.2 Distribution of Post-Enrollment Consumption 110
A.7.3 Pricing Hour Event Studies in kWh Consumed 116
A.7.4 Effect of Marginal Price in kWh Consumed 117
A.7.5 Persistence of Price Response in kWh Consumed 118
A.8.6 Example Distributions of Individual Household Consumption 120
A.9.7 Point Balances by Week After Enrollment 121
A.9.8 Initial Point Balance Variation 122
A.12.9 Distribution of Event Calls Over Time 127
A.13.10 Screenshot of First Page of Survey 130
A.13.11 Results for Survey Questions 136

B.2.1 Temperature Effect (Balanced Panel) 143

B.2.2 Temperature Effect on Days Prior to and After Exam (Balanced Panel) 144

C.2.1 ITE Distribution 147
C.3.2 Persistence of Targeting Effect on Incentives and Consumption 149

List of Tables

1.3.1 Summary Statistics and Balance of Treatment Assignment 21
1.4.2 Event Period Summary Statistics 24
1.4.3 Effect of Pricing Events 26
1.4.4 Effect of Price vs. Moral Suasion (No Price) 34
1.4.5 Effect of Encouragement on Automation Take-up 36
1.4.6 Effect of Automation 38
1.5.7 Elasticity and Attention Estimates 45

2.2.1 Summary Statistics by Test Date 62
2.4.2 Temperature Effect on Scores by Neighborhood Income Quartile . . . 73
2.4.3 Temperature Effect on Pass Rates by Neighborhood Income Quartile . 74

3.4.1 Pre-period Balance and Summary Stats 90
3.4.2 Effect of Targeting Strategy 92

A.3.1 Additional Sample Summary Statistics 105
A.3.2 Sample vs. Eligible Population 106
A.5.3 Estimates of Sample Attrition 109
A.6.4 Estimates of Pricing Event Effect using Post Data Only 111
A.7.5 Effect of Pricing Events on kWh Consumed 113
A.7.6 Effect of Price vs. Moral Suasion (No Price) on kWh Consumed . . . 114
A.7.7 Effect of Automation on kWh Consumed 115
A.10.8 Price Response by Point Balance at Time of Event 123
A.11.9 Effect of Heuristics on log(kWh) 125
A.13.10 Difference in Means between Survey Responders and Non-Responders 129

B.1.1 Temperature Effect on Scores with Varying Controls 138

B.1.2 Temperature Effect on Pass Rates with Varying Controls 139
B.2.3 Temperature Effect on Scores (Balanced Panel) 141
B.2.4 Temperature Effect on Pass Rates (Balanced Panel) 142

C.3.1 Effect of Targeting Strategy (Balanced Panel) 148

To Connie and Mick Gillan

I would never have accomplished this without your unwavering support.
I love you both.

Acknowledgments

I have been humbled by the generosity of so many in the process of completing this PhD. My advisors and mentors, Meredith and Catherine, spent countless hours in their offices and on calls, helping me refine my projects and ideas. Jim continually broadened my thinking or sharpened my language in our frequent coffee talks. Stefano's class was the reason I chose to style myself a behavioral economist and his infectious enthusiasm gave me confidence that my ideas had general appeal.

I can easily say that the Energy Institute at Haas provided the most intellectually stimulating home an energy economist could ever ask for. Severin, Lucas, Max, Reed, Dmitry, and Andy's were always willing to provide constructive feedback on my well-formed and less than well-formed questions on regulatory policy or identification strategies. I can only hope my future work environments will have as exciting of lunch conversations as I came to expect at the 'tute.

My research would never have been possible without the support of the many others who assisted me with the many logistical challenges and bureaucratic hurdles I discovered along the way. Karen, Paula, Casey, Carmen, Diana, Maggie, Max, Datong, and Claire I am so appreciative of your patience and expertise. I would never have had as good of a time doing all the smart things without the brilliant graduate students at UC Berkeley. Walter, Kenny, Ceren, Zhimin, Ed, Matt, Dave, Laura, Andrew, Sophie, and so many more were key to keeping me sane through the PhD process.

Many from my pre-graduate school world were also instrumental in my education. Mike and his lab got me excited about science and shaped my thinking about posing research questions and pitching grants. Jens gave me the invaluable advice to "stay liquid" and was always willing to teach an RA more than was required. And Bevin convinced me to continue down the path of environmental economics and apply to ARE.

Lastly, my family encouraged me to pursue graduate school and their emphasis on the importance of education has made me who I am today. Their tireless love, support, and expectations helped me in more ways than I could ever describe.

To all of you, I would like to say thank you. I am deeply grateful for devoting so much of your time to my academic and personal development and I will continue striving to never let you down.

Chapter 1

Are Households Attentive to Time-varying Marginal Incentives? Evidence from a Field Experiment[1]

1.1 Introduction

An increasing number of consumer decisions feature dynamic pricing—the application of marginal cost pricing to goods with costs that vary over time. Examples include ride-hailing applications like Lyft and Uber (Cramer and Krueger 2016), e-commerce websites such as Amazon (Borenstein and Saloner 2001), and end-use electricity tariffs (Joskow and Wolfram 2012). Neoclassical models imply that dynamic pricing improves efficiency if demand is elastic and buyers are fully informed, yielding a real-time efficiency benchmark where prices and consumption instantaneously adjust to cost fluctuations. However, evidence from a variety of decision settings finds that consumers can be inattentive to marginal financial incentives such as value-added taxes (Chetty, Looney, and Kroft 2009) or shipping costs (Hossain and Morgan 2006).[2] Thus, a necessary component to determining the actual efficiency of dynamic pricing is understanding how attentive consumers are to dynamic marginal incentives. This paper addresses this question within the context of residential electricity demand.

[1]The material in this chapter is from Energy Institute at Haas Working Paper #284. The original version can be found online at `https://ei.haas.berkeley.edu/research/papers/WP%20284.pdf`.

[2]See DellaVigna (2009) for a review and more examples of consumer inattention.

One solution to the inherent tension between dynamic pricing and consumer attention costs is automation. The trend in dynamic pricing follows the wide deployment of internet-connected "smart" devices (e.g., phones and thermostats). In addition to allowing sellers to update prices and monitor consumption at lower cost, smart devices can also enable buyers to automate decisions.[3] If automation accurately represents preferences, it provides a substitute to a more active (and cognitively costly) response to frequent price changes. It is standard to assume that enabling automation has an important role to play when discussing the efficiency potential of dynamic pricing in electricity (Borenstein and Holland 2005, Joskow 2012), but there remains little research testing its performance in the field.

This paper explores three lines of inquiry to assess the effect of dynamic pricing and automation on residential electricity consumption. First, how large is the response of demand to dynamic price changes? Second, do consumer responses depart significantly from the behavior implied by standard neoclassical models that assume full information? Third, if information frictions generate substantive differences between observed behavior and neoclassical assumptions, does enabling automation reduce them in the context of dynamic pricing?

In order to answer these questions, I conducted a field experiment with 5,491 households that face real economic stakes. The experiment was implemented by a private company that pays its customers to reduce their electricity usage when demand reductions are particularly valuable.[4] During the experiment, households continued to pay a monthly utility bill, but were also exposed to one-hour events that featured a linear financial incentive to reduce consumption. The incentive was offered in the form of points that could be redeemed for dollars, but was specifically designed to replicate changes in the per unit price of electricity.[5] Households were given less than one hour ahead notice each time an event occurred and notified of

[3]Examples include technologies that generate automatic price reminders for e-commerce purchases where prices vary such as airline tickets, hotel rooms, and used automobiles. There are also applications that allow for automatic bidding into online auctions platforms like eBay.

[4]Following the taxonomy developed in Harrison and List (2004), this is a *natural field experiment* because households did not know they were part of a study and experimental variation occurred within the company's product.

[5]Households were paid for reducing consumption relative to an individual forecast and penalized for exceeding that forecast at the same marginal rate. This is similar in spirit to the critical peak rebates studied in past dynamic pricing experiments such as Wolak (2006) with the important distinction that there was substantially less asymmetry in the incentive around the forecast. Households faced a linear incentive except when penalizing the household resulted in them losing money from the program overall. This edge case is the only departure from the incentive being financially equivalent to changes in marginal price.

their performance a few days later. Any net earnings could then be cashed out via the company's website.

Each time an event was called during the first 90 days of enrollment, each household was exposed to a price increase that was randomly varied between \$0.05-\$3/kWh. These levels represent dramatic changes in the marginal price given the non-event average price of electricity was \$0.16/kWh. The experiment also featured a recruit and delay design that randomly assigned a subset of households to serve as a control group upon signup. Together, these two sources of variation allow me to identify the impact of marginal financial incentives on consumption over a wide range of prices. To my knowledge, this is the only dynamic pricing study to use within event random price variation.[6]

Households were also offered an automation service that would shut off connected devices when an event was called. Among the enrolled, half of the households were randomly selected to receive a full-cost rebate for purchasing a smart device and connecting it to the automation service. The rebate subsidized the purchase of one smart thermostat or two smart plugs with retail values of up to \$240 and \$80, respectively. This caused an 83 percent increase in uptake from 5.3 percentage points in the unsubsidized group to 9.7 percentage points in the encouraged group. I use this variation to estimate the local average treatment effect of automation on energy consumption for those who adopted as a result of the rebate.

I present three main empirical findings that address the lines of inquiry stated above. First, treated households reduced consumption by 13 percent on average during pricing events relative to control households. The response is precisely estimated and 77 percent of the reduction is explained by households without automation, meaning that the results reflect active decision-making. Second, households were extremely insensitive to the size of the price change. I find that the average reduction during events increased from 12 percentage points to 14 percentage points when the price change increased from the lowest level \$0.05/kWh (a 31 percent price increase) to the highest level \$3/kWh (a 1,875 percent price increase). Third, among households who took up the automation as a result of the rebate, the automation technology increased average response to any price change by an additional 48 percentage points during events, but did not alter the insensitivity to marginal price. Back of the envelope calculations suggest the additional savings from automation are substantial and the average payback period for purchasing a device at the retail price range from 3 to 13 years, depending on the technology.

[6]Ito, Ida, and Tanaka (2017) is the closest existing study within the dynamic electricity pricing literature since it uses random price variation between events. The within event randomization I use in this paper has a broader support and identifies the price responsiveness during each event.

To offer an interpretation of the findings, I propose a model of rational agents with limited attention and heterogeneous costs of information, drawing from the framework introduced in Chetty (2012). In the model, agents choose to respond to dynamic prices with a heuristic or exert costly effort to become fully informed. For example, households could be using the expected price change each event instead of the actual price change. The model delivers the standard choice primitive estimated in dynamic pricing experiments, the price elasticity of demand, but relaxes the condition that all households are perfectly attentive.

I provide empirical evidence in support of the model's assumptions over competing mechanisms using a separate set of moral suasion interventions that occurred after the pricing interventions. During these events, households were randomly assigned to receive moral suasion messages promoting the environmental attributes of reducing electricity usage instead of dynamic price changes. Reductions from price messages are larger and significantly different from reductions from moral suasion, which are not significantly different from zero. This suggests that households observe the message content and that responses are driven by price rather than preferences for the environmental attributes of consumption.

I use my experimental variation to show that price elasticity estimates derived from the limited attention model differ substantively from those obtained from the neoclassical assumptions standard to the dynamic pricing literature. I make several simplifying assumptions to give the model empirical tractability and estimate bounds on the price elasticity of demand by assuming a plausible range for the agents' heuristic response. The bounds are -0.470 and -0.050, generally covering the range of estimates from past studies with longer term price variation. Importantly, the range implies the consumer is substantially more elastic than the estimate recovered using the standard method from the literature, -0.045.

Using my estimates of the price elasticity of demand, I examine an important implication of inattention: the misoptimization cost a household experiences by being inattentive to the marginal price. I estimate the average misoptimization cost per one hour event falls in the range of \$0.03 to \$0.43, representing 10 to 141 percent of the average savings. In the program studied here, households experience about 100 events per year, implying the nominal stakes of this dynamic pricing program are small for the average household and supporting the interpretation that inattention could be rational with modest information costs. However, small mistakes could quickly add up if applied to the 130 million customers that make up the \$178 billion U.S. residential electricity sector, over 90 percent of whom do not pay time-varying rates (EIA 2015).

This paper relates to the larger literature on attention and information provision in economic decision-making. In the energy context this includes investigations into

consumer attentiveness to the non-salient energy attributes of durable goods such as cars (Sallee 2014), refrigerators (Houde n.d.), and light bulbs (Allcott and Taubinsky 2015). There is also a connection to the literature on rational inattention in the presence of costly information acquisition (Sims 2003; Bartoš et al. 2016). Within this literature, I make three contributions.

First, the paper provides novel evidence on how information costs manifest in dynamic pricing. This contributes to the nascent economics literature that seeks to empirically identify how consumers actually *perceive* non-salient or complex financial incentives. For example, Rees-Jones and Taubinsky (2016) provide strong evidence that individuals facing nonlinear tax schedules exhibit behavior consistent with "ironing" or linearizing the schedule using their own average rate. Ito (2014) finds similar results in electricity consumption, showing that households respond to lagged average rates rather than marginal incentives when facing nonlinear tariffs in their monthly bills. Within the context of this study, households respond to a positive price change, but not to the marginal rate. This hierarchical simplification of information is consistent with the heuristic phenomena of *scope neglect* shown in contingent valuation studies (Kahneman 2003).[7] These phenomena could also exist in other dynamic pricing settings such as ride-hailing applications if consumers do not check the price before making the decision to use the service.

My finding that demand is insensitive to the magnitude of the price change extends and complements the existing literature on dynamic pricing in electricity. In previous experiments, Jessoe and Rapson 2014 (henceforth JR) find information provision makes demand more elastic during price changes and Ito, Ida, and Tanaka (2017) (henceforth IIT) find consumers respond much more to price than the non-price motivator moral suasion.[8] JR study six events with two price increases of 200 and 600 percent and find an average response of 16 percent for informed households. IIT study 31 events with three price increases between 260 and 420 percent and uses random variation in the price of each event to estimate the casual effect of the price level. IIT finds increasing the price change from the lowest to the highest level increases reductions from 14 to 17 percent, implying a price elasticity of -0.14. By using five price increases between 31 and 1,875 percent, I characterize the response to marginal incentives over a substantially broader support and thus test for consumer inattention more completely.

[7] Findings from psychology and economics show that individuals typically respond to more readily accessible categorical characteristics of a decision (a price change occurring), but not its extensive characteristics (the magnitude of the change).

[8] A number of other studies have used experiments to evaluate consumer responses to dynamic electricity prices (Wolak 2006, 2010, 2011, Faruqui and Sergici 2010, Allcott 2011a, Fowlie et al. 2017). See Harding and Sexton (2017) for a comprehensive review of the literature.

Second, the paper is the first to provide conclusive evidence on the price elasticity of residential electricity demand in the very short-run. Previous studies have focused almost exclusively on "day-ahead" dynamic pricing where households were given notice on the day prior to a price change.[9] The fact that households respond to price changes with less than one hour notice (the average was 5 minutes) has special importance in electricity due to the structure of wholesale electricity markets. The timing of the events in this study mimics the deployment of "real-time" price interventions that respond to the spot market for electricity procurement. Previous studies have used theory and simulations to study the potential for real-time pricing to provide efficiency using elasticity estimates from longer-term price variation (Borenstein and Holland 2005). My results provide the first experimental evidence on the efficiency potential of real-time residential pricing programs that better reflect marginal costs from the wholesale market.[10]

Third, the paper provides novel evidence on how automation affects dynamic price responses. To the best of my knowledge, Bollinger and Hartmann (2017) is the only other study of automated price responses using a field experiment, but their research design, like others in the literature, evaluate technology regimes that have been assigned to entire treatment groups. The encouragement design is advantageous because it allows for the invoking of revealed preference when interpreting adoption decisions.[11] This information is valuable because uptake is generally low and households can be sensitive to defaults when making decisions about dynamic electricity tariffs (Fowlie et al. 2017). The results also provide suggestive evidence that automation, while preferred to a manual response, does not perfectly reflect preferences.[12] This derives from the fact that the company's service provided a blunt instrument that did not allow household to enter a reservation price for being automated. While households could override the response, automated households respond significantly

[9]To the best of my knowledge, JR is the one exception. Three of the six events they study give 30-minute notice. While their results for these periods are statistically insignificant, their point estimates are similar in magnitude to what I find.

[10]Allcott (2011) studies a program that features hourly price variation, but despite including the term real-time pricing in its title, the hourly price variation households were exposed to in that study came from the day-ahead forward market.

[11]The encouragement rebate also reflects the tools policymakers typically use to deploy new technologies. Examples in energy include subsidizing home energy retrofits (Fowlie, Greenstone, and Wolfram n.d.), fuel-efficient vehicles (Sallee 2011), and rooftop solar panels (Hughes and Podolefsky 2015).

[12]Given that automation generally functions as a default, ensuring it accurately captures preferences poses a difficult task as evidenced by the literature exploring default bias in retirement savings (Carroll et al. 2009, Bernheim and Rangel 2009). I do not address this tension in this study and leave the welfare consequences of automation defaults to future work.

to price *and* moral suasion, suggesting a potential default effect. This suggests future dynamic pricing research should address the framing of decisions carefully if automated responses are to be interpreted as preferences.

The paper proceeds as follows: Section 1.2 develops the theoretical model, Section 1.3 reports the experimental design and describes the data, Section 1.4 reports the empirical results, Section 1.5 applies the model to obtain estimates of demand parameters, Section 1.6 discusses and interprets the results, and Section 3.5 concludes.

1.2 Theoretical Model

In this section, I develop a theoretical model to guide interpretation of the empirical results.[13] It draws upon the model from Chetty 2012 which derives bounds for the price elasticity of demand when consumers misperceive prices.[14] The model's main feature is that households have limited attention when facing an unexpected change in price and must exert costly effort to become completely informed to its value. The structure delivers two sharp tests of whether there are inattentive households in a population with smooth preferences. It also recovers the price elasticity of demand. I demonstrate the model's application to my setting to recover estimates of the price elasticity in Section 1.5.

1.2.1 Limited Attention and Dynamic Pricing

Consider a heterogenous population of consumers, where the utility of agent i is given by:

$$U_i(y, s) = \frac{y^{1+1/\eta}}{1 + 1/\eta} - p_s \cdot y - \kappa_i \cdot \mathbf{1}(s = c). \tag{1.1}$$

y is a dynamically priced consumption good (i.e., electricity) and preferences for y follow a constant elasticity form governed by the parameter η.[15] p_s is the perceived per unit price of y and is a function of the agent's chosen information set s. Agents

[13] The model was not pre-specified and was developed after preliminary results were obtained.

[14] Chetty 2012 derives bounds on price elasticity under a more general set of adjustment costs which can manifest as price misperceptions or information costs. My model elaborates on the interpretation as price misperceptions.

[15] The utility form specified in 1.1 is isomorphic to the common assumption in the literature that households have quasilinear preferences over y and a numeraire good with a static price. The relevant first-order condition is equivalent.

choose between a "bounded" (b) and "complete" (c) information set such that $s \in \{b, c\}$ and the complete information set is the actual dynamic price. I refer to agents who choose the bounded set as "inattentive" and those who choose the complete set as "attentive". Consumers experience a positive utility cost $\kappa_i \sim F_\kappa$ for being attentive, which is the only form of heterogeneity between consumers.

Suppose the model has two stages. In the first stage, the price changes unexpectedly from p_0 to p_c and agents choose their information set by maximizing expected utility following von Neumann-Morgenstern preferences. In the second stage, they optimize consumption using the chosen information and utility is experienced. Consumption can be determined by solving the model backwards. Conditional on an information choice, the first-order condition for y implies $y^*(p_s) = p_s^\eta$. Agent i chooses to be attentive if and only if

$$\mathrm{E}[U_i(y^*(p_c), c)|\widehat{F}_p] > U_i(y^*(p_b), b) \tag{1.2}$$

where $\widehat{F}_p$ denotes the agent's prior over the distribution of the dynamic price. Note that this framework is isomorphic to situations where the price is clearly stated, but consumers face a cognitive cost to map the stated price to the cost of services rendered by an input such as electricity.[16]

To map individual decisions to aggregate demand, let k denote the cost at which agents are indifferent between the two information sets. Agents with $\kappa_i \leq k$ are attentive to actual price change and those with $\kappa_i > k$ are inattentive in that they may respond to the price change, but not its actual value. For a unit mass of consumers, observed aggregate demand is given by

$$Y(p_b, p_c) = \sigma y^*(p_c) + (1 - \sigma) y^*(p_b) \tag{1.3}$$

where $\sigma \equiv F_\kappa(k)$ represents the fraction of attentive agents. Alternatively, if we interpret (1.3) as the demand for a representative agent and κ_i as a stochastic information cost, σ represents the probability of being attentive. This does not change the model's testable implications.

I focus on two quantities of interest: the elasticity of demand with respect to the dynamic price level and the change in demand from the counterfactual of no price change. The elasticity of demand with respect to the actual dynamic price is given by $\varepsilon \equiv \frac{\partial Y}{\partial p_c} \cdot \frac{p_c}{Y} = \sigma\eta$. Thus, it is a function of the fraction of attentive

[16]For example, households may observe the price per unit kWh conveyed in the signal m, but mapping this to the operating cost of lighting versus cooling versus TV or computer use could require substantial cognitive effort. The model can be easily modified to include these scenarios by introducing the costly information decision between m and p_c, where m contains the electricity price, but p_c is the cost of services.

agents and the preference parameter. The change in demand that occurs when price changes from p_0 to p_c is given by $\Delta Y(p_c, p_b) = \sigma \Delta y^*(p_c) + (1-\sigma)\Delta y^*(p_b)$ where $\Delta y^*(p_s) \equiv y^*(p_s) - y^*(p_0)$ is the change for an agent of type $s \in \{b, c\}$.

To gain intuition, Figure 1.2.1 plots three cases. If no price change occurs, all agents consume $y_0 \equiv y^*(p_0)$. If the entire population is made up of attentive agents ($\sigma = 1$), consumption changes from y_0 to $y_p \equiv y^*(p_c)$ when the price increaes to p_c and the elasticity of aggregate demand identifies the underlying preference parameter because $\varepsilon = \eta$. However, if any fraction of agents are inattentive ($\sigma < 1$), η is not identified by the elasticity alone. In the extreme, if all agents are inattentive ($\sigma = 0$), aggregate demand is perfectly inelastic to the dynamic price level at level $y_b \equiv y^*(p_b)$, indicated by the dashed gray line in Figure 1.2.1. Demand for a mixture of households ($0 < \sigma < 1$) is plotted in the dashed black line which slopes down slightly with a discrete change from y_0 for dynamic prices in the neighborhood of p_0.

The model provides two sharp tests of whether there are inattentive agents in a population.

Proposition 1. *If $\sigma < 1$ and $p_b \neq p_0$, then $\lim_{p_c \to p_0} \Delta Y(p_c, p_b) \neq 0$.*

Proof: See Appendix.

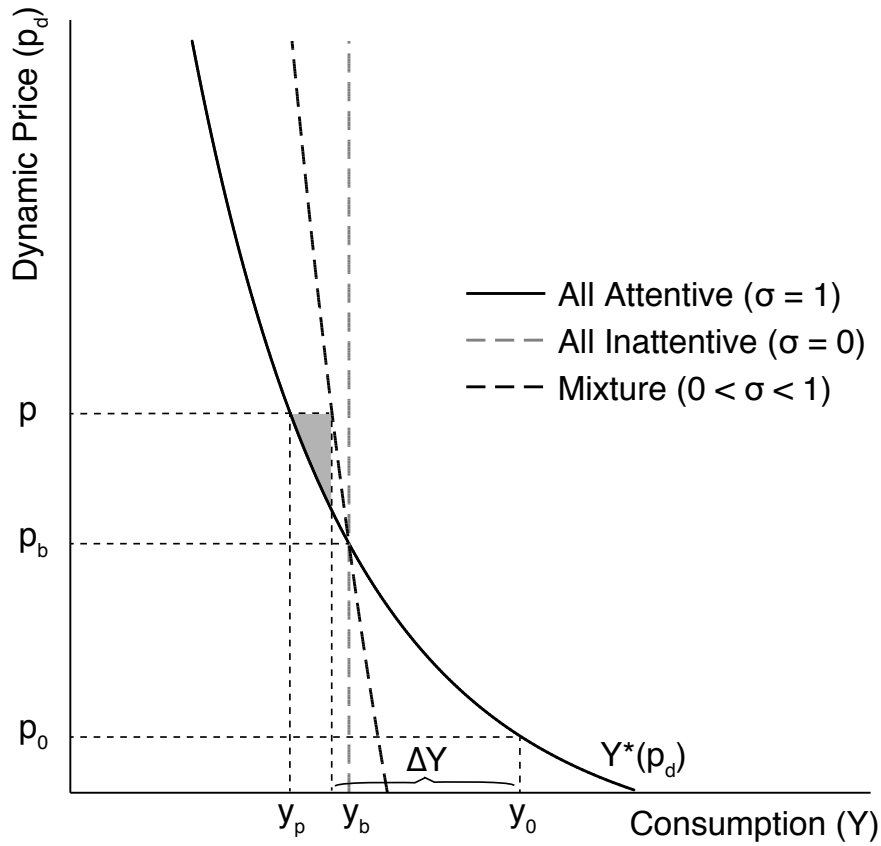
Proposition 2. *If $\sigma = 0$, then for all η, $\varepsilon = 0$.*

Proof: If $\sigma = 0$, then $\varepsilon = \sigma\eta = 0$. □

Proposition 1 states that there will be a discrete shift in demand for very small price changes if there is some amount of inattentive agents using a coarse information set instead of the true marginal price. Proposition 2 states that if all households are inattentive, demand will be completely inelastic with respect to the level of the price change, regardless of the taste parameter. The proof for Proposition 2 follows trivially from the definition of ε and the proof of Proposition 1 is included in the appendix. The model does not explicitly consider the effect of automation on the decision setting, but can be extended to incorporate it.[17]

[17] We can think of automation as affecting two dimensions of the household's decision. First, it reduces the physical adjustment cost of effort for responding to a price change. For example, automating the response of an appliance that is not readily accessible such as an electric water heater or an entertainment console is likely to make the household more elastic. Second, the automation provides a substitute for attention, allowing for price conditional responses. While the company's automation is overly simple in this respect, it provides a default response we can interpret as responding as if an extreme price change occurred. These features can be incorporated explicitly through the utility function for adjustment costs and the information choice set.

Figure 1.2.1: Model Intuition



The figure plots the effect of a price change for different fractions of attentive households for the model developed in Section 1.2. The vertical axis plots the dynamic price and the horizontal axis aggregate consumption. The two types of households are indicated by the solid black and dashed gray lines. The solid black line shows demand if all households were attentive ($\sigma = 1$), the gray dashed line shows demand if all households were inattentive ($\sigma = 0$), and the black dashed line shows demand for a hypothetical mixture with $\sigma \in (0, 1)$. Consumption at the non-dynamic price p_0 is denoted y_0. A hypothetical price change p is indicated and shows that attentive households consume y_p. Inattentive households consume according to the heuristic price, p_b, at y_b. For the mixture case, the observed aggregate demand for a price change p is not labeled, but falls in the range $[y_p, y_b]$. In this case, the consumption change from the price change is shown by the difference ΔY and the shaded gray area indicates the attention costs of misoptimizing using the heuristic.

1.2.2 Empirical Elasticity Estimates

The model developed above has implications for the standard empirical estimation of the elasticity of demand. To see this, consider the approximation of the elasticity using a single price change and the aggregate change in consumption defined as:

$$\widehat{\varepsilon} \equiv \left(\frac{\Delta Y}{p_c - p_0}\right)\frac{p_0}{y_0} \approx -\eta\left(\sigma + (1-\sigma)\left(\frac{p_b - p_0}{p_c - p_0}\right)\right) \tag{1.4}$$

where I use the approximation that $\Delta y_s \approx -\eta\left(\frac{y_0(p_s - p_0)}{p_0}\right)$ for $s \in \{b, c\}$. For $\sigma = 1$, we can see that the empirical elasticity approximately recovers the taste parameter η. For $\sigma < 1$, the taste parameter is scaled by the potentially confounding term in parentheses. If $p_b \neq p_c$, we can see that $\widehat{\varepsilon} \neq -\eta$. In the simplified model above, p_b is the expected price. In this case, the model implies that the empirical elasticity will be larger (smaller) in magnitude than implied by the taste parameter if expectations are higher (lower) than the true price.

This highlights an important implication of the model. If there are any inattentive households, recovering the true preference parameter η is contingent upon knowing or making assumptions about the household's information set. While verifying households beliefs is an empirically burdensome exercise, it is important to note that p_c is a random variable in true dynamic pricing and this model. If p_c is a consistent price change with little uncertainty, there would be more support for assuming beliefs are accurate such that $p_b = p_c$. This is the case for many existing dynamic pricing programs, where price changes are of a consistent magnitude during select hours of the day. However, the structure of these programs do not reflect the uncertainty of the real-time pricing ideal.

1.2.3 Misoptimization Costs

Using the model, I examine another implication of inattention: the misoptimization costs a household experiences by using bounded information. These costs represent the utility loss from reducing consumption for which the household has a high (low) willingness-to-pay (WTP) when the price is actually low (high), thus consuming more (less) than is optimal given the true price change. The misoptimization cost for the households can have direct implications for the efficiency of dynamic pricing when the price change p_c is equal to the marginal cost. In such a case, the mismatch of consumption with marginal WTP also reflects an allocative inefficiency.

Quantifying the misoptimization requires me to make an assumption about the household's true preferences. Namely, I assume that households *experience* utility

as though they had no information costs such that $\sigma = 1$. That is, preferences are revealed by their consumption level if they observed the actual price change. Under this assumption, the average utility misoptimization for given price change p is

$$a(p, p_b) \equiv \left| \int_{y^*(p)}^{Y(p,\, p_b)} p(y) - p \; dy \right| \tag{1.5}$$

where $p(y)$ is the inverse demand curve of $y^*(p)$. Visually, the integral in (1.5) corresponds to the area between the true demand curve and the dynamic price within the range of the observed quantity to the implied fully attentive quantity. Figure 1.2.1 shows this area shaded in gray. The order of integration is arbitrarily defined and I take the absolute value since any deviation from y^* is counted as a positive cost. Note that $a = 0$ for $p_b = p_c$ or if $\sigma = 1$ because either would imply $Y = y^*$.

1.3 Design and Data

This section provides a description of the experimental design and summary statistics.[18]

1.3.1 Recruitment

The experiment was implemented through the product of a private company that serves as a third-party demand response provider (DRP) in the state of California during the 2017 calendar year. The DRP offers a service where households who enroll are paid for reducing their electricity consumption during hour-long dynamic pricing events. The company also allows customers to automate their consumption during events to facilitate responsiveness by connecting their smart home devices via a web portal. Events and automated response are described in detail below. Enrolled households continue to receive a bill from their utility for their electricity consumption, but by enrolling in the program they have the opportunity to earn money from the DRP and reduce electricity expenditures overall. Households within the service territories of the three major California investor-owned electric utilities (IOUs) were eligible to enroll in the study.[19]

[18] The section below describes the complete design. At the time this draft was written, enrollment had been completed, but interventions had not concluded. I have collected intervention data through September 2017. The final paper will include data from October 2017- December 2017 and will represent approximately 90 percent of the final data.

[19] The IOUs are Pacific Gas & Electric, Southern California Edison, and San Diego Gas & Electric.

Recruitment was conducted by the DRP on an ongoing basis from January 1, 2017 to August 15, 2017. During this period, households were contacted through a variety of outreach channels and told they could earn up to $300 a year for reducing their electricity consumption.[20] Since households chose to enroll in the program, the study sample will not necessarily be representative of the population of eligible households. However, programs like these are unlikely to be mandatory in the near future, suggesting that volunteers are the relevant population. These factors should be noted when interpreting the external validity of the results.

A total of 13,782 households were recruited for the study. I define recruited households as those who created an account on the DRP's website by entering an email address and password. Of these, 6,169 households completed sign-up by providing enough information to connect their utility data so that the DRP could monitor their consumption and calculate payouts.[21] I remove households with erroneous meter data and those who appear to generate electricity on-site to arrive at the study sample of 5,491 households.[22]

1.3.2 Experimental Platform

A description of the DRP's product provides context that is useful for interpretation and assessment of construct validity. The DRP markets its product as a game where users can earn points during pricing events. Points are awarded with a linear incentive mechanism, where households gain points for using less electricity than the level the DRP forecasts and lose points for exceeding the forecast.[23] The reward (or loss)

[20]Recruitment channels included digital advertising, social media, in person outreach, referral bonuses for existing customers, and coverage on a national radio program. Households were also offered rewards for signing up that varied during the recruitment period from $20 to no reward. There were also periods during which the DRP used referral bonuses to existing customers which may have driven recruitment.

[21]According to the DRP, the 45 percent follow-through rate is consistent with what they normally see and can be attributed to the requirements of the enrollment pipeline. In order to enroll, a household is required to sign a series of forms after entering their email in order to connect their utility information and created an account. These forms pertain to disenrolling from any other DR programs run by their utility and entering into an exclusive contract with the DRP. Each IOU runs its own unique DR program and households were required to end participation in all other DR programs as a condition of enrolling in the study.

[22]I identify on-site generation as having a negative meter read during any period prior to or during enrollment.

[23]Forecasts are calculated for each household using the well-established "10 in 10" methodology defined by the California Independent System Operator as a weighted average of past consumption during similar hours of the day (CAISO, 2014). I verified the accuracy of the forecast methodology independently on pre-treatment data and the control group. The forecast produces estimates of

for an event is calculated as $Reward = Incentive * (Forecast - Usage)$.[24] Points earned (lost) during an event are then added (subtracted) to a balance the household maintains with the DRP.

The reward mechanism functions financially equivalent to an increase in the per unit price of electricity due to the symmetry of the reward around the forecast except for one exception: the company does not allow negative point balances. Thus, the incentive to consume may be asymmetric around the baseline if the household has zero points in their balance. If the household is at the zero lower bound, the effective price change would be lower than the stated reward level. About 20 percent of households have zero balances during events, but the results are not driven by these consumers.[25] While the reward mechanism is certainly different from a simple price change, it provides the closest example of real-time variation in the marginal incentive to consume electricity of any experiment to-date.

Households were informed of an impending event with an hour or less notice via email and an SMS text message to their mobile phone. Figure 1.3.2 shows messaging for an example hour for both SMS and email. In both forms of contact households are told the specific period of the event, the incentive level, and the company's forecast for their usage. The incentive level was communicated as saliently as was possible within the constraints of the product infrastructure. 1 to 3 days after an event occurred, the household's performance was evaluated and communicated via email and the DRP's website. Upon logging in to their account, the resulting number of points gained or lost for the past four events was displayed and the household could scroll to view the complete history of event performance.[26]

Points could be electronically cashed out via a secure online payment system or donated to cause of their choosing. A household's point balance was shown on the upper right section of the home page. One point was equivalent to \$0.01 when cashed out and households could only cash out after they had \$10 worth of points in their

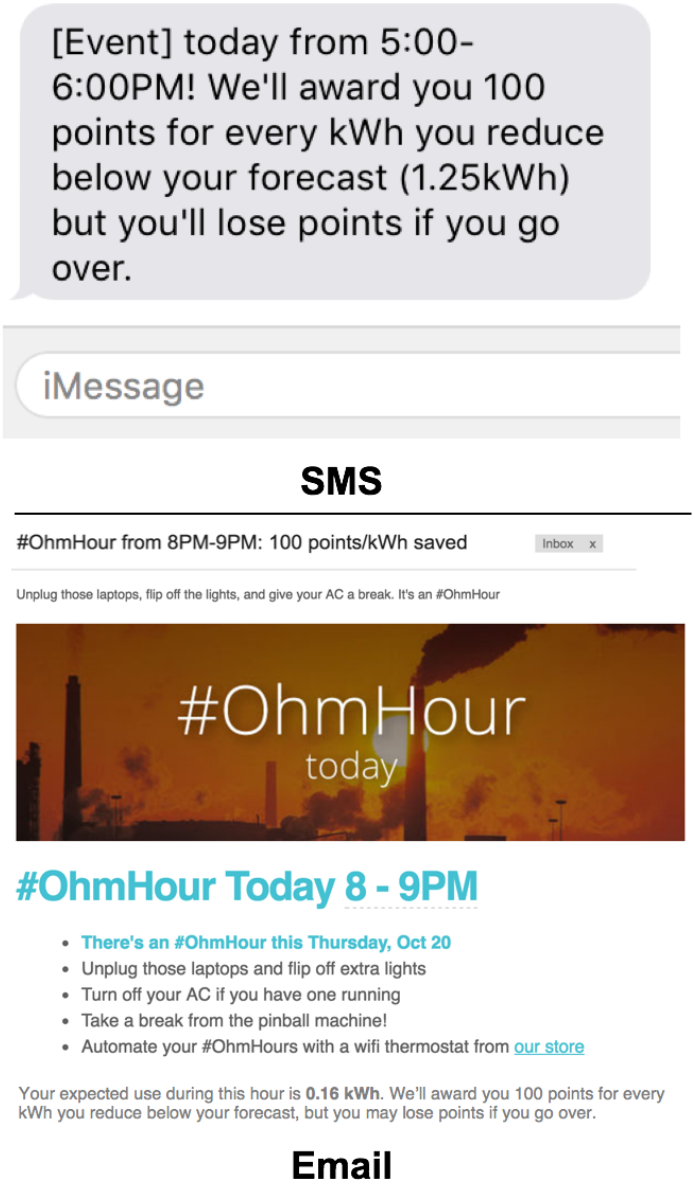
consumption which are slightly negatively biased by less than 5 percent of consumption. While the susceptibility of these types of forecasting methodologies to gaming is often noted in the literature, it often pertains to situations where the user has ample time to change usage prior to an event such as with day-ahead notice. The hour-ahead notice for the events studied in this paper plus their unpredictability make it less likely gaming is an issue.

[24]For example, if the incentive level is 50 points per kWh, the household's forecast is 1.5kWh, and they use 1kWh, they will earn 25 points for the event. Conversely, if they use 1.7kWh, they will lose 10 points.

[25]I provide evidence in the Appendix that the main findings are not driven by the zero balance lower bound. The majority of households start with a positive balance and within a few weeks of enrollment less than 30 percent of households have a zero balance. The Appendix also provides information on how the point balances evolve over the course of the program.

[26]The Appendix shows a screenshot of the home page.

Figure 1.3.2: Example of SMS and Email Event Messaging



The figure shows messaging for two example events that occurred during the experiment. Both messages convey the time of the event, the incentive level, and the company's forecast. The top pane shows messaging via SMS and the bottom pane shows messaging via email for a different event. The email contains more information on suggested strategies for reducing consumption. The yellow boxes cover the company's proprietary term for events which I have redacted for the purpose of confidentiality.

balance. The mapping between points and dollars is important since it allows me to interpret responses to variation in price as representative of demand for electricity. I provide evidence that households understood the value of the points in Section 1.6.

The automation service the company provides remotely shuts off connected devices during a pricing event and turns them back on when the event is over. For example, if an appliance is plugged into a connected smart plug, the company will remotely break the circuit between the plug and the appliance so no more load can be drawn from the socket. With connected smart thermostats, the service turns the thermostat off so that the central air-conditioning (AC) or heating does not cycle. This automates the decision to reduce consumption for connected appliances. Households can override the automation by turning their thermostat or plug back on manually or remotely if they have connected their smart device to their mobile phone. The process for connecting a smart device involves going on the company's website and entering appliance information. Importantly, the automation service is not a function of the price level. That is, households cannot set a trigger price above which they would like to be automated. This has implications for the interpretation of results I discuss in Section 1.6.

It is also important to note that enrolled households had the option to opt-out of certain times of the week they did not want to receive event periods. Specifically, a household could choose via the web portal which hours of the day they wanted to receive event hours for weekdays and weekends separately. For example a household could choose to receive events from 5PM-8PM on weekdays and 12PM-6PM on weekends. Households were defaulted to receive messages between 7AM-10PM for both weekdays and weekends and the majority receive events during all possible hours for this experiment (11AM-10PM). I address this event-level noncompliance directly in the empirical analysis presented in Section 1.4.

1.3.3 Treatments

Upon recruitment, households were randomized into three treatment groups: Control (C), Standard (S), and Encouraged (E). Each treatment was exposed to a unique product experience for the first 90 days of enrollment as follows:

- **Control (C):** 1,043 households were informed that their enrollment would be delayed for 90 days. These households did not receive any event messages or other contact during this time.
- **Standard (S):** 2,246 households received *price treatments* and were allowed to connect their smart devices to automate their response.

- **Encouraged (E):** 2,202 households received the same *price treatments* as the Standard group, but also received an *automation encouragement* in the form of a rebate for the full cost of purchasing a smart home device connecting it to the automation service.

Households were probabilistically assigned to C/S/E with a ratio of 20/40/40 percent. The assignment ratios were calculated to maximize the statistical power for comparing Standard to Encouraged, while ensuring there were enough Control households to have a minimum detectable effect of 1−2 percent between pricing levels. The product experience randomization occurred at the time of sign-up to ensure each day's enrollment cohort had comparable households in each treatment arm. This makes the internal validity of the study robust to changes in the recruitment strategy that took place during the study.

Control Group – Control households were granted access to the web-portal, but it did not display any information about their usage and they did not receive any event messaging. Instead, upon enrollment they received an email with the following language: *"Due to overwhelming demand for our service, there will be a delay before we can send you #OhmHours. We estimate this delay will last approximately 3 months. In return for your patience, we'll issue you an extra $10 bonus when your account delay is over."* The extra bonus was meant to combat sample attrition. There is evidence of a small amount of attrition in the control group between the point of recruitment when households enter an email and are assigned and when they connect their utility data. While this poses a potential threat to identification, I show below that the sample is balanced across treatment assignment for a large set of observable characteristics.[27]

Automation Encouragement – In order to measure the causal effect of adopting an automation technology the Encouraged households were offered a rebate for the full purchase price of a new smart home device up to $240 in value. Upon creating an account, these households were shown a pop-up notification on the web-portal in addition to being sent an email notifying them they had been selected to receive a rebate for purchasing a new smart device. The household was offered a choice

[27]Random assignment happened prior to households connecting their utility account due to technical constraints within the DRP's product infrastructure. The probability of connecting utility data was 41.6 percent for control households as compared with 45.5 percent in the treatment groups, indicating a small amount of attrition in the control group. The difference is statistically significant and should be noted when interpreting the results. I report the regressions estimating attrition at the two stages of data construction in the Appendix. The results show no difference in attrition between the point of connecting utility data and having appropriate data to be in the study sample. Further, the sample is balanced in the electricity usage and temperature data used in the empirical analysis.

between 3 smart thermostats ranging in retail prices from $198 to $240 or one package of two smart plugs with a retail price of $80.[28]

Price Treatments – In order to measure the effect of the price level households were exposed to *price treatments* where the financial incentive to consume was randomly varied. During the first 90 days, events were called 1-3 times per week and households were notified with one hour ahead notice of an increase in the effective price of electricity consumption via SMS and email. The SMS message said *"Reduce consumption today from 6PM-7PM! We'll award you 100 points for every kWh you reduce below your forecast (1.35 kWh), but you'll lose points if you go over."* The email contained the same information plus additional general strategies on how to reduce consumption. For each event, the incentive level was randomly assigned within household to one of five incentive levels: 5, 25, 50, 100, and 300 points per kWh with equal probability. The only variation between message content was the event period, the price level, and the forecast.

Moral Suasion Treatments – To further investigate the degree to which message content affected responses, households were also exposed to *moral suasion treatments* instead of a financial incentive during events.[29] An additional 90 days after the completion of the *price treatments*, households in all both treatment groups were pooled.[30] Within each event, households were randomly assigned to one of four groups with equal probability. Three treatment groups received events with different messaging and a fourth served as a control. Among the treated, households were assigned to one of the following messages:

- **Moral Suasion Only:** Messages reminded households that reducing energy consumption could improve environmental outcomes instead of offering a financial incentive. An example SMS message was *"Environmental event today from 6PM-7PM! Saving energy now could keep a dirty power plant turned off!"*
- **Price + Moral Suasion:** Messages had the same financial incentive as *price only* plus language from *moral suasion only*. An example SMS message was *"Pricing event today from 6PM-7PM! Get 100 points per kWh saved below your*

[28]The households were told that they would have the purchase price equivalent of points added to their balance when they connected the device as to ensure rebates encouraged automation.

[29]Ideally, I could have tested the effect of presenting the same language as the *price treatments* with a zero price, but this was deemed to be too confusing by the DRP. The product generally features messaging on the environmental benefit of reducing, so it is possible that moral suasion is already driving consumption decisions to an extent.

[30]The 90 days between the completion of the *price treatments* and the start of the *moral suasion treatments* featured a different, cross-randomized experimental intervention. Data from days 91-180 are not included in the study.

forecast (1.35 kWh). Saving energy now could keep a dirty power plant turned off!"

- **Price Only:** Messages used the same language as the *price treatments* with an incentive level of 100 points ($1) per kWh.

Households assigned to the control group received no message or financial incentive for the hour.

Survey – The study also utilized a voluntary survey to gain further insight into the mechanisms driving the household decision. Two weeks after the *moral suasion treatments*, households were contacted via email with a link to the survey website and offered $10 to provide feedback on the product.[31] 2,250 households had been contacted with the offer to complete the survey. 587 households completed the survey—a 26 percent response rate.

1.3.4 Data and Summary Statistics

Experimental Data – The DRP was the the primary data source for the study. The company receives Advanced Metering Infrastructure (AMI) data for its customers from their respective utilities that allows me to observe each household's hourly electricity consumption during the study and up to a year prior to enrollment.[32] The company provided me with household-specific information on the experimental assignment, the time and content of messages sent to households, the type and number of devices connected to the automation service, and the financial outcome of each event. I also observe each household's 5-digit ZIP code and census block group which I use to merge in hourly temperature data and demographics on the area where the household is located.

Weather Data – I use outdoor temperature data for the centroid of each ZIP code in my sample from the Dark Sky API which combines data from multiple government sources and uses a meteorological model to predict weather for a given geocode.[33]

[31] A complete list of survey questions is provided in the Appendix along with summary statistics comparing the responders to the full sample for interpretation.

[32] This data is used to settle the company's wholesale market activity and capacity contracts so the company and the external counterparties independently verify the accuracy of the data. As a further precaution, the CEO of the company has signed a memorandum of understanding that the data has not been tampered with and the design was executed consistent with the research design described here.

[33] Information on the sources used in the Dark Sky API model are summarized here http://darksky.net/dev/docs/sources

Census Data – I observe the Census Block and Block Group for 84 percent of households.[34] I use household which I use to merge in demographics on the location of the household from the 2015 American Community Survey 5-year estimates.

Balance and Summary Statistics – Using the above data sources I construct a panel dataset where the unit of observation is a household by hour-of-sample. The company scrapes historical data on usage that allows me to observe household consumption prior to enrollment in the program. Not all households have the same amount of data available due to movers and the creation of new accounts so the panel is unbalanced.

The randomization was performed within the DRP's product infrastructure and assignment could not be feasibly stratified. I test for balance in observables characteristics between treatment groups to provide evidence on the validity of assignment. Table 3.4.1 reports summary statistics by treatment group. Column (1), (2), and (3) report means and standard deviations in parentheses for the Control, Standard, and Encouraged groups calculated using pre-enrollment data. Columns (4)-(6) report p-values on the t-test on the difference in means permuted between each of the treatment groups.[35] Rows $1-6$ report statistics for the data used in the empirical analysis and row 7 and above for the census variables, where available. Only one comparison is significantly different at the 10 percent significance level: the maximum consumption between the Encouraged and the Control group. This provides strong evidence that the assignment was random.

Figure 1.3.3 provides further evidence of balance by showing a map of where the households in the study are located in the state of California. Regions shaded green have both Enrolled (assigned to Standard or Encouraged group) and Control households. Regions with only Enrolled are shaded blue and regions with only Control are shaded orange. The map shows that there is considerable coverage over populous regions of the state and that there does not appear to be any correlation between region and assignment. The significant geographic area covered by the sample also provides rich spatial variation not available in previous dynamic pricing studies. However, it should be noted that the sample is observably different from the eligible population which likely reflects the voluntary selection into the experiment.[36]

[34] I was unable to obtain the census geographic identifiers for the outstanding 16 percent.

[35] Standard errors for the balance tests are assumed to be *iid* between households.

[36] I report more complete summary statistics on the sample and the eligible population in the Appendix.

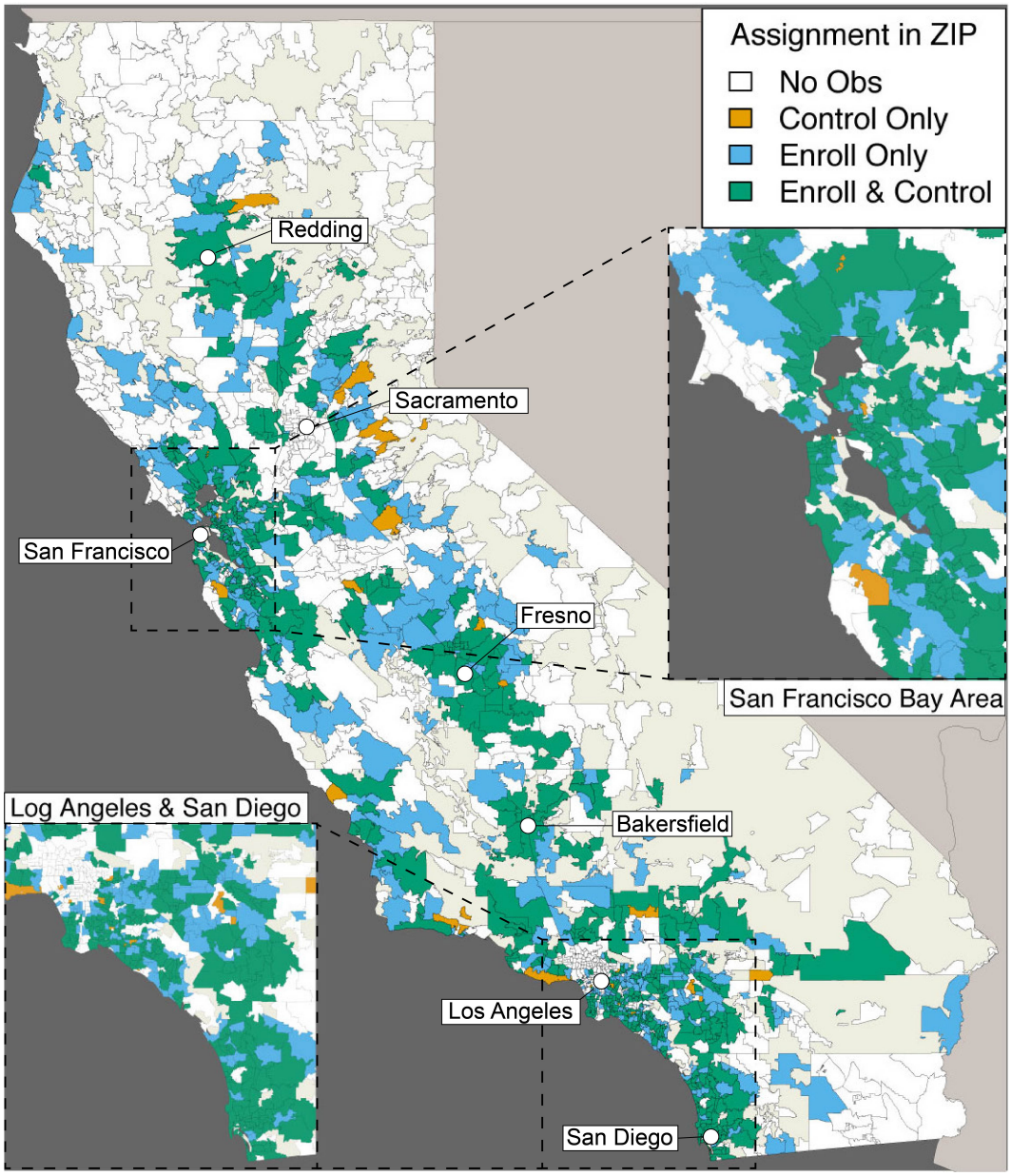
Table 1.3.1: Summary Statistics and Balance of Treatment Assignment

	Control Group	Treatment Groups		p-value on t-test with H_0:		
		Standard	Encouraged	$S = C$	$E = C$	$E = S$
	(1)	(2)	(3)	(4)	(5)	(6)
Energy & Weather						
Daily Cons. (kWh)	16.5	16.0	16.0	0.23	0.21	0.94
	(11.2)	(10.8)	(10.7)			
Max Cons. (kWh)	5.0	4.9	4.8	0.14	0.09*	0.71
	(2.7)	(2.6)	(2.7)			
Mean Daily CDHs	33.8	34.0	33.4	0.90	0.69	0.51
	(30.6)	(31.8)	(31.7)			
Mean Daily HDHs	113.8	113.7	115.5	0.98	0.27	0.15
	(41.5)	(41.8)	(43.2)			
Pre-Enroll Obs.	8755.6	8720.5	8686.5	0.83	0.68	0.79
	(4342.7)	(4362.4)	(4399.8)			
Census Block Group Demographics						
% HH Inc. < $25K	18.8	19.2	19.4	0.59	0.40	0.69
	(14.2)	(13.7)	(13.8)			
% Age 21-39	30.1	29.6	29.8	0.36	0.56	0.68
	(11.8)	(12.1)	(12.1)			
% Family HHs	66.3	66.1	66.7	0.85	0.66	0.42
	(19.1)	(19.0)	(18.5)			
% Pop. w/ Bach.	22.8	22.5	21.8	0.68	0.14	0.18
	(13.0)	(13.0)	(12.4)			
Med. Year Built	1973.5	1973.5	1973.0	0.92	0.54	0.50
	(17.2)	(16.9)	(16.9)			
% Renters	48.7	48.3	48.4	0.79	0.81	0.98
	(26.7)	(26.0)	(26.3)			
% Detached Units	54.2	54.2	54.4	0.99	0.91	0.89
	(33.9)	(33.0)	(32.2)			
% Elec. Heating	28.2	27.9	27.2	0.80	0.31	0.34
	(17.3)	(17.8)	(17.0)			
Households	1,043	2,246	2,202			

Standard deviations in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The table reports summary statistics by treatment assignment using household-level data. The energy, temperature, and observation counts are calculated by averaging over hourly variation within a household using pre-enrollment energy and weather data. The demographic statistics are calculated using a subset of the study participants for whom we have census block group designations. The census variables are constructed from block-group-level data from the 2015 American Community Survey 5-year estimates. Columns (1)-(3) report the means for each treatment group with standard deviations in parentheses. Columns (4)-(6) report p-values on the hypothesis tests for the difference in means permuted between each treatment group with C =Control, S = Standard, and E = Encouraged.

Figure 1.3.3: Enrollment and Assignment by ZIP Code



The map plots the ZIP codes with households participating in the study. Regions are shaded green have both Enrolled (assigned to Standard or Encouraged group) and Control households. Regions with only Enrolled are shaded blue and regions with only Control are shaded orange.

1.4 Empirical Strategy and Results

I present the empirical results of the *price* and *moral suasion treatments* in this section. Before proceeding with the statistical analysis, I report summary statistics on the events to assist in interpretation. There were a total of 101 events called between January 1, 2017 and November 15, 2017. The average number of events called per household during the first 90 days was 26 with a standard deviation of 4. The minimum number of messages sent was 7 and the maximum was 30. Table 1.4.2 reports summary statistics for the events by hour-of-day. Column (1) reports the number of household-hour observations. Column (2) shows that the majority of events were called between 1pm and 8pm[37] Column (3) shows the proportion of households called during any given event was 0.89. I address this imperfect compliance in my analysis below. Since I did not have explicit control over price randomization, I verify that the average incentive in each event is consistent with the expected value of \$0.96/kWh.[38] Column (6) and (7) show the mean consumption for the control and treatment groups, respectively. From this we can see the enrolled households consumed less electricity than the control households using the raw experimental variation.[39]

1.4.1 Effect of Pricing Events

Pricing events reduce consumption – As a first exercise, I test whether household consumption responds to pricing events. I separate the effect of enrollment into event hours and non-event hours by estimating the following difference-in-differences (DD) regression on the unbalanced panel of hourly household electricity consumption

$$Y_{it} = \delta D_{it} \times H_t + \xi D_{it} \times N_t + X'_{it}\beta + \varepsilon_{it}. \tag{1.6}$$

The dependent variable is hourly electricity consumption in log(kWh) for household i in hour-of-sample t. I report results in log terms to assist in interpretation.[40] D_{it} is an indicator variable that household i had enrolled in the program during hour t. H_t is an indicator that an event was called during hour t and N_t is a complementary indicator that an event was *not* called during hour t. The DD specification in (1.6) is not necessary for identification, but I include a vector of controls X_{it} to improve

[37] These are the hours often associated with peak capacity and the ramping challenges of high renewables penetration.

[38] The unweighted average is calculated as $(0.05 + 0.25 + 0.5 + 1 + 3)/5 = 0.96$.

[39] I provide further information on the raw experimental variation by plotting the empirical CDFs of enrolled and control households in the Appendix.

[40] Results in terms of kWh are reported in the Appendix and are qualitatively similar.

Table 1.4.2: Event Period Summary Statistics

Event Period	Obs.	Events Called	Frac. Called	Mean Incentive ($/kWh)	Mean Temp. (°C/°F)	Consumption for: Control (kWh)	Consumption for: Treated (kWh)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
11:00-12:00	3,838	2	0.89	0.97	22.2 (71.9)	0.80	0.68
12:00-13:00	8,061	4	0.90	0.94	25.3 (77.6)	0.93	0.76
13:00-14:00	25,152	14	0.90	0.96	27.4 (81.3)	1.04	0.84
14:00-15:00	22,444	13	0.90	0.96	25.5 (77.9)	1.09	0.88
15:00-16:00	13,338	8	0.88	0.97	28.7 (83.7)	1.27	1.01
16:00-17:00	24,645	18	0.89	0.98	25.6 (78.2)	1.06	0.87
17:00-18:00	19,345	12	0.90	0.96	24.3 (75.8)	1.14	0.95
18:00-19:00	11,256	9	0.89	0.96	19.3 (66.8)	0.83	0.74
19:00-20:00	16,329	12	0.89	0.96	21.5 (70.8)	1.23	1.02
20:00-21:00	9,570	8	0.89	0.95	18.7 (65.6)	1.09	0.99
21:00-22:00	1,425	1	0.89	0.95	11.2 (52.2)	0.71	0.69
All Events	155,403	101	0.89	0.96	24.4 (76.0)	1.08	0.89

This table reports summary statistics for the pricing events by hour-of-day called. Column (1) reports the number household-hour observations, column (2) reports the total number of events called, column (3) reports the fraction of treated households sent messages, column (4) shows the average $/kWh offered each event, and column (5) shows the mean temperature. Column (6) and (7) shows the mean consumption in kWh for control and enrolled households, respectively.

the precision of the estimates. These include household by hour-of-day fixed effects to remove time-invariant household factors for each hour of the day, hour-of-sample fixed effects to remove factors that effect the entire sample each period, and a flexible function to control for outdoor temperature at household i's location in period t.[41] ε_{it} is the structural error term.

The primary coefficient of interest in (1.6) is δ and captures the average treatment effect (ATE) of being enrolled in the program during a pricing event on electricity consumption. Since there are multiple pricing events called while a household is enrolled, the effect averages variation *within* and *between* households. Another coefficient of interest is ξ, which captures the ATE of being enrolled on consumption during all non-event hours. I interpret estimates of δ and ξ causally because they are identified by comparing enrolled households to the pure control households that randomly experience a delay. Formally, I assume $\mathrm{E}[\varepsilon_{it}|D_{it}] = 0$ because enrollment is randomly assigned.

While the price level was randomly assigned among households that received messages, not all enrolled households were called each time a pricing event occurred. On average, about 11 percent of households were not called each event. Event noncompliance was driven by a household's decision to opt-out of certain times of the week they did not want to receive events. This poses an identification challenge because I cannot observe who would have opted-out in the control group. Importantly, households decide to opt-out before they are called with an event so selection is not driven by the price level.

The event noncompliance means δ in (1.6) can be thought of as the average intent-to-treat (ITT) effect where treatment is defined as receiving a pricing event. While the ITT is important for the overall evaluation of the program, I also estimate the local average treatment effect (LATE) for households who received messages to characterize the behavioral response to pricing events.[42] In order to estimate the LATE, I modify (1.6) and estimate the following regression:

$$Y_{it} = \eta C_{it} \times H_t + \xi D_{it} \times N_t + X'_{it}\beta + \varepsilon_{it} \tag{1.7}$$

where C_{it} is an indicator variable that household i was contacted with a change in marginal price during period t. Estimates of η obtained using OLS cannot be interpreted causally due to the selection bias potentially introduced by households

[41] In my preferred specification I parametrically control for linear dependence on cooling-degree hours (CDH) and heating-degree hours (HDH) separately. CDHs and HDHs are defined as the deviations in outdoor temperature above and below 20°C, respectively. Estimating the model with a quadratic control function in outdoor temperature does not change the results.

[42] Note this is a different LATE than for the automation compliers, which I discuss below.

opting-out of certain event time periods. To address the identification issue, I estimate (1.7) using 2SLS and instrument the first term $C_{it} \times H_t$ with the randomly assigned enrollment indicator D_{it}. The resulting estimate of η captures the LATE of receiving a pricing event on consumption for the households who do not opt-out. Interpreting the LATE causally relies on the same identifying assumption that $\mathrm{E}[\varepsilon_{it}|D_{it}] = 0$ so that enrollment only affects behavior during events through the calling of pricing events.[43]

Table 1.4.3: Effect of Pricing Events

	OLS	OLS	IV
	(1)	(2)	(3)
Panel A: Effect Estimates			
Enrolled×Event ($D_{it} \times H_t$)	-0.117^{***}		
	(0.009)		
Called×Event ($C_{it} \times H_t$)		-0.131^{***}	-0.132^{***}
		(0.009)	(0.010)
Enrolled×Non-event ($D_{it} \times N_t$)	-0.017	-0.017	-0.017
	(0.006)	(0.006)	(0.006)
Panel B: 1st Stage			
Enrolled×Event ($D_{it} \times H_t$)			0.889^{***}
			(0.005)
Households (HHs)	5,491	5,491	5,491
Observations	22,926,631	22,926,631	22,926,631

SEs clustered by HH and hour-of-sample in parentheses.

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

Panel A reports estimates of the effect of enrollment during pricing event periods and non-event periods on hourly energy consumption measured in log(kWh). Each column report coefficients estimated from separate regressions and all include temperature controls, hour-of-sample fixed effects, households by hour-of-day fixed effects, and pre-enrollment data. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. Column (1) reports estimates for the DD specification in (1.6). Column (2) reports estimates for the TOT for called households using OLS. Column (3) reports the 2SLS estimates where called being called is instrumented with enrollment. Panel B reports the first stage coefficient on the instrument.

Table 1.4.3 reports estimates of the ITT and LATE of pricing events on log electricity consumption obtained from estimating (1.6) and (1.7), respectively. Each

[43]Identification also relies on a significant first stage (indicated by the Panel B of Table 1.4.3), monotonicity of the instrument, and SUTVA to hold.

column reports coefficient estimates from separate regressions which include fixed effects and controls. Standard errors are reported in parentheses and clustered by household and hour-of-sample to account for arbitrary correlated errors within household and across sample within each time period.[44] For the remainder of Section 1.4, standard errors and confidence intervals are estimated using this structure unless otherwise noted. Column (1) reports the estimates of the ITT and columns (2) and (3) report the estimates of the LATE using OLS and 2SLS, respectively.[45]

The results in column (1) of Table 1.4.3 indicate average consumption falls by 0.117 log points (11.0 percent) during pricing events as a result of enrolling in the program. The estimate for the non-event periods is reported in the third row and indicates enrollment reduces electricity consumption by 0.017 log points (1.7 percent), but is not significantly different from zero. Column (3) shows calling households with an event reduced consumption by 0.132 log points (12.4 percent) and column (2) shows the LATE is not substantively different from the OLS estimation. This is the first main result of the paper. The LATE is slightly larger than the ITT which follows from the intuition that the ITT is attenuated by households whose consumption was closer to the controls because they did not receive an event. Panel B reports the coefficient on the instrument from the first-stage for the regression in column (6), which confirms the noncompliance rate of 11 percent.

Event responses are precise with no substantial spillovers – I offer graphical evidence of the consumption behavior around an event by using an event-study specification to characterize changes in energy consumption in the 8 hours leading up to and following a pricing event. Specifically, I estimate the following regression:

$$Y_{it} = \sum_{\tau=-8}^{8} \delta_\tau D_{it} \times \mathbf{1}(H_{t+\tau} = 1) + \xi D_{it} \times \mathbf{1}(t \notin \mathcal{H}) + X_{it}'\beta + \varepsilon_{it}. \tag{1.8}$$

[44]The first dimension of clustering are robust to arbitrary correlated errors within households. This is meant to account for serially dependent errors hourly electricity consumption. The second dimension of clustering are robust to arbitrary correlation within an hour-of-sample. This is meant to account for correlated weather shocks in time across the region due to weather or grid operations. The results are not qualitatively affected by clustering at the day-of-sample level, which is significantly more conservative.

[45]In the Appendix I show how adding controls, fixed effects, and pre-enrollment data change the point estimates in the Appendix. The results are qualitatively similar for all regressions and the DD specification is preferred because it controls for time-invariant characteristics of households. These may be systematically different between treatment groups by chance as I was unable to stratify assignment. The DD specification is more robust to these concerns than the raw experimental variation.

The coefficients of interest are the δ_τ's that capture the average change in electricity consumption caused by enrolling in the program in the period τ hours relative to an event. I also include an indicator for all post-enrollment periods that do not lie within $\mathcal{H}$, the set of the sample periods within an 8 hour neighborhood of any event.

Figure A.7.3a plots the results from estimating (1.8) using OLS with log consumption as the dependent variable. The event period is shaded in dark gray. The plot shows point estimates for δ_τ using the period two hours prior to an event (-2) as the reference category. Households are notified of an event during the hour before an event (period -1) which is shaded light gray. The vertical bars plot 95 percent confidence intervals.

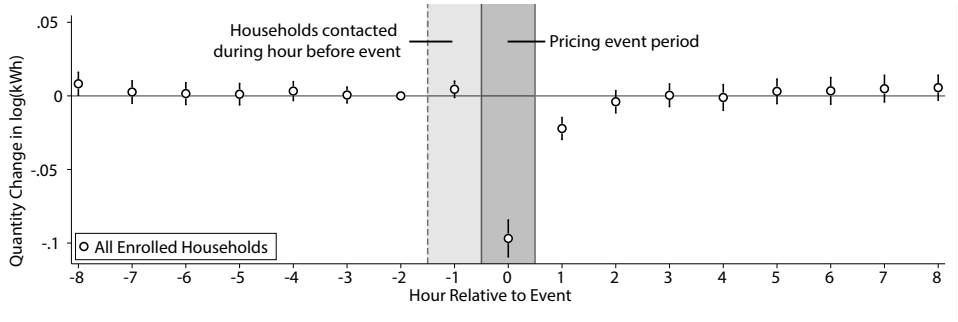
The results show that being enrolled in the program during pricing event period causes electricity consumption to decrease by 0.097 log points (9.2 percent) relative to the period two hours before an event. Further, there are no distinguishable pre-trends in consumption and the reversion to the pre-event consumption level happens quickly. By the second hour following the event, electricity consumption is statistically indistinguishable from pre-event levels. This is notable because there is a symmetric price change in the periods after the event that provide a second test of household's responsiveness to price changes. This provides evidence that, on average, the *price treatments* caused households to reduce their electricity consumption.

Households reduce with and without automation – Since automation is a feature of the setting, I estimate effect heterogeneity by whether households chose to automate. These estimates provide non-causal information on the degree to which automated versus manual decisions explain the results. The comparison is non-causal because households who choose to adopt the automation technology are likely to be different than those who do not in ways the econometrician cannot observe. For instance, they may engage in additional conservation behaviors during pricing periods if they are motivated by environmental concerns or they may have different preferences over appliance usage.

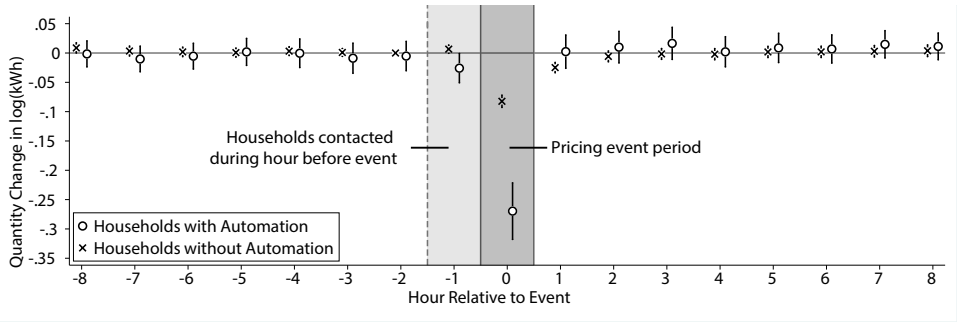
To investigate heterogeneity, I estimate the same event-study specification as (1.8) using OLS except I interact the enrollment indicator an indicator i household was automated in period t, denoted A_{it}. The variable is constant post-enrollment and zero in the pre-enrollment period for all households. Figure A.7.3b plots the results from estimating the effect of enrollment event-study and decomposing the effect for automated versus non-automated households. I estimate the coefficients for both groups jointly and omit the period two hours prior to the event for non-automated households as the reference category. Point estimates for automated and non-automated households are plotted with "o" and "x" markers, respectively.

Figure A.7.3b shows automated households reduce consumption significantly more during pricing periods. The point estimate for automation is 0.267 log points (23.4

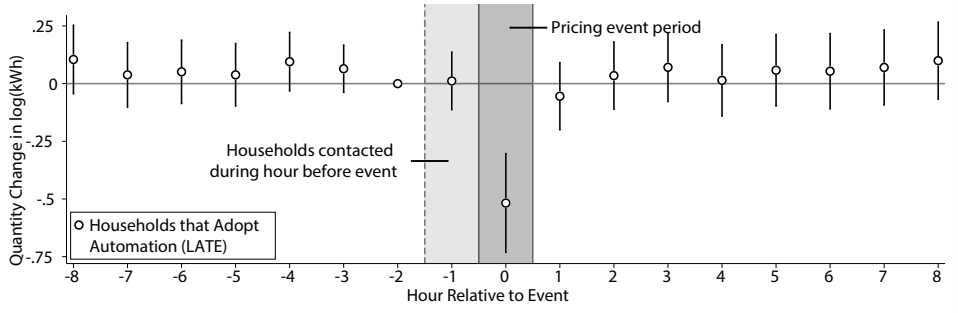
Figure 1.4.4: Pricing Hour Event Studies



(a) Enrolled vs. Control Households



(b) Automated vs. Non-Automated Households



(c) Additional Effect of Automation for Enrolled Households

Note the vertical axis scale varies between the three figures. Panel (a) plots event-study point estimates of the ITT effect that compares the consumption of enrolled households to control households in the hours leading up to and following a pricing event (shaded dark gray). Panel (b) plots the non-causal decomposition of the ITT estimates in (a) by whether households automated their decisions. Panel (c) plots event study estimates of the LATE of adopting automation on consumption during pricing periods. Households are contacted during the hour prior to the event (shaded light gray). I use the period two hours prior as the reference category so the vertical axis is the quantity change relative to period -2 in log points. For (b) the reference category is period -2 for non-automated households. Vertical bars indicate 95 percent confidence intervals clustered by household and hour-of-sample.

percent) lower consumption relative to the two hours prior for non-automated households and 0.273 log points (23.9 percent) lower relative to non-automated households during pricing events. The consumption for non-automated households is 0.082 log points (7.9 percent) lower than pre-event consumption. The results also indicate that there are no significant pre-event differences in trends between automated and non-automated households.

Two important results are worth noting. First, consistent with expectations, households who choose to automate their decisions are significantly more responsive than those who do not. Second, the non-automated households still respond significantly to pricing events. Only 29 percent of the of the ITT effect is attributable to automation, suggesting households are making active decisions to reduce consumption during pricing periods.[46] Automation contributes only a small amount to the effect of enrollment because only 8 percent of all enrolled households adopt the technology.

1.4.2 Effect of Marginal Price

Responses are very insensitive to the marginal incentive – The estimate of η in (1.7) pools the effect of receiving a message with any positive price level. Define ΔMP_{it} as the change in marginal price household i would receive in period t regardless of whether they were called or not.[47] To recover the effect of the level of the marginal price change on energy consumption, I modify (1.7) by interacting C_{it} with a vector of indicators for each price level and estimate the regression

$$Y_{it} = \sum_{s \in \mathcal{S}} \eta_s C_{it} \times \Delta MP_{it}^s \times H_t + \xi D_{it} \times N_t + X_{it}'\beta + \varepsilon_{it} \tag{1.9}$$

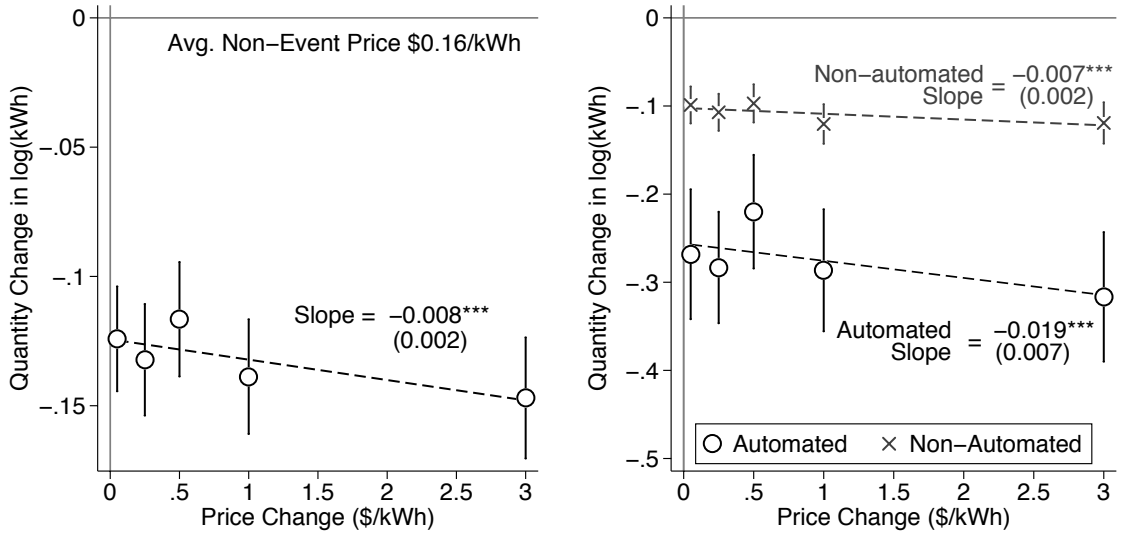
where $\Delta MP_{it}^s \equiv \mathbf{1}(\Delta MP_{it} = s)$ and $\mathcal{S} \equiv \{0.05, 0.25, 0.5, 1, 3\}$ is the set of price levels. To deal with the selection issue, I estimate (1.9) using 2SLS and instrument each interaction term $C_{it} \times \Delta MP_{it}^s \times H_t$ with $D_{it} \times \Delta MP_{it}^s \times H_t$, the randomly assigned

[46] I come to this number by subtracting the non-automated component from the enrollment effect and dividing it by the total effect: $(0.117 - 0.082)/0.117 = 0.29$.

[47] There is a subtle technical point here regarding the generation of the ΔMP_{it} variable. Households were not assigned a price level by the DRP until they were called, so the raw experimental data does not have a price level for households who were not called. To populate these, I randomly assign a counterfactual marginal price to the opt-out households consistent with the design. This can be thought of as randomly assigning prices for all future events at the time of enrollment, but not informing the household. Households decide generally whether to opt out of the program during certain hours, but since they do not observe the price assignments, the opt-out decision is independent of the price level of a given event.

enrollment indicator interacted with the randomly assigned price level indicator. Again the key identifying assumption is that the enrollment only affects behavior through the calling of pricing events during those periods. The coefficients of interest in (1.9) are $\{\eta_s\}_{s\in\mathcal{S}}$ which recover the TOT of receiving a message with marginal price s in points per kWh on electricity consumption for the called households.

Figure 1.4.5: Effect of Marginal Price



(a) Effect of Price Change in log(kWh)

(b) Auto vs. Non-Auto in log(kWh)

The figure plots the point estimates for the change in the conditional mean of consumption as a result of changes in the effective price of electricity in dollars per kWh. Note the difference in the scales of the axes between the left and right panels. The left panel shows the effect with log consumption on the vertical axis and the price change on the horizontal axis. The right panel shows the treatment effect heterogeneity by households with and without automation. Vertical bars in both panels show 95 percent confidence intervals estimated using standard errors clustered by household and by hour-of-sample. The dashed line plots the linear parametric estimation of the response as a function of the price change. The figure also reports the estimates for the slope and intercept of the parametric estimation with two-way clustered standard errors in parentheses.

Figure A.13.11d plots estimates of (1.9) using the 2SLS estimator described in the previous paragraph and log(kWh) consumed as the dependent variable. The vertical axis represents the change in consumption relative to the control group in log points and the horizontal axis represents the price change in $ per kWh equivalent and ranges from $0-$3/kWh. The leftmost point estimate indicates calling an event with

a \$0.05/kWh effective price increase caused an average reduction in consumption of 0.124 log points (11.7 percent). The rightmost point estimate indicates an event with a \$3/kWh effective price increase caused an average reduction of 0.147 log points (13.7 percent). Vertical bars plot 95 percent confidence intervals.

I can confidently reject the null that the effects for the smallest and largest incentives are equal ($\eta_{0.05} = \eta_3$) because the p-value is less than 0.01. However, the change in effect size is extremely small given the difference in incentive levels. The results indicate a 60-fold increase in the price increase yields 2.0 percentage points lower consumption, an increase in the size of the treatment effect of only 17 percent. This is remarkable since the average utility price for households is approximately \$0.16/kWh during non-event periods. This is the second main result of the paper.

The dashed line in Figure A.13.11d plots the results of a parametric estimation of the price responsiveness with an intercept for any event being called and a slope for the incentive level within an event.[48] The slope can be interpreted as indicating a \$1/kWh increase in the level of price change causes an additional 0.008 log points (0.008 percentage points) reduction in electricity consumption. The estimate is statistically significant from zero with a p-value less than 0.01. The 95 percent lower confidence interval indicates we can reject slopes more negative than -0.012 which translates roughly to a demand elasticity with respect to the dynamic price level of no larger than -0.001.

Figure A.13.11c shows the effect of marginal price decomposed by automation status. Automated and non-automated households are plotted with "o" and "x" markers, respectively. I also plot the parametric estimation of the price responsiveness within each group. The differences cannot be interpreted causally, but the figure provides evidence that the price insensitivity is not driven mechanically by the fact that the automation service does not depend on price. In fact, automated households appear to be marginally more price responsive than non-automated households, but this could be driven by selection.

Households respond less to messages without a price change – To investigate the degree to which the financial incentive in messages drives the results, households were randomly assigned messages with no price change during the *moral suasion treatments*. The interventions happened 90 days after households completed the *price*

[48] Specifically I estimate the regression

$$Y_{it} = \eta^{int} C_{it} \times \mathbf{1}(\Delta MP_{it} > 0) \times H_t + \eta^{slope} C_{it} \times \Delta MP_{it} \times H_t + \xi D_{it} \times N_t + X'_{it}\beta + \varepsilon_{it}$$

with 2SLS and instrument the first two terms with $D_{it} \times \mathbf{1}(\Delta MP_{it} > 0) \times H_t$ and $D_{it} \times \Delta MP_{it} \times H_t$, respectively. The dashed line in Figure A.13.11d plots $\widehat{Y}_{it} = \hat{\eta}^{int} + \hat{\eta}^{slope} \Delta MP$ using the estimated parameters.

treatments (180 days after enrollment). Thus, the sample for these interventions is composed of 1,255 households who signed up early enough in the program to reach 180 days enrolled. The sample for these interventions may not be representative of the full sample because recruitment in time was not random, so the *moral suasion treatments* included messages identical to the *price treatments* to benchmark the results. The benchmark results for messages with price are generally consistent across the two treatments, suggesting the results are not substantially confounded with selection in time of recruitment.

To recover the effect of the message content on electricity consumption I estimate the model

$$Y_{it} = \delta_S \, C_{it}^{Suasion} + \delta_P C_{it}^{Price} + \delta_B C_{it}^{Both} + X_{it}'\beta + \varepsilon_{it} \tag{1.10}$$

where $C_{it}^{Susaion}$, C_{it}^{Price}, and C_{it}^{Both} are indicator variables for whether the household received a message with moral suasion and no price change, a standard message with the price change, and a message with both, respectively. I use the same pre-enrollment data as the *price treatments*, but limit the post-enrollment data to periods when a *moral suasion treatment* event occurred. The coefficient δ_S captures the ATE of receiving a message with moral suasion and no financial incentive relative to receiving no message. δ_P captures the ATE of receiving a message with the same language as the *price treatments* and an incentive level of \$1/kWh. δ_B captures the ATE of receiving a message with moral suasion and an incentive level of \$1/kWh. I include the same controls as the *price treatments* estimation.

Table 1.4.4 reports the results from estimating (1.10) with log(kWh) as the dependent variable in column (1). The first coefficient of the first column indicates that sending a message with environmental priming language and no price change causes households to reduce consumption by 0.023 log points (2.3 percent) relative to being enrolled and receiving no event, but the result is not statistically significant. The second coefficient indicates that sending a message with a price incentive of \$1/kWh causes households reduce consumption by 0.062 log points (6.0 percent). The third coefficient indicates a message with a price incentive of \$1/kWh that also includes moral suasion causes households to reduce consumption by 0.074 log points (7.1 percent). We can confidently reject the null that the messages with price are significantly different from zero at the 1 percent level.

Column (2) of Table 1.4.4 breaks out the effect by households with and without automation. The coefficients can be interpreted as the causal effect of sending each message type on consumption for automated and non-automated, separately. Comparing automated to non-automated responses cannot be interpreted causally and the encouragement instrument does not have enough statistical power with the

Table 1.4.4: Effect of Price vs. Moral Suasion (No Price)

	(1)	(2)	
Panel A: Effect Estimates			
Moral Suasion Only (δ_S)	-0.023		
	(0.016)		
× Non-Automated		-0.003	
		(0.016)	
× Automated		-0.202^{***}	
		(0.056)	
Price Only (δ_P)	-0.062^{***}		
	(0.017)		
× Non-Automated		-0.044^{**}	
		(0.017)	
× Automated		-0.208^{***}	
		(0.059)	
Moral Suasion + Price (δ_B)	-0.074^{***}		
	(0.019)		
× Non-Automated		-0.046^{**}	
		(0.019)	
× Automated		-0.311^{***}	
		(0.057)	
Panel B: Coefficient Tests	*Pooled*	*Non-automated*	*Automated*
H_0: $\delta_S = \delta_P$	0.002	0.002	0.886
H_0: $\delta_S = \delta_B$	0.001	0.004	0.010
H_0: $\delta_P = \delta_B$	0.303	0.886	0.008
H_0: $\delta_S + \delta_P = \delta_B$	0.619	0.951	0.139
Households	1,255	1,255	
Observations	2,579,728	2,579,728	

SEs clustered by HH and hour-of-sample in parentheses.

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

Panel A reports estimates of the casual effect of message content from the *moral suasion treatments* on hourly energy consumption measured in log(kWh). Each column report coefficients estimated from separate regressions and all include temperature controls, hour-of-sample fixed effects, households by hour-of-day fixed effects, and pre-enrollment data. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. Column (1) reports the model estimated using the DD specification in (1.10). Column (2) reports estimates for the same model interacting each message type indicator with automation status. Estimates comparing within automation type can be interpreted causally. Panel B reports p-values on tests of whether treatments are equivalent and additive. δ_S representing the parameter for moral suasion only, δ_P representing the parameter for price only, and δ_B representing the parameter for both interventions together. Column (1) reports the p-values for the coefficients that pool automated and non-automated. Column (2) breaks out the p-values by automated and non-automated.

smaller sample. Still the results confirm intuitive patterns. Households that made decisions manually responded differently to price as compared with moral suasion. Panel B reports the p-value for the hypothesis test moral suasion and price alone have the same effect is equal to 0.002 for pooled and non-automated households. Automated households on the other hand show significant effects for all three message types. This provides evidence in favor of that price is driving the impact on non-automated active consumption decisions rather than environmental altruism.

1.4.3 Effect of Automation

Adopting automation causes larger reductions – To estimate the causal effect of adopting the automation technology on household energy consumption I use a randomized encouragement design within the subsample of enrolled households. Specifically, I randomly select half of enrolled households to receive rebate to take-up the automation technology. The encouragement provides an instrument to identify the causal effect of adopting automation.

Consider a model that only uses enrolled households. Using 2SLS, I estimate the regression

$$Y_{it} = \alpha_H A_{it} \times H_t + \alpha_N A_{it} \times N_t + X'_{it}\beta + \varepsilon_{it} \tag{1.11}$$

instrumenting A_{it} with Z_{it}, an indicator that household i received the encouragement offer prior to period t.[49] The coefficient of interests is α_H, which captures the LATE of automation on consumption during events for the compliers nudged into adopting by the encouragement. The coefficient α_N captures the LATE of automation on consumption during non-events, which may be affected by spillovers. Importantly, the effects are additional to the level of being enrolled without automation, but can be interpreted causally if the instrument is valid.[50]

[49]Since households receive the encouragement upon enrollment, $Z_{it} = D_{it}$ if $Z_{it} = 1$.

[50]Specifically, the effect can be interpreted causally if four assumptions hold. First, the encouragement must affect uptake of the automation technology, which I verify below. Second, the encouragement must not affect consumption in any way except through the automation technology. I cannot formally test this assumption's validity and if the offer of a rebate caused households to change their investments in other electricity consuming durables, this may be a concern. This omitted variation could lead to bias in the estimates of the LATE and should be noted when interpreting results. The sign of this bias is uncertain. For example, if the rebate offer causes the household to invest in other smart home appliances, the overall efficiency of the home could increase if new appliances are more energy efficient than the older ones. Conversely, the new appliances could have more features and consume more energy or the new appliance could lead to higher utilization rates. Third, the stable treatment unit value assumption (SUTVA) must hold. This may be violated if

Table 1.4.5: Effect of Encouragement on Automation Take-up

	Standard (1)	Encouraged (2)	Difference (3)
Panel A: Types and Counts			
Households (HHs)	2,246	2,202	-
HHs w/ 1+ device	122	218	96
Connected devices	295	554	259
Thermostats (subsidized)	143	210	67
Plugs (subsidized)	143	327	184
Home systems (not subsidized)	0	10	10
Electric Vehicles (not subsidized)	8	5	-3
Panel B: Take-up			
HHs w/ any automation	0.053	0.097	0.045***
			(0.008)
Take-up by consumption:			
1st quartile (0-0.34kWh)	0.022	0.103	0.080***
			(0.015)
2nd quartile (0.34-0.55kWh)	0.064	0.103	0.039**
			(0.016)
3rd quartile (0.55-0.85kWh)	0.066	0.098	0.031*
			(0.016)
4th quartile (0.85-6kWh)	0.060	0.091	0.031*
			(0.016)

SEs in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

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

Panel A reports counts for the home devices connected to the company's automation service in the Standard enrolled and Encouraged groups. Panel B reports the proportion of households who took up automation and the difference between the two treatment groups. It also reports the proportion take-up by pre-enrollment consumption quartile. The differences are calculated by regressing an indicator for any automation on the encouragement indicator. Standard errors are reported in parentheses and observations are assumed to be independent.

Table 1.4.5 reports the summary statistics for the number and type of connected automation devices in the Standard enrolled and Encouraged groups. Panel A reports counts for the number and types of devices connected in each group. The results show that 96 more households connected at least one device in the Encouraged group and that they connected 259 more devices. The encouragement rebate was offered for thermostats and smart plugs, which both increased significantly in number in the Encouraged group. There is also an increase in automated home systems among encouraged participants, despite no rebate offer, which suggests potentially small complementarities. Electric vehicle connections were not incentivized and and show little difference in take-up between the groups.

Panel B reports the proportion of households who took up at least one device by the treatment group. Uptake is 9.7 percent in the Encouraged group and 5.3 percent in the Standard group. The difference is statistically significant with a p-value that is less than 0.0001. This shows the encouragement caused higher adoption in the automation technology. Panel B also reports uptake broken out by pre-enrollment consumption quartile to provide information on the type of household taking up automation. Column (1) shows that households with the lowest consumption are less likely to connect to the automation service when they are not offered the rebate. Offering the rebate increases take-up by 8.0 percentage points for the bottom quartile as compared with 3.1 percentage points in the highest quartile. This is important for interpreting the LATE since it provides evidence that the always-takers are observably different from the compliers.

Table 1.4.6 reports the results for estimating (1.11) with log electricity consumption as the dependent variable using only enrolled households. Column (1) reports the OLS estimates of the average effect of automation on post-enrollment consumption broken out by event and non-event periods. The results show estimates of the difference between automated and non-automated households consistent with the difference shown in Figure A.7.3b. Column (2) reports the reduced form specification, regressing log consumption on the encouragement instrument directly, interacted with the even-period indicators. Column (3) reports the IV estimation. The first stage is reported in Panel B with both regressions collapsed to a single column. I report the coefficients for the first-stage and instrument corresponding to the endogenous variable by event and non-event period. The other coefficients are statistically insignificant from zero. Figure A.7.3c plots the point estimates from the analogous

social interactions cause spillovers between treatment and control, but seems unlikely given the low uptake overall. Fourth, the encouragement effect on uptake must be monotonic such that there are no “defiers" who would adopt automation, but choose not to as a result of the rebate.

Table 1.4.6: Effect of Automation

	OLS	OLS	IV
	(1)	(2)	(3)
Panel A: Effect Estimates			
Adopted × Event ($A_{it} \times H_t$)	-0.222^{***}		-0.668^{***}
	(0.029)		(0.160)
Adopted × Non-Event ($A_{it} \times N_t$)	-0.005		-0.112
	(0.015)		(0.127)
Encouraged × Event ($Z_{it} \times H_t$)		-0.048^{***}	
		(0.011)	
Encouraged × Non-Event ($Z_{it} \times N_t$)		-0.007	
		(0.008)	
Panel B: Selected First Stage Estimates			
Encouraged × Event ($Z_{it} \times H_t$)			0.073^{***}
			(0.009)
Encouraged × Non-Event ($Z_{it} \times N_t$)			0.074^{***}
			(0.007)
Households	4,448	4,448	4,448
Observations	18,559,930	18,559,930	18,559,930

Standard errors clustered by household and hour-of-sample in parentheses

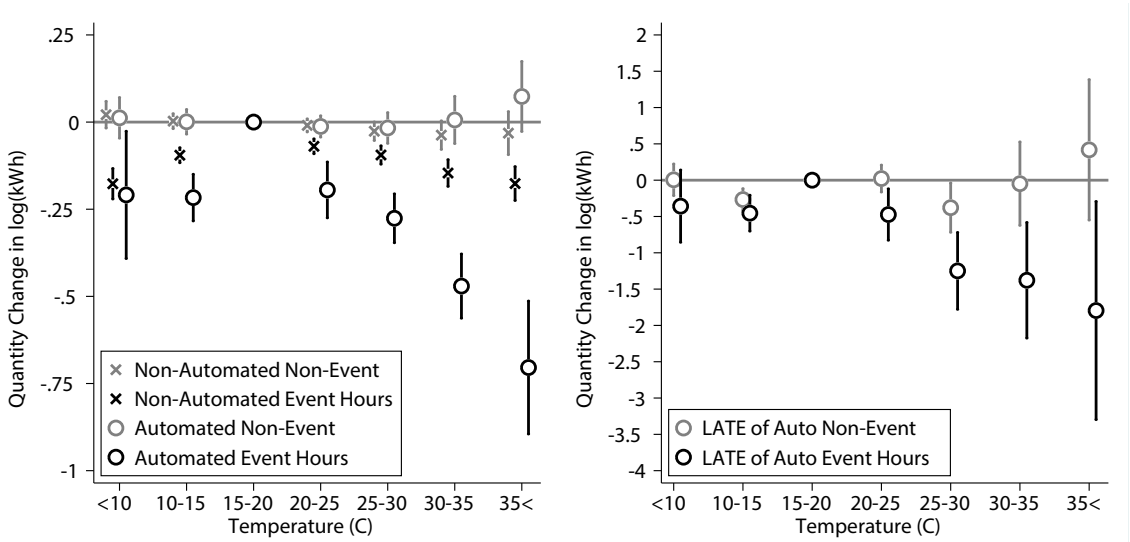
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Panel A reports estimates of the effect of automation on log energy consumption for the subsample of enrolled households (excluding controls). Each column report coefficients estimated from separate regressions and all include temperature controls, hour-of-sample fixed effects, households by hour-of-day fixed effects, and pre-enrollment data. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. All regressions include household by hour-of-day and hour-of-sample fixed effects and temperature controls. Column (1) reports the OLS estimate for the effect of automation on consumption. Column (2) reports the reduced form specification regressing consumption on the encouragement indicator. Columns (3) reports the 2SLS estimates instrumenting the automation indicator with the encouragement. Panel B reports results of the first-stage regressions for the coefficient on the instrument corresponding to the endogenous variable.

event-study specification, where the period two hours prior to the event is the omitted category.

The results show that the causal effect of automation on price response is substantial. The first coefficient in column (3) indicates automation causes 0.668 log points (48.7 percent) lower consumption during pricing events relative to the non-automated counterfactual for compliers. I cannot identify the consumption of the

Figure 1.4.6: Effect Heterogeneity by Temperature



(a) Auto vs. Non-auto HHs (non-causal) (b) Additional Causal Effect of Automation

This figure shows how *price treatment* effects change as a function of outdoor temperature. The estimates are broken out by 5°C temperature bins and normalized to the 15-20 bin which represents milder temperatures. Vertical bars show 95 percent confidence intervals constructed from standard errors two-way clustered by household and hour-of-sample. Each panel represents estimates from a separate regression. Panel (a) plots non-causal estimates of the event price response for households with and without automation. Panel (b) plots the same parameters for the LATE of adopting automation. For both panels, estimates for event (non-event) hours are plotted with black (gray) symbols. Automated households are shown with circles and non-automated households with exes in panel (a).

compliers in the Standard group with the experimental design, but if we use the price event ITT as a benchmark, the results imply 3 to 5−fold increase in responsiveness.[51] This is the third main result of the paper. These results are dramatic, but it is worth noting that the take-up due to the rebate was disproportionately driven by households with smaller consumption. Further, the fact that the OLS estimate of the difference is smaller suggests that the compliers reduce more than always-takers.

[51]If we use the pooled result across all automated non-automated, the increase is from 13 percentage points and if we only use non-automated households, the increase is from 8 percentage points.

This suggests that within the range of adoption explored here, 5 to 10 percent, the marginal treatment effect is increasing.

Decomposing the consumption effects by temperature shows that cooling load is a major margin of adjustment and automation is likely to be applied to it as well. I estimate effect heterogeneity for event and non-event hours by interacting treatment indicators with a set of temperature bin dummy variables. I omit the bin associated with the mildest climates so point estimates are interpreted relative that effect. Figure 1.4.6 shows the results of two regressions non-causally decomposing the ATE by automation in panel (a) and showing heterogeneity in the LATE estimate by temperature panel (b). The estimates show a strong association between temperature and the size of the response that increases nearly linearly among automated and non-automated households alike. Automated households show larger reductions and the LATE estimates show that adopting automation causes greater reduction at higher temperatures as well.[52] Interestingly, there do not appear to be strong responses to temperature during non-event periods, suggesting little evidence of pre-setting thermostats.

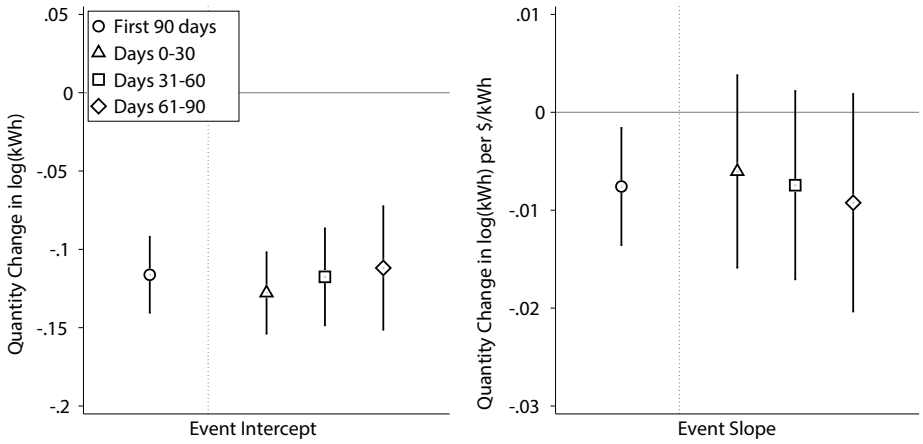
1.4.4 Persistence

This section reports results showing how the effect of pricing events changes over time. Household's understanding of the program and the technologies may change as they learn new strategies to reduce or make adaptive longer-term behaviors. In order to characterize the effect over time, I examine the *price treatments* for each 30 day period of their 90 day duration. I use the parametric specification of the response to marginal price shown in Figure 1.4.5 to see if households become more or less price elastic as they become more familiar with the product.

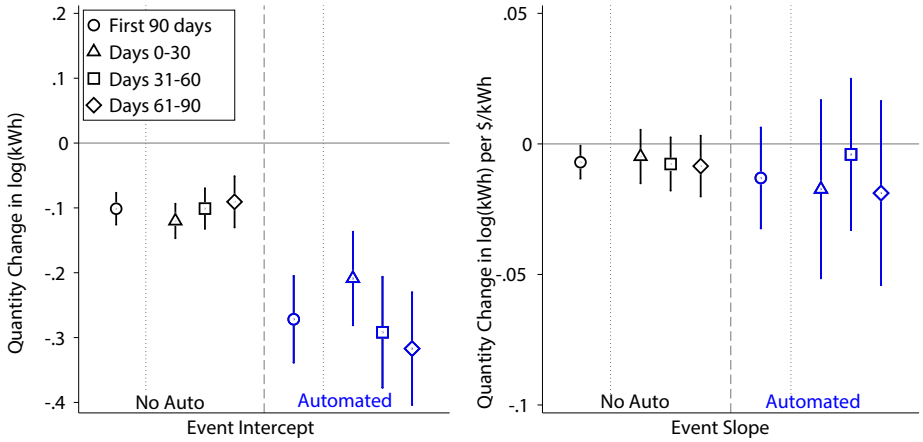
Figure 3.4.2 plots estimates for how the effects change over time during the *price treatments*. The estimates are broken out by 30 day periods during the 90 days for which *price treatments* occurred. Estimates for days 0-30, 31-60, and 61-90 are represented by triangles, squares, and diamonds, respectively and circles indicate estimates for the full 90 days for reference. Vertical bars show 95 percent confidence intervals. Figure A.7.5a shows the causal estimates of the event price response for all enrolled households. Figure A.7.5b plots the non-causal effect heterogeneity for households with and without automation in blue and black, respectively. Each panel shows estimates from a single regression. Estimates labeled “Event Intercept" and

[52]This suggests the thermostats may be driving larger responses, but the plugs could also be associated with cooling loads so we cannot say conclusively.

Figure 1.4.7: Persistence of Price Response



(a) Event Price Response Effect



(b) Automated vs. Non-automated Households (non-causal)

This figure shows how *price treatment* effects change as a function of time in the program. The estimates are broken out by 30 day periods during the 90 days for which *price treatments* occurred. Triangles represent days 0-30, squares days 31-60, and diamonds days 61-90. Circles represent the estimate for the full 90 days as reference. Vertical bars show 95 percent confidence intervals constructed from standard errors two-way clustered by household and hour-of-sample. Each panel represents estimates from a separate regression. Panel (a) plots causal estimates of the event price response and non-event enrollment response for all enrolled households regardless of automation choice. Panel (b) plots the same parameters non-causally decomposed to show differences between non-automated (black) and automated (blue) households. Estimates labeled "Event Intercept" and "Event Slope" represent the parametric estimation of the price response of consumption for called households.

“Event Slope" represent the parametric estimation of the price response of consumption for called households.

The left set of estimates Figure A.7.5a shows the intercept decreases slightly from -0.128 log points to -0.112 log points, but the estimates are not statistically distinguishable at conventional levels. The right estimates show the slope estimate does not change substantially. Figure A.7.5b shows that the results from A.7.5a are generally explained by households without automation, suggesting there does not appear to be significant learning or habituation during the 90 day period.[53]

1.5 Model Application

The salient presentation of a random price change provides the rare opportunity to estimate the elasticity of household electricity demand using experimental price variation. However, the constant elasticity form commonly used by the literature does not seem wholly appropriate given the insensitivity to change size shown in the empirical results. As a result, I adapt the theoretical model developed in Section 1.2 to the experimental setting to make it empirically tractable and demonstrate the additional information necessary to recover structural demand parameters. The model application illustrates how adding attention to the model affects elasticity estimates. It also allows me to quantify the potential costs of inattention which gives a measure of the degree to which it impacts household welfare.

1.5.1 Empirical Specification

I make additional simplifying assumptions in order to make the model developed in Section 1.2 empirically tractable. First, I assume a household in any given event can be characterized by the two types introduced in Section 1.2: attentive and inattentive.[54] Second, I assume the taste parameter η is uncorrelated with the information

[53]The estimates for the automated households show the price response intercept increases significantly from days 0-30 to days 31-60, but then levels off. This is likely due to the fact that there was a lag between enrollment, purchase, delivery, and installation of the automation technologies. Further the slope for the automated households does not follow a clear pattern and while there does appear that spillovers for automated households are gradually more negative, the effects are statistically insignificant.

[54]Since I average over variation within households, this does not assume household type is time-invariant. Instead, I assume for any given event, households could be attentive. Given that I have variation in the price level within and between households, I have explored whether households exhibit time-invariant types. The regression approach of estimating individual parametric responses yields noisy responses with no clear differentiation, but the analysis is severely underpowered. 2018

costs that determine σ.[55] Third, I allow for an additive social preferences component to utility to control for the effect of moral suasion explicitly in the model.[56]

I model a household i's log consumption in period t as

$$\log(Y_{it}) = \eta\left(\sigma\log(MP_{it}) + (1-\sigma)\log(HP)\right) + \gamma G_{it} + \alpha_i + \varepsilon_{it} \tag{1.12}$$

where MP_{it} is the marginal price the household faces in period t, G_{it} is an indicator that household i was primed with moral suasion in period t, and ε_{it} is a stochastic taste shock. I also allow for a household specific taste parameter α_i which reflects time-invariant differences in tastes.

Consumption in (1.12) is governed by four choice primitives: η, σ, HP, and γ. η captures the elasticity of demand with respect to a fully attentive household. σ governs the attention of the household to the marginal price change. Given that I am averaging over households and events, the model follows the representative agent formulation where σ can be interpreted as the probability of being attentive during an event or the average fraction of attentive households. HP governs the heuristic price response of a household with a bounded information set, which is assumed to be the same for all event periods. γ governs a demand shift due to social preferences.

In order to jointly estimate all model parameters, I pool data from the *price treatments*, the *moral suasion treatments* and pre-enrollment consumption and drop post-enrollment non-event periods.[57] I use the resulting panel to estimate the following DD specification:

$$\begin{aligned}\log(Y_{it}) = {} & (\delta_{MP}\log(MP_{it}) + \delta_{HP}\log(HP) \\ & + \delta_G G_{it}) \times C_{it} \times H_t + X_{it}'\beta + \varepsilon_{it}\end{aligned} \tag{1.13}$$

where Y_{it} is consumption for household i in period t in kWh and MP_{it} is the the marginal price in dollars per kWh (including an assumed utility price of \$0.16/kWh).[58]

explores the classification of household types in more detail by characterizing individual responses using machine learning techniques.

[55]This assumption is perhaps the most unrealistic as we might assume information costs could be lower for households that are generally more price elastic. However, given that I do not have exogenous shocks to the information set, I cannot separately identify the two with my current design.

[56]This is roughly equivalent to assuming utility follows the form $U_i(y,s) = \frac{y^{1+1/\eta}}{1+1/\eta} - p_s \cdot y - \gamma y - \kappa_i \cdot \mathbf{1}(s=c)$. The fact that I am unable to reject the null that $\delta_S + \delta_P = \delta_B$ from the *moral suasion treatments* provides some support for this assumption.

[57]I drop the post-enrollment non-event periods due to the fact that control units in the moral suasion treatments vary by event and thus spillovers are not defined as cleanly as in the price treatments. These periods provide additional variation for precisely estimating the effect of the control variables, but is not necessary for identification.

[58]I estimate the model using 2SLS and instrument $\log(MP_{it}) \times C_{it} \times H_t$ and $\log(HP) \times C_{it} \times H_t$ with $D_{it} \times H_{it}$ interacted with the corresponding price variable.

G_{it} indicates the event message included moral suasion. The value for HP is not identified so I assume it is the expected value of the marginal price in my base specification and explore the sensitivity to that assumption. The coefficient δ_{HP} technically recovers the intercept of price response scaled by the log of the heuristic price. I also include the same controls X_{it} as the preferred reduced form specification.

The parameters in (1.13) map to the choice primitives specified in (1.12) as follows

$$\delta_{MP} = \eta\sigma \tag{1.14}$$

$$\delta_{HP} = \eta(1-\sigma) \tag{1.15}$$

$$\delta_{G} = \gamma \tag{1.16}$$

We can recover the elasticity primitive by adding the two price coefficients because $\eta = \delta_{MP} + \delta_{HP}$. We can determine whether demand is perfectly inelastic (with no shift) by testing the null that $\eta = 0$. If rejected and $\eta \neq 0$, we can recover $\sigma = \delta_{MP}/(\delta_{MP}+\delta_{HP})$ which represents the fraction of the price response attributable to the marginal price change versus the heuristic price change.

The model requires $\mathrm{E}[\varepsilon_{it}|HP, MP_{it}] = 0$ for the parameters to be identified. The random assignment of the price level ensures MP is uncorrelated, but if there is unobserved heterogeneity in HP that is correlated with consumption, the elasticity and attention parameter estimates will be biased. For example, if households that have engaged in other conservation behaviors that make them generally more price responsive are also more to use a heuristic price that is above the expected value of marginal price, then η will overestimate the elasticity of demand. Estimates of the elasticity should be interpreted with caution if this is a concern.

1.5.2 Demand Estimates

Panel A of Table 1.5.7 reports estimates of the demand parameters η and σ obtained from estimating (1.13) using the same instrumental variables approach as the marginal price exercise.[59] The exercise pools households such that automation is not explicitly modeled. While raising identification concerns, removing automated households from the sample yields similar parameter estimates. I discuss incorporating automation into the model below.

The estimates are obtained by taking the sum and ratio of regression coefficients in MP and HP described in the previous section. Each column reports estimates obtained from separate regressions with a different assumed value of HP, noted in

[59]Namely, the model is estimated with 2SLS with $C_{it} \times H_t$ instrumented with the $D_{it} \times H_t$.

Table 1.5.7: Elasticity and Attention Estimates

	$HP = \mathrm{E}[MP_{it}]$ ($1.12/kWh) (1)	$HP = MP_{low}$ ($0.21/kWh) (2)	$HP = MP_{high}$ ($3.16/kWh) (3)
Panel A: Demand Estimates			
η (elasticity)	-0.073^{***}	-0.470^{***}	-0.050^{***}
	(0.005)	(0.036)	(0.004)
σ (fraction attentive)	0.110^{***}	0.017^{***}	0.160^{***}
	(0.033)	(0.006)	(0.045)
Panel B: Costs per Event			
Average Attention Cost	$0.03	$0.43	$0.06
Average Electricity Cost	$0.16	$0.16	$0.16
Average Reward	$0.15	$0.15	$0.15

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

Panel A estimates of the demand parameters obtained from the coefficients estimated using the model in (1.12). The estimate of γ the moral suasion component of utility is the same for all three specifications and equal to 0.014 (0.020). Standard errors are reported in parenthesis and are calculated using a covariance matrix that is clustered by household and hour-of-sample along with the delta method. Panel B reports estimates on the expected financial outcomes of a pricing event during the *price treatments*.

the column heading. Standard errors are reported in parentheses.[60] For all of the regressions the estimate of γ is 0.014 and is statistically insignificant with a p-value of 0.46.

Column (1) reports the results assuming HP is equal to the expected value of the marginal price plus the average price of electricity, $0.96/kWh + $0.16/kWh = $1.12/kWh. The number in the first row shows that the elasticity of demand implied by the model is -0.075 and is significantly different from zero. Further, the estimated fraction of attentive households during any given event is about 10 percent, meaning about 9 out of 10 households are inattentive for any given price change. The fraction is statistically different from zero which reflects the slight downward slope of demand recovered from the marginal price exercise. If the model structure is assumed to be true, households appear to be responding to price, but only a small fraction of the slope explains the overall shift.

[60]Standard errors are estimated in the regression using the same two-way clustering by household and hour-of-sample as the main results. The standard errors for the parameters recovered from combining regression estimates using the delta method.

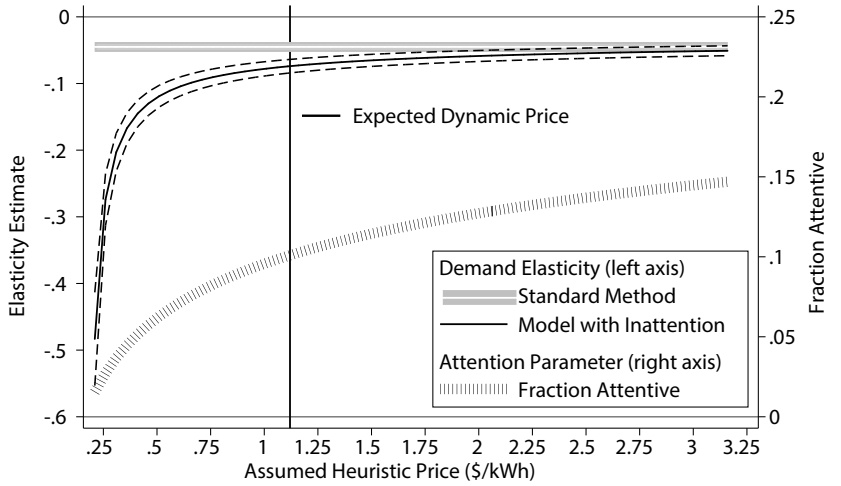
For reference, running the double log specification and omitting the heuristic price variable yields an elasticity estimate of -0.045, which is 38 percent smaller and statistically significantly different. To show this more concretely, I estimate the model for the full range of the possible assumptions about HP over the price changes in the study. Figure 1.5.8 plots these results with the horizontal axis showing the range of possible estimates. The solid white line shows the model estimated using the standard double log formulation with the gray bars indicating the 95 percent confidence intervals plotted against the left axis. The solid black line shows the estimates of η according to the model demand elasticity with dashed lines indicating 95 percent confidence intervals. The hashed line shows the estimates of the σ plotted against the right vertical axis and is generally less sensitive to the change in assumption about HP. The elasticity estimates converge for prices above \$2/kWh, but for the majority of the price distribution, the estimates are substantively different. This suggests the model with inattention recovers meaningfully different estimates and that the standard approach may mischaracterize underlying preferences.

Columns (2) and (3) of Table 1.5.7 report the sensitivity of the results to assumptions about the value of HP at the extremes. Column (2) shows estimates obtained from a regression that assumes the heuristic is equal to the minimum of the pricing distribution, \$0.21/kWh, and column (3) the maximum, \$3.16/kWh. These map to the extremes for estimates of η and σ under the model structure. The elasticity results follow general intuition. If the true heuristic is assumed to be at the low extreme of the distribution, then households preferences are very elastic with an estimate of about -0.5, significantly larger than the empirical literature that uses billing data. The results also show only 1 percent of households appear to be attentive since a smaller fraction of the elasticity is represented in the empirical slope. Alternatively, the high price heuristic shows at most 15 percent of households are attentive and demand appears slightly less elastic than the results in column (1). I interpret these results as evidence of a price response that is largely consistent with values of other electricity demand estimates. However, the low estimate for fraction attentive suggests the constant elasticity form alone does a poor job explaining the price response.

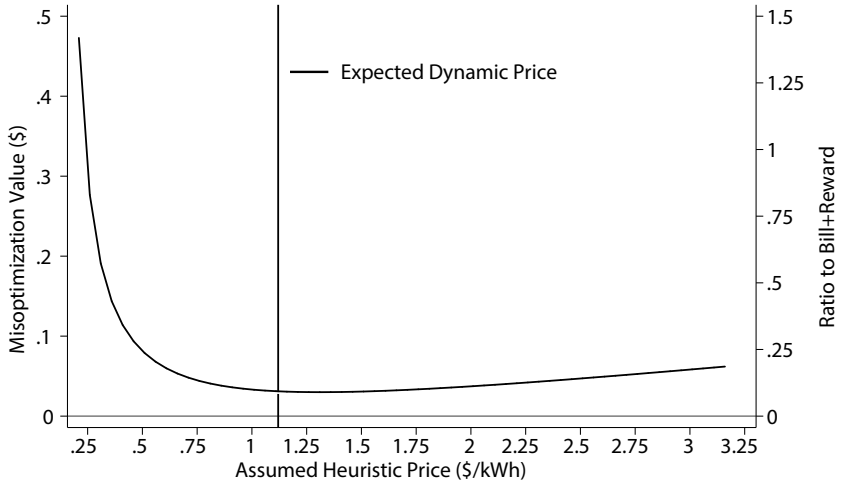
1.5.3 Attention and Misoptimization

The structural estimates presented in Panel A of Table 1.5.7 suggest inattention is a factor in decision-making. Panel B of Table 1.5.7 reports estimates of $\mathrm{E}[a(MP, HP)]$ using the definition from (1.5) and the distribution of MP from the *price treatments*. This recovers the expected attention cost of a single event. While the price distribution is not reflective of the true cost of electricity in this setting, it is informative

Figure 1.5.8: Demand Estimation



(a) Demand Elasticity and Attention Estimates



(b) Attention Cost Per Event

This figure shows how the estimates of the elasticity (η) and fraction attentive (σ) vary with assumptions about the heuristic price plotted in \$/kWh on the horizontal axis. The solid black line shows the estimates of the own-price demand elasticity plotted against the left vertical axis. The dashed black lines show the 95 percent confidence intervals for the elasticity parameter. The hashed line shows the estimates of the salience parameter which is the fraction of attentive households against the right vertical axis. The solid white line shows the model estimated using the standard double log formulation with the gray bars indicating the 95 percent confidence intervals.

as to the impact of attention on the financial outcomes of the study. The results in the first column show that the average misoptimization cost is \$0.03. The misoptimization is minimized at the expected value so this provides a lower bound to the attention cost using this framework.

The nominal value is small, but sensitive to the assumption about the heuristic. Consider the extreme values to gain intuition. A household whose heuristic is the low extreme consistently under-responds. The low heuristic implies true utility is very elastic, so overconsumption has little value. Thus, large price changes yield large mistakes. The results are significantly less sensitive for the high heuristic. In this case, households consistently consume too little, but are much less elastic so the differences between true utility and what is observed are small. Figure 1.5.8b plots the attention cost estimates for the range of HP assumptions showing the costs are much higher for households whose heuristic is below the expected value.

Table 1.5.7 also shows estimates of the financial outcomes of an event to provide context. I report the average cost of electricity consumed during the event along with average expected reward. These are calculated individually using the demand estimates, but rounding makes them appear the same. On average households are paying \$0.16/kWh because the level of consumption among treated households during events is about 1kWh and I calculate the average reward to be \$0.15 per event.[61] Column (1) shows the attention costs represent between 10 to 19 percent of the financial stakes depending on the inclusion of the reward.[62] Columns (2) and (3) show the estimate can represent up to 142 to 273 percent of the financial outcome. Aggregating these mistakes over larger consumption suggests potentially large attention costs, but ultimately depends on the number of events called and the distribution of price changes.

1.5.4 Automation and Attention

This section discusses how to incorporate automation explicitly into the framework developed above. The fact that automation is not price responsive complicates the welfare interpretation since it opens the potential for overcorrection. Further, automation could change the relative cost of information as well as the cost of effort, affecting both the σ and η parameters. For example, otherwise attentive households could become inattentive and engage in “set it and forget it" behaviors. Given these complications, future work should seek to understand the degree to which automation

[61] I assume the forecast for an event is given by the control group. This provides a good approximation for illustrative purposes, but may not perfectly reflect the rewards.

[62] Including the reward may not be appropriate for all dynamic pricing programs, for example those that use price.

reveals a preference for inattention. This would require a more rigorous evaluation of WTP or the introduction of automation that actively elicits reservation prices as an input.

1.6 Discussion

This section interprets the results of the empirical analysis and the model application. I provide justification for the simplifying assumptions of the model, but also comment generally on the mechanisms implied by the reduced form results. I interpret the response to a price, but not the price level, as suggesting that households are making decisions using a simplifying heuristic consistent with a rationally inattentive agent. I show the results are unlikely to be due to moral suasion, incorrect or incomplete valuation of the reward currency, observation of the price incentive, and physical discontinuities in the consumption choice. I also interpret the automation effect as suggesting there are substantial costs to taking action that are unrelated to attention to the price level which affect the overall elasticity of demand.

1.6.1 Improbable Mechansisms

Moral suasion – While there is evidence that non-price factors such as social comparison (Allcott 2011b) and moral suasion (Ito, Ida, and Tanaka 2017) affect residential electricity consumption decisions, the finding that households do not respond to the moral suasion alone provides evidence that the effect is driven by the presence of a financial incentive rather than non-price motivations. Given that households choose to enroll in the program, the concern that non-price motivations could be driving the result is reasonable, but evidence from the survey shows 61 percent of respondents rank "getting financial rewards for saving" as the top reason for joining the program, compared with 21 percent who chose "environmental or sustainability concerns."[63]

Observation of price level – If households are simply reacting to the receipt of a message, there may be a Hawthorne effect where the shock to salience affects energy consumption regardless of the message content. Again, the difference in response to

[63]The other choices were "making my home Smarter and more efficient", "entertainment/gamifying of energy use", "grid reliability", and "donating rewards to a specific cause". While the survey is a non-random subset of the sample, it provides suggestive evidence that financial rewards are the primary motivator for participating households.

moral suasion versus price messaging goes against this interpretation and suggests households are paying attention to message content.[64]

Misunderstanding of monetary value of points – The fact that households respond significantly to pricing events and not to moral suasion again shows that households are making economic tradeoffs between consumption and points. However, if there is confusion over the dollar value of points, then large changes in the points per kWh rate may not map completely to the economic significance when households are making decisions. The monetary value of a point is communicated consistently throughout the program through the rebate and point balance. The survey also tested whether households understand the dollar value of points, by asking them the question: "How much money is 100 points worth?" Given a multiple choice of four numeric answers and a fifth "I don't know" option, 70 percent of survey respondents gave correct answers.[65] This suggests misunderstanding about the economic value is not driving my price insensitivity finding.

Electricity consumption is a discrete choice – If electricity consumption depends on a set of discrete choices, the shift in demand I find may simply be the shape of a step-wise demand curve. This explanation is unlikely because demand represents a full hour of electricity consumption and utilization provides a continuous margin to respond. For example, a household that decided to turn off the lights and take a walk when an event was called might extend the walk for higher prices. Further, the distribution of consumption within household for pre-enrollment periods does not appear to be discrete.[66]

[64]The survey also asks two questions designed to test whether households observe message content by asking them to recall the specifics of their last pricing event. 40 percent of households correctly recall their last incentive level (as compared with a 17 percent chance of guessing the correct answer randomly). Households are asked to recall their last incentive level in points per kWh from a set of six numeric choices: 0, 5, 25, 50, 100, and 300 points per kWh. Since recall is likely to be a perfect approximation of observation, the survey also asks a simpler question of when their last event period occurred to gauge recall accuracy unrelated to the price level. The question gave six answers to choose from (today, yesterday, 2, 3, and 4 days ago, and 5 or more days ago). Only 32 percent correctly recalled how many days had passed since their last event (compared with a random chance of about 17 percent). The question was incentivized for correctness by offering $0.25 per correct answer. The incentive size was chosen to be small enough as not to encourage individuals to look up their last event, but large enough for them to put some thought into the response. This provides further evidence that people are observing the message content, but suggests attention to the message content is likely to be imperfect.

[65]20 percent answered "I don't know" and 10 percent answered incorrectly.

[66]In the Appendix I show example distributions that show the continuousness of consumption during a single hour of the day during the pre-enrollment period conditional on temperature.

1.6.2 Scope Neglect and Price Insensitivity

The discussion above makes the case that households are responding to a price change, but are responding much less to its size. This leads to the question: what information are households using to make their decisions? I argue households are using simplified decision rules that I will refer to as heuristics. There is precedent for such behavior in electricity consumption when facing nonlinear tariffs (Ito 2014), but my findings suggest that there are also information costs for linear incentives.

The finding that households respond to price, but not the size can be explained by the phenomena of *scope neglect* (Kahneman 2003). The concept comes from evidence in psychology that individuals make decisions based on simple accessible *prototype attributes* and often neglect the scope of *extensional attributes*. [67] In this setting, we can think of the prototype attribute as a price change occurring since households do not respond to the moral suasion intervention.[68] This categorical information is readily available for a quick decision about whether or not to reduce consumption. However, contemplating the extensional attribute, the size of the price change, requires additional cognitive effort.[69] If the effort cost is sufficiently high, households may be making decisions based on the prototype attribute of whether a price change occurred.

1.6.3 Automation Reduces Effort Costs

The significant effect of automation in response to a pricing event provides evidence that the technology makes it easier for households to respond to price changes by reducing effort costs. Two empirical findings support this interpretation. First, Figure A.7.3c shows clearly that automation causes a significant reduction in consumption during pricing events and no evidence of displacement to other periods. During the event period, the effect is negative and large. During the periods prior to and following the event, the effect of automation is not consistently positive or negative and

[67] For example, households have been documented in stated willingness to pay (SWTP) elicitations as revealing preference to save an endangered species of bird, but having the same SWTP for 2,000 birds as they do for 200,000 (Kahneman, Ritov, and Schkade 1999).

[68] I also investigated whether households were more sensitive to several candidate heuristics: the previous price level, a Bayesian prior over past price levels, and whether the last price was higher or lower. The price insensitivity persists for all candidates, but some suggest more price sensitivity than others with there being some evidence of Bayesian updating taking place. I report the results of this exercise in the Appendix.

[69] Again, the effort need not map to observing the price, but rather making the intermediate decisions required to make a marginal response that includes the extensional information.

the estimates are statistically insignificant.[70] Second, rebate compliers consume less than always-takers, on average. If some component of effort costs are independent of quantity of electricity consumed, the benefit of automation would be greater for households who consume more. This implies lowering the technology's cost should disproportionately induce smaller consumers to take-up the automation. Table 1.4.5 reports evidence supporting this case since households in the lowest pre-enrollment consumption quartile take-up the automation significantly more than larger consumers. This is also suggestive evidence of rationality on the adoption margin.

To provide a ballpark on the costs of effort automation addresses, I calculate a back-of-the envelope measure of the additional electricity expenditure savings. Households who adopted due to the rebate reduced an additional 0.57 kWh on average during pricing events.[71] If the average reward level is \$0.96/kWh, this suggests a additional reward of about \$0.54 per event.[72] This suggests additional savings on their electricity bill of about \$0.09, leading to a total of \$0.63 per event. Over the course of 3 months, households received 26 events on average, implying a total of about \$16 additionally saved due to the automation.

The study period paid incentives that were not necessarily reflective of the rewards during typical program operation so these numbers should be interpreted with caution when extending them to future periods. If we assume real-time wholesale prices reasonably reflect the future reward rates, a reasonable upper bound would be \$0.30/kWh for the year which gives about \$0.26 per event from the program and bill savings. For a program with 100 events per year, the annual savings are \$26, giving two smart plugs a payback period of about 3 to 4 years with a discount rate of 6 percent. The payback period for a smart thermostat is between 10 to 13 years, depending on upfront cost. Including the additional non-event savings makes these periods even shorter, suggesting the subsidized price is particularly beneficial to the average household.

While the automation suggests household preferences are more price elastic, the fact that price responsiveness does not substantively change suggests the automation does not address inattention. The automation did not differentiate response based on the price change and thus can be interpreted as default response that households could override via their smart phone. This raises questions over whether the automated price response can be interpreted as perfectly reflective of preferences, but the

[70] I also fail to reject the null that the sum of changes in all 8 hours following an event, the 8 hours before and 7 hours prior (excluding hour -2), or the sum of all adjacent hours is different from zero.

[71] This number comes the automation IV with kWh as the dependent variable reported in the Appendix.

[72] I assume the control group proxies for the forecast in this calculation.

fact that households chose to adopt suggests there is a revealed preference for lower effort costs. Given that the automation was marketed for this purpose, this seems a reasonable conclusion.

1.7 Conclusion

This paper documents how households respond to dynamic prices in residential electricity consumption. I show evidence that households reduce consumption when price increases, but are insensitive to extreme changes in the marginal price in a manner indicative of scope neglect. I develop a model of limited attention to explain the findings which suggests 9 out of 10 households are inattentive to the size of the price change during any given event and elasticities recovered from models that do not account for attention frictions may give substantively different estimates. I find the misoptimization costs of inattention are nominally small, but represent meaningful fractions of the financial stakes. Lastly, I report novel evidence that automation can make households generally more responsive to price changes by addressing substantial costs of effort, but does not necessarily resolve inattention unless.

Given the increasing amount of consumer dynamic pricing and the developing landscape of smart devices, this paper highlights several areas for continued research. If attention costs are prevalent, how can dynamic pricing be designed to minimize misoptimization? If automation is the primary strategy, how can reservation prices be elicited in a way that is consistent with preferences? Carefully designed field experiments can contribute to these lines of inquiry and inform future dynamic pricing strategies that take into account the attention costs consumers face in their decisions.

Chapter 2

The Effect of Temperature on Student Test Performance: Evidence from High School Exit Exams[1]

2.1 Introduction

The earth's climate will be altered dramatically in the coming century due to the effect of rising greenhouse gas levels. One risk climate change poses is the increased frequency and severity of heat waves (USCCSP 2008). Thus, studying the effect of thermal stress on human and economic outcomes provides valuable information for assessing the impact of this potential threat and the degree to which adaptation technologies can mitigate any negative effects. Within economics, there is an established body of work quantifying the sensitivity to exposure to extreme temperature on health, but less on how it affects human capital development.[2] This paper examines this important dimension by estimating the effect of extreme outdoor heat adversely affects educational outcomes as measured through test scores and whether climate control technologies mitigate the effect.

The psychology and human factor literature has long studied the impact of indoor environment on psychometrics and human productivity in controlled laboratory settings (Hancock and Vasmatzidis 2003), and there is a growing body of evidence documenting the effects of thermal stress in the field (Graff Zivin, Hsiang, and Neidell 2018; Park 2018). The lab evidence shows an inverted U-shape relationship where an

[1]The material in this chapter is from a working paper coauthored with Maximilian Auffhammer and Catherine Wolfram.

[2]See Deschênes (2012) for a review of the literature.

individual's perception, cognition and reaction time are inhibited above and below an optimal range of comfortable temperatures (Hancock, Ross, and Szalma 2007). Seppänen, Fisk, and Lei (2006) use a meta-analysis framework to characterize this relationship for office task performance and find performance increases in indoor temperature up to 21°C (71.6°F) and decreases for temperatures above 24°C (75°F). These results are interpreted physiologically as a breakdown in the body's ability to regulate its core temperature or that thermal stress poses a distraction in information processing (Staal 2004).[3]

This paper tests these laboratory findings using outcomes from the field with real educational stakes for millions of students. Specifically, we look at how performance on the California High School Exit Exam (CAHSEE) differed for students who experienced extreme outdoor air temperature on the day of the test as compared with those who experienced temperatures around 19°C. For the graduating classes of 2006 through 2015, passing the CAHSEE was a requirement for California students to receive their high school diploma, an achievement that is significantly correlated with higher lifetime earnings (Cullen et al. 2013).[4] Students first took the exam in tenth grade and it assessed their ability in English-language arts (ELA) and mathematics. While it is unclear whether students who failed their first attempt were less likely to graduate (Reardon et al. 2010), the exam existed as a barrier to at-risk students and thus provides a performance measure with real implications for future economic outcomes.

Using data on California public high schools, we observe school-level mean scores and pass rates in addition to the demographic composition of the tested cohort. The sample spans an an eight year period with sixteen test dates per subject. We match the school location to daily temperature data to approximate the maximum outdoor temperature at the testing site on the day of the test. We then exploit plausibly exogenous year-to-year daily variation in temperature within a school to obtain causal estimates of the contemporaneous effect of outdoor temperature on a school's aggregate performance.

Our results indicate small, but statistically significant negative effects for schools experiencing maximum daily outdoor temperatures above 25°C (77°F) as compared with those around 19°C (66°F). The effect is apparent in both the mean scores and

[3]A recurring counter mechanism mentioned in the literature posits that extremes may increase arousal and attention, but this is thought to be limited to primary tasks such as psychomotor skills rather than more complicated cognitive tasks (Hocking et al. 2001).

[4]Administration of the CAHSEE was discontinued with the class of 2016 with the passage and governor approval of Senate Bill 172 due to changes to academic standards. While the CAHSEE was discontinued, standardized testing remains common in the education system. Thus, our findings may be relevant to the broader administration of standardized testing.

pass rates of the mathematics section, but only significant in the ELA section for pass rates. We also find that the effects are larger for schools in areas with lower incomes, suggesting building quality and access to climate control technologies such as air conditioning (AC) may be a factor. The temperature response functions appear to follow the pattern documented in the lab, but the point estimates for the colder extremes are generally not significant. This provides quasi-experimental evidence that extreme outdoor heat has a causal impact on student test performance.

Park (2018) provides the most relevant work in the literature to date, documenting negative effects of high temperatures for students in New York City taking the New York State Regents Exam during years 1999-2011. The Regents Exam is similar to the CAHSEE in that it is a hight-stakes exam and passing it was a requirement for receiving a high school diploma. Park (2018) finds comparable estimates of high temperatures decreasing pass rates and scores using student-level performance data. Our paper extends these findings to another educational setting, providing evidence of external validity for a region with different weather and a significantly larger population of students.

Our results also support recent work by Graff Zivin, Hsiang, and Neidell (2018) that matches outdoor temperature data to the National Longitudinal Survey of Youth for days when a standardized psychometric exam was administered. They use within student variation and estimate similar negative effects for hot temperatures on a mathematics assessment, but find no effect for other subjects such as reading recognition and comprehension. Our results complement these findings by providing similar evidence using an exam that directly maps to education and potential labor market outcomes.

The difficulty in any of these studies is identifying the pathways by which outdoor temperature affects individuals taking an exam indoors. While the granularity of our study does not allow us to shed light on this issue, there are several possible explanations. First and foremost, indoor temperature may have been highly correlated with outdoor temperature if the settings where the tests were administered were not outfitted with climate control technologies.[5] Since the effects we estimate are much less significant for lower maximum temperatures, this may suggest operation of indoor heating was more prevalent than air conditioning. Another possible explanation is that students were exposed to outdoor extreme temperature before the exam and this carried over into the test setting. Increased exposure could have had physiolog-

[5]Conversations with several CAHSEE test site coordinators reveal that many schools chose to administer the exam in the school gym because typically the entire grade is tested and the students are allowed the entire school day to finish the exam. This made setting aside dedicated spaces difficult because it required the rescheduling of other classes.

ical impacts as noted above or increased the likelihood of social disruptions such as domestic violence and conflict (Carleton and Hsiang 2016). [6]

In addition to testing the laboratory research from psychology in the field, this paper contributes to the microeconomic literature that documents the effect of exposure to extreme temperature on economic outcomes such as excess mortality (Barreca et al. 2013) and labor supply decisions (Graff Zivin and Neidell 2014). The paper also pairs well with a larger body of work that looks at how human capital development is affected by other environmental factors such as air quality. One of the earliest works, Ransom and Pope III (1992) show student absenteeism increases with levels of ambient pollution. Other recent work characterizes the effect of air pollution on standardized test score performance and also finds negative effects (Ham, Zweig, and Avol 2014; Lavy, Ebenstein, and Roth 2012). To the extent that many of these papers instrument for pollution levels with daily temperature variation, some of the effects I measure could be a result of pollution intensity and the opposite could be true as well.

The paper proceeds as follows: Section 2 describes our data sources, Section 3 explains the estimation strategy, Section 4 reports the main results on the effect of temperature, Section 5 concludes.

2.2 Data and Summary Statistics

The models we estimate in this paper use school-level aggregates on CAHSEE performance and cohort demographics from the California Department of Education (CDE). From test score data we construct a panel for schools that spans the years 2004-2015.[7] We match school locations and test dates to daily maximum temperature from the PRISM climate mapping system to capture the outdoor temperature on the day of the exam. The demographics provided in the CAHSEE data do not report income, so we merge in American Community Survey (ACS) household income data by each school's ZIP code. Due to the Family Education Rights and Privacy Act (FERPA), the performance data for small schools is censored, but these omissions should not significantly affect our results, as we discuss later. The sample is comprised of first attempts by tenth grade students in the two exam sections: mathematics and ELA. Since the math and ELA sections are administered on dif-

[6]For higher maximum temperatures, exposure seems less likely since the state recommends schools start the exam fairly early in the morning when intra-day temperatures should be lower. However, such an effect may explain the negative point estimates we recover for lower maximum temperatures.

[7]Note 2004 corresponds to the tenth grade assessment for the class of 2006.

ferent days, we separate the sample into a dataset for each section. The unit of observation within each section's dataset is a school-date combination, but since we only consider first attempts schools only appear in the dataset once per year. This means each school has at most 12 observations, but only 43 percent of schools have performance data for every year in the sample period. Our primary analysis uses the unbalanced panel for each section where each contains 1,916 schools with 13,466 school-date observations from 24 different testing dates per section.

2.2.1 Data Sources

Testing Data — Passing the CAHSEE was a requirement for receiving a high school diploma in California from 2004-2015. State law mandated that students take the exam for the first time in the second half of tenth grade and students could only take it once that year. Failure to pass on the first attempt did not necessarily imply failure to graduate since there were multiple opportunities to retake the exam in later years. However, the education literature suggests that failure may have had negative consequences for persistence due to student discouragement and lowered expectations (Papay, Murnane, and Willett 2010).

The exam was administered on specific dates designated each year by the state. Thus, schools could not postpone exams based on day-to-day variation in weather. Each school's first assessment of tenth graders was defined by the state as the census administration. The state required school districts to offer each year's census administration on one of two sets of dates in February or March. All tenth graders were required to take the exam during the census administration.[8] We restrict our analysis to first attempts to avoid selection problems for absentees and complications that may arise from students re-taking the exam.

The content of the exam was based on tenth grade state curriculum standards so many students were able to pass it before completing their high school education. For instance, over 65 percent of tenth graders completed the CAHSEE requirement in 2007 (CAHSEE Fourth Biennial Report, 2008). Results reported for other years are of a similar magnitude. The mathematics and the ELA section both consist of multiple choice questions. The ELA section also contains a writing assessment. A passing score entailed answering 55 percent of the questions correct on the mathematics section and 60 percent for the ELA section. Students' raw scores for each section were converted into a normalized "scale score" that accounts for variation in

[8]Students retained in grade ten were not allowed to take the exam during the census administration. If students were absent during the census administration, districts offered make-up opportunities for students on subsequent March and May test dates.

the test across years. The minimum score for a student to pass was 350 for each section.

The data we obtained on the CAHSEE is publicly available from the CDE website.[9] We use school-level aggregates from academic years 2003-2004 to 2014-2015 for a sample of all California public high schools.[10] We refer to the group of students who took the exam on a given date at a given school as a testing cohort and we refer to the latter year in a school year as the year of the data. For every school, we observe the date the exam was offered, the number of students tested, the number who passed, their mean scale score, and the demographic breakdown of the tested cohort. We use two performance statistics in the analysis, the percentage of the cohort who passed and their mean scale score.

We cannot observe the aggregate performance statistics for cohorts or demographic groups of less than 11 students due to FERPA compliance. This has two consequences. First, it censors small schools, omitting 1,091 schools and leaving 2,291 to be included in the analysis. The censored schools do not appear to have any spatially correlated determinants of omission and the geographic coverage of the included schools looks complete. Further, the number of students censored by FERPA amounts to less than 0.5 percent of the total tested population for the census administrations.[11] Thus, school censorship should not significantly alter our estimates of the population effect.

Second, we cannot observe performance statistics for demographic groups within a cohort that are smaller than 11 students. This hinders our ability to condition effects on demographics. For example, a cohort of 20 students with an equal number of male and female students would have a performance statistic for the total cohort, but not for male and female students individually. we recover information from the data available by using the number of students within each demographic cut to construct the percentage of tested students belonging to each demographic. Specifically, we calculate the fraction of students in the cohort who are female, black, Latino or

[9]Downloaded on May 1, 2018 from http://cahsee.cde.ca.gov/datafiles.asp.

[10]We restrict the sample to schools in what we traditionally refer to as school districts. Specifically, schools that are members of Unified School Districts, High School Districts and Elementary School Districts, which contain charter high schools and specialty academies that administer the CAHSEE. The complete dataset includes non-school entities that administer the CAHSEE such as county offices of education, juvenile detention facilities, and state specialty schools, but we omit these due to potential differences in the tested populations that will be difficult to control for.

[11]Our data is comprised of 5,231,505 student test scores and only 24,298 were censored due to FERPA.

hispanic, Asian, economically disadvantaged[12], English-language learners[13], and participating in special education programs. This is better than restricting the sample to schools for which demographic groups all number greater than 11, which would dramatically diminishes the number of schools included in the analysis.

One last fact to note is that the exams were administered over a two-day period that always took place on a sequential Tuesday and Wednesday. The ELA section is taken on the first day and the math section on the second. In order to guard against potential differences in student populations between sections we only use performance data for schools that survive the FERPA censorship in both sections. This drops about 150 observations and 20 schools from the sample.

Temperature Data — We use the fine-scale gridded daily temperature data from the PRISM climate mapping system[14] to measure the maximum outdoor temperature on the day of the exam. The model from which the data are constructed gathers information from a wide array of monitoring stations and performs sophisticated quality control measures to map weather variables to a 2.5 $\times$ 2.5 arc-minute grid (about 2.25mi $\times$ 2.75mi rectangles for California).[15]

Matching Temperature to Schools — We obtained latitude and longitude coordinates for a majority of the schools in our dataset from a CDE directory of all public schools.[16] For schools with no reported coordinates, we used the directory address and the Google Geocoding API[17] to obtain coordinates for nearly all of the remaining schools.[18] Using these coordinates we performed a nearest-neighbor match with the centroids from the grid cells in the PRISM weather data. Grid points were allowed to be matched to multiple schools, but schools were only matched to a single grid point. With these matches we then merged in the weather on the date of the exam reported in the CAHSEE dataset.

[12]This status is constructed from participation in the National School Lunch Program and is a commonly used proxy for economic disadvantage in the education literature.

[13]These are students for whom English is not designated as their primary language.

[14]Downloaded on May 1, 2018 from the PRISM Climate Group website http://www.prism.oregonstate.edu/recent/.

[15]The data is widely used by government agencies (Daly and Bryant 2013) and in economic climate analyses (Schlenker and Roberts 2009). The specific climate measure we use in our analysis is daily maximum temperature.

[16]Downloaded on November 19, 2015 from http://www.cde.ca.gov/ds/si/ds/pubschls.asp.

[17]See https://developers.google.com/maps/documentation/geocoding/ for an explanation of the service

[18]We were unable to successfully geocode 20 schools for which we have CAHSEE data due to errors processing the provided addresses. These were omitted from the analysis.

ACS Income Data — The CAHSEE data reports the fraction of economically disadvantaged students, but does not report income. In order to augment the data along this dimension, we merge in the 2008-2012 American Community Survey 5-year estimates for median household income in the past 12 months for the ZIP code tabulation area for the school address.[19]

2.2.2 Summary Statistics

Testing Data — Table 2.2.1 reports summary statistics for the 48 testing dates used in the analysis – 24 dates per exam section. The total number of schools or students for each year can be obtained by adding the February and March dates together. This gives 1,397 schools for 2004 and 1,667 schools for 2015. More schools tend to administer the exam in March than February and the number of schools administering the exam in February declines throughout the sample. About a third of schools in the sample that change the month of their exam throughout the sample and about a third of those change it more than once.

Several other secular trends are apparent from Table 2.2.1. Notably, for a given administration (row) there are fewer students taking the math section than the ELA, possibly due to student fatigue or frustration. While this implies there are differences in the composition of the math and ELA cohorts, they are less than 0.6 percent of the population for every year. Further, the effect on the average cohort size is negligible.[20] There also appears to be a shift to smaller cohorts. Since schools were required to have students take the exam during the census administration, it is unlikely that schools are limiting their census administration cohorts to more able students. Rather, the change could be a result of schools becoming smaller over the course of the sample.

The CAHSEE performance statistics exhibit a large amount of variation both within a school over time and between schools. To show this more clearly we plot the kernel density distributions of the mean scale scores and pass rates for the first and last year of the sample. Figure 2.2.1a and Figure 2.2.1b plot kernel density distributions for the raw ELA and math performance statistics in the first and last years of the sample in black and gray, respectively. The left panels of both figures plot the distributions for mean scale scores and the right panels plot pass rates. The distribution of both performance statistics clearly shifts to the right from 2004 to 2015 and the years in between display intermediate phases of this pattern. This is

[19]Downloaded on February, 21, 2014 from http://factfinder2.census.gov. Estimates reported in 2012 inflation-adjusted dollars.

[20]The numbers are the same when rounded to the nearest integer. Hence, the third column reports one number for both exam dates.

Table 2.2.1: Summary Statistics by Test Date

		ELA Administration				Math Administration			
Test Dates	# of Schools	# Tested (000s)	Pass Rate	Temp. (°C) Mean	SD	# Tested (000s)	Pass Rate	Temp. (°C) Mean	SD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Feb. 3-4, 2004	483	139	0.75	13.7	3.5	138	0.73	12.6	2.8
Mar. 16-17, 2004	914	269	0.78	25.5	3.5	268	0.78	27.9	3.4
Feb. 8-9, 2005	497	130	0.76	14.1	2.4	129	0.72	15.6	3.0
Mar. 15-16, 2005	969	289	0.80	21.3	2.3	287	0.78	22.4	2.0
Feb. 7-8, 2006	624	163	0.78	21.3	4.2	163	0.75	23.2	3.3
Mar. 25-26, 2006	921	279	0.80	22	4.2	278	0.79	18	3.8
Feb. 6-7, 2007	617	160	0.78	22.8	5.0	159	0.79	20.7	3.8
Mar. 20-21, 2007	947	284	0.79	20.5	3.4	283	0.77	16.9	2.8
Feb. 5-6, 2008	581	141	0.81	12.5	2.9	140	0.80	13.7	2.4
Mar. 11-12, 2008	980	300	0.80	24.9	3.0	299	0.80	24.3	3.6
Feb. 3-4, 2009	596	140	0.81	21.7	3.5	139	0.81	22.5	3.5
Mar. 17-18, 2009	1,022	304	0.81	20.9	3.0	303	0.82	23.7	3.6
Feb. 2-3, 2010	527	120	0.84	14.9	2.8	120	0.82	14.4	2.2
Mar. 16-17, 2010	1,123	326	0.81	24.2	3.1	324	0.82	26.4	3.5
Feb. 1-2, 2011	473	107	0.84	14.9	2.5	107	0.84	16.4	2.6
Mar. 8-9, 2011	1202	335	0.84	16.5	2.8	333	0.84	19.6	2.8
Feb. 7-8, 2012	478	109	0.85	18.2	2.2	109	0.85	15	2.5
Mar. 13-14, 2012	1,189	329	0.84	18.3	2.9	328	0.85	17.3	3.4
Feb. 5-6, 2013	436	102	0.84	17.1	3.0	101	0.86	14.5	2.3
Mar. 12-13, 2013	1,233	331	0.85	23.1	2.3	330	0.85	25	3.1
Feb. 4-5, 2014	485	115	0.84	13.5	2.3	115	0.87	13.2	2.3
Mar. 18-19, 2014	1,177	318	0.84	23	3.3	317	0.86	22.7	2.5
Feb. 3-4, 2015	459	107	0.86	20.6	3.1	107	0.86	20.9	2.5
Mar. 17-18, 2015	1,208	324	0.86	27.8	4.9	323	0.86	25	3.1
All Dates	1,916	3,724	0.82	20.8	5.2	3,761	0.81	20.8	5.1

The table reports summary statistics for each test date. The number of students tested is reported in thousands. Aggregate pass rates are calculated as the total fraction of students who pass on that test date. The temperature statistics reported are the mean and standard deviation (SD) of the maximum outdoor temperature on the day of the exam.

also notable from the aggregate pass rates in Table 2.2.1 which improve considerably through the sample period from around 78 percent to over 85 percent.

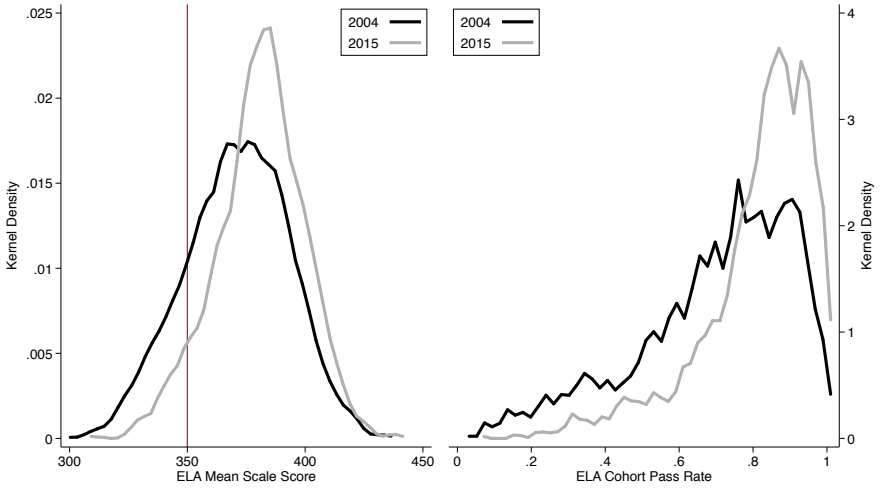
The distribution of scores and pass rates are also sharper in 2015 than in 2004, indicating that there is not simply a parallel shift in school performance. The narrowing of the distribution is driven largely by schools at the lower performing end of the distribution improving faster than schools toward the high end of the distribution. For instance, the pass rates for schools performing in the bottom quartile for math in 2004 where on average by 8 percent higher in 2015, while schools in the top quartile did not show any change.[21] A similar pattern is visible in the ELA distributions.

Temperature Data — Despite the mild climate in California during the months of February and March, the geographic coverage of the schools included in the analysis provides a good amount of temperature variation. Figure 2.2.2a shows the average number of schools per year that fall within the 8 temperature bins used in estimation. We top- and bottom-code maximum temperatures above 27.5°C (81.5°F) and below 12.5°C (54.5°F), respectively. We partition the remaining variation into six 2.5°C (4.5°F) bins. The lowest and highest maximum temperatures observed over the sample period were -4.9°C (23.2°F) and 35.3°C (95.5°F), respectively. The discrepancy between the bin heights for each section comes from the ELA and mathematics sections being administered on separate days. On average, schools span between three and four bins, but some span as many as seven over the sample period. The width of the bins were chosen to allow for a flexible estimation of the temperature sensitivity while also ensuring a sufficient amount of schools fell within each bin each year. The choice of the bin width was made to balance the need to estimate the temperature sensitivity flexibly while maintaining a sufficient number of schools in each bin to ensure estimates were precise. We discuss robustness to different temperature bin specifications in the results section, but the results are qualitatively similar for different bin specifications.

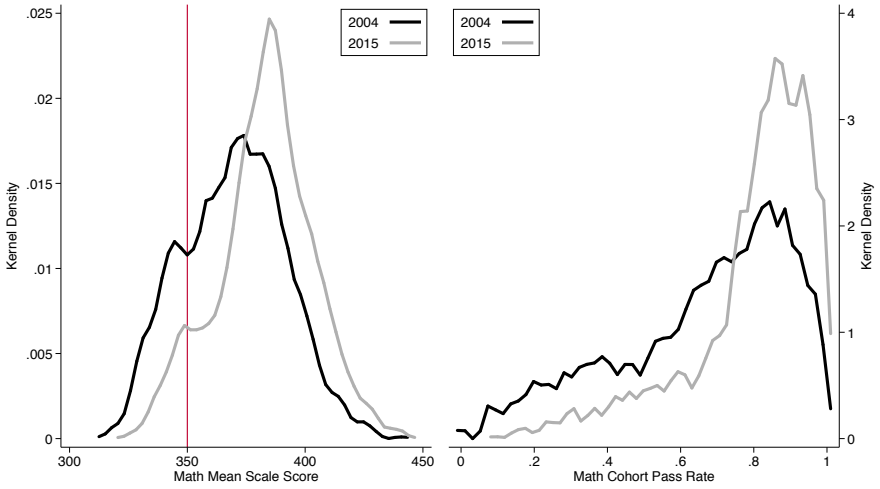
We break out the bins in Figure 2.2.2a to show the temperature variation by income quartile in Figure 2.2.2b. The figures show good coverage in the temperature variation. While the number of schools in each bin is roughly cut in four, but the distribution for each income quartile is of a similar shape. That is, there does seem to be significant differences in the coverage of the temperature variation by income quartile.

[21] We calculate these numbers by limiting the sample to schools with data in every year and estimating a model with school and year interacted with 2004 performance quartile fixed effects.

Figure 2.2.1: Performance Metric Distributions in 2004 and 2015



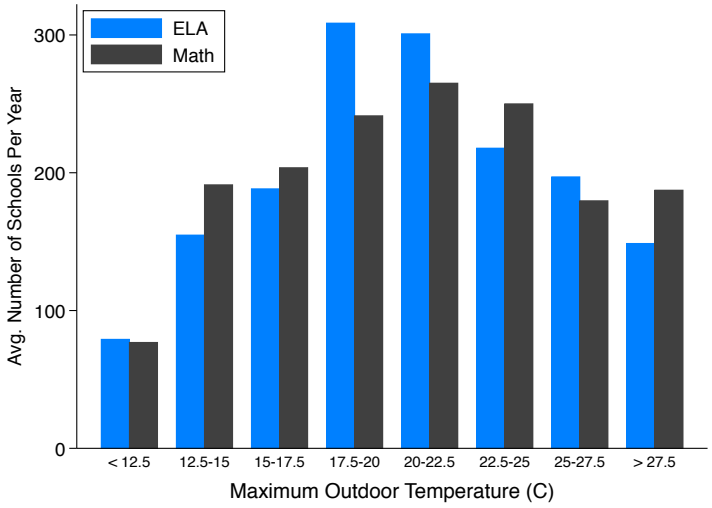
(a) ELA Administration



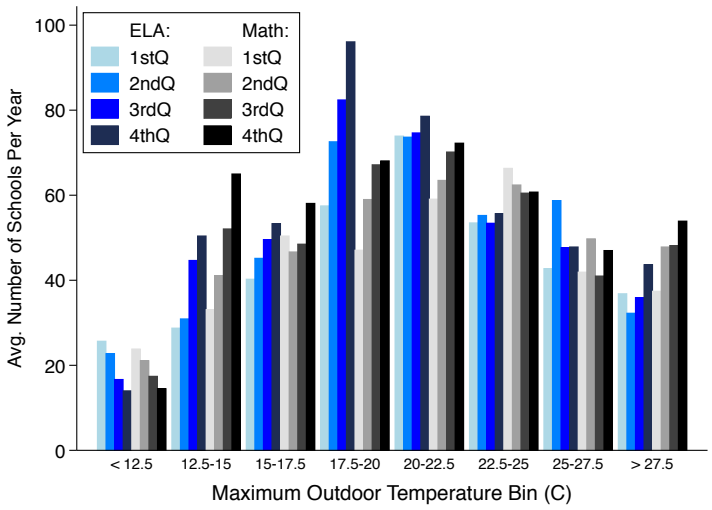
(b) Math Administration

The top (bottom) panel plots changes in the distribution of performance metrics over time for the ELA (math) administration. Both panels plot kernel density estimates for the mean scale scores (left panels) and cohort pass rates (right panels) on the CAHSEE for years 2004 (black) and 2015 (gray). The red lines in the left panels mark the minimum pass score of 350. Kernel bandwidths are 2.5 and 0.01 for scores and pass rates, respectively. All distributions are calculated using Epanechnikov kernels.

Figure 2.2.2: Temperature Variation



(a) Full Sample



(b) By Income Quartile

The vertical axis represents the average number of schools within each bin per year over the 2004-2015 sample for each subject. The difference in heights between the ELA (blue) and mathematics (black) numbers is due to the sections being administered on different days. The top panel reports results for all schools. The bottom panel breaks the temperature variation out by income quartile. The shades indicate the income quartile with the lightest indicating the lowest income quartiles and the darkest indicating the highest. Income is measured as median household income for the school's ZIP code.

2.3 Estimation Strategy

We exploit plausibly exogenous year-to-year within-school daily weather variation in order to measure the reduced form effect of maximum outdoor temperature on the day of the exam on aggregate student performance. To do this we use a semi-parametric specification with daily temperature bins, a common strategy within the literature quantifying the effect of temperature on economic outcomes (Barreca et al. 2013; Deschênes and Greenstone 2011). The bins allow for a more flexible estimation of non-linear effects than parametric methods while preserving as much of the temperature variation as possible.

Effect of Temperature — To recover the effect of temperature we estimate the following econometric model:

$$Y_{it} = \sum_{\tau} \theta_{\tau} 1(T_{it} \in B_{\tau}) + X_{it}\beta + \alpha_i + \delta_t + \varepsilon_{it} \tag{2.1}$$

where i indexes schools, t indexes the test date, and τ indexes the temperature bins. The dependent variable Y_{it} is the aggregate CAHSEE performance statistic for school i on date t and ε_{it} is the stochastic residual term. The temperature bins are indicator variables that the maximum outdoor temperature at school i on date t, denoted T_{it}, fell within the range of bin B_{τ}. We estimate (1) for both the cohort pass rate and mean scale score as the dependent variable and with the bin specification shown in Figure 2.2.2a. While the intention is to characterize the effects of extreme heat, we estimate the model over a complete set of temperature bins to test whether there are effects for colder temperatures as well.

The coefficients of interest are the θ_{τ} values which represent the average effect of maximum outdoor temperature in bin τ on deviations from the school-specific mean performance over the sample period, the mean for all schools that administered the test on that date, and the mean for the school's county in that year. We omit the temperature range that contains 19°C (66.2 F) so that we interpret a negative estimates as a decrease in the performance due to a deviation in maximum outdoor temperature from what is generally considered comfortable.

We include a full set of school fixed effects α_i to control for unobservable time-invariant factors. These take out the mean performance of each school over the entire sample period, controlling for differences in school quality so that we do not compare the raw performance levels of different schools. We also include a full set of date fixed effects δ_t to capture aggregate time-varying factors that affected test score performance such as school districts' familiarity with the exam, changes in test difficulty and administration protocol, and any advantage or unobserved difference

in performance between schools taking the test in February versus March. The date fixed effects control for the notable improvement in score performance that occured during the sample period. However, they do not control for heterogeneous variation between schools over time.

Covariates — In order to increase the precision of the estimates we include a vector of covariates X_{it}. These capture the percent of the tested cohort that belonged to various demographic groups. The demographic breakouts we use are percent female, black, Latino or hispanic, asian, Filipino, American Indian, EL, economically disadvantaged, and participating in special education.[22] Including higher order terms of the percentages does not change the results.

Heteroskedasticity and Observation Weighting — Heteroskedastic disturbances are a concern due to the fact that school sizes vary significantly. That is, we might expect an aggregate of 500 students to provide a more precise estimate on the effect than an aggregate of 50 students. Means and pass rates from schools with large cohorts on average should have lower variances than schools with small cohorts. We confirmed the presence of heteroskedastic disturbances for all specifications at 99 percent confidence using a White Test.[23] We then confirmed that average cohort size for a school is correlated with the variance of the residuals for each specification using a Breusch-Pagan Test.

We correct for this by estimating the model using weighted least squares. Specifically, we weight each school-date observation by the school's average cohort size across years. To be concrete, let w_i denote the average cohort size for a school so that

$$w_i \equiv \frac{1}{T_i} \sum_{t=2004}^{2015} n_{it} \tag{2.2}$$

where T_i denotes the number of periods for which we have data for school i and n_{it} denotes the number of students who took the CAHSEE at school i on date t. We weight each observation by $\sqrt{w_i}$. The results are qualitatively similar for both the weighted and OLS estimations, but weighting tightens the standard errors.

[22]This is slightly different from the specification used in the Reardon et al. (2010) study of the effect of the CAHSEE implementation on graduation rates, but their data is student-level so the samples aren't exactly comparable.

[23]I regress the residuals on the full set of school, date, and county by year fixed effects along with squares of all of the controls and average cohort size for each school.

2.4 Results

This section reports the main results of the paper. The first subsection reports the estimation results for effect of temperature on test scores, the second reports estimates for subsamples partitioned by income quartile, and the third discusses the results and their robustness to alternate specifications.

2.4.1 Effect of Temperature

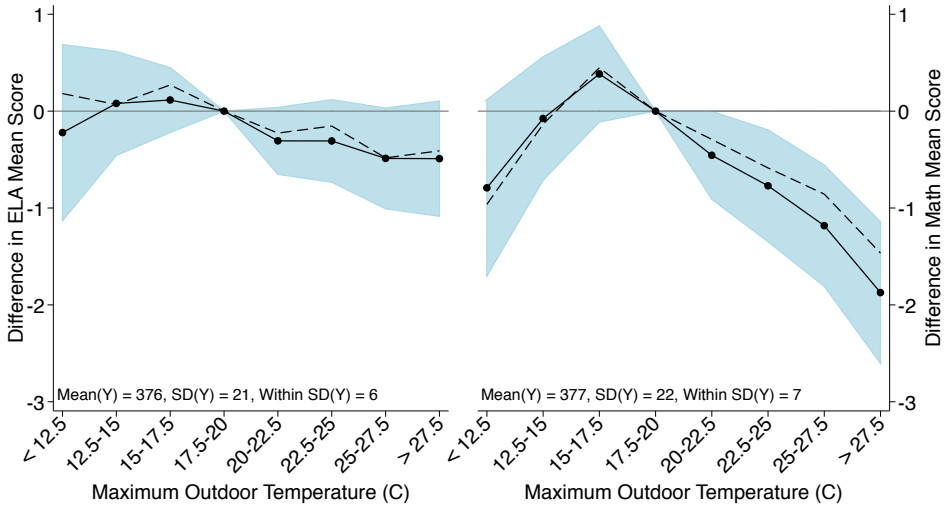
Figure 2.4.3 shows the main results of the paper by plotting temperature response functions for both exam metrics and administrations. Figure 2.4.3a reports results for scale scores and Figure 2.4.3b for cohort pass rates. The left panels reports estimates for ELA and the right panels math. The default bin spans 17.5°C-20°C so all regression coefficients can be interpreted as a difference in performance for schools experiencing maximum outdoor temperature outside of this range. All estimates come from separate regressions that weight observations by number of students tested and include school and date fixed effects and demographics as controls. The shaded regions show 95 percent confidence intervals constructed using a covariance matrix two-way clustered by school and county by year to account.[24] We provide a full set of regression tables showing the effect of adding controls to the estimation in the Appendix.

The results show extreme heat has a significant negative impact on exam performance, but varies by subject. Specifically, we find negative impacts of maximum temperatures above 27.5°C (81°F) on pass rates in both subjects and negative impacts for math test scores above 20°C (68°F). The results for ELA scores are statistically insignificant. Generally the response functions exhibit the U-shape documented in the lab, although the results for colder temperatures are not significantly different from zero. The same pattern is also apparent in OLS regression estimates that do not weight observations, as indicated by the dashed lines in Figure 2.4.3. Lastly, our effects exhibit linearity in the magnitude of the effect similar to Graff Zivin, Hsiang, and Neidell (2018).

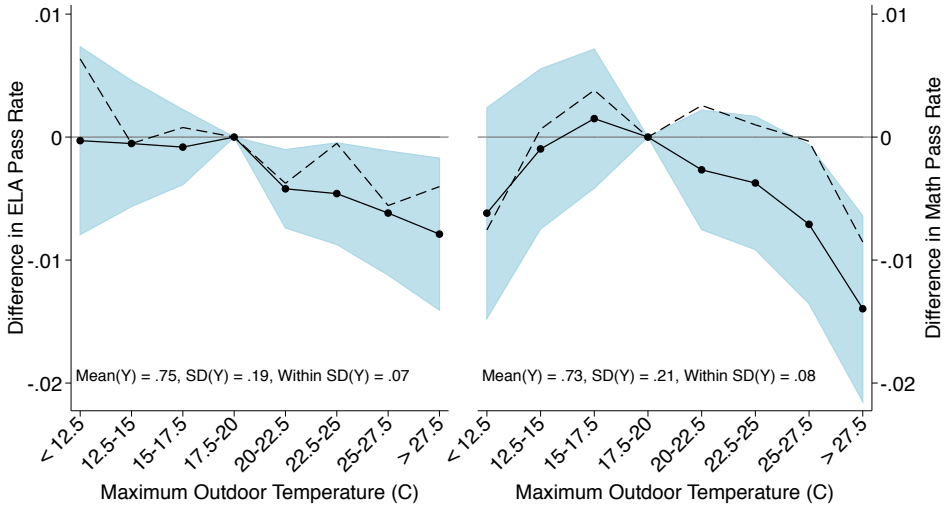
The magnitude of the estimates presented Figure 2.4.3 are modest with respect to the mean, but represent a non-trivial amount of the within-school variation. Specifically, our confidence intervals for the effect of the hottest temperatures indicate pass rates were 0.1-1.2 percentage points lower for ELA and 0.6-2.0 percentage points lower for math. The mean pass rates were between 70-80 percent, but the within school standard deviation for pass rates was 7 and 8 percentage points for ELA and

[24]This is meant to account for arbitrary correlation within schools over time and shocks to a county within a year that may be driven by an abnormally warm year or change in district policy.

Figure 2.4.3: Effect of Temperature



(a) Scale Scores



(b) Pass Rates

These figures plot causal estimates for the effect of maximum outdoor temperature on exam performance. Response functions are reported for mean scale scores (top panels) and pass rates (bottom panels), and administrations of ELA (left panels) and math (right panels) sections. Each panel's estimates come from a separate regression estimated with weighting, school and test date fixed effects, and demographic controls, omitting the bin spanning 17.5-20°C. Shaded areas indicate 95 percent confidence intervals constructed using a covariance matrix two-way clustered by school and county by year. The dashed line indicates the unweighted OLS point estimates.

math, respectively. This indicates the effect pf extreme heat accounted for between 1-17 percent and 8-25 percent of the variation in pass rates. Given the size of the testing population, a one percent change in pass rates maps to roughly 37,000 fewer students passing their first attempt over the 12 year period. The magnitude of the effects with respect to variation are similar for the mean scale scores of the math administration.

2.4.2 Exam Day Event Study

To investigate whether the effects are contemporaneous or driven by cumulative exposure, we estimate an event-study specification around the exam day. We also estimate the effect of days following the exam to provide a placebo test of our estimation strategy. To simplify the interpretation, we estimate a parametric specification of exam outcomes regressed on cooling degree days (CDDs) and heating degree days (HDDs).[25] Higher values of CDDs are associated with hotter temperatures and higher values of HDDs cooler temperatures. We average the CDDs and HDDs on the 3 days prior to and following the exam and include these on the right-hand side along with values for the exam day.[26]

Figure 2.4.4 plots the results for the effect of an additional CDD on exam outcomes for days prior to and following the exam. The top panels reports results for scores and bottom panels pass rates. The left and right panels report results for ELA and math, respectively. The middle point estimate shows the effect of an additional CDD on the day of the exam on performance. The estimates are generally consistent with the temperature response functions. Math scores and pass rates are negatively impacted by an additional CDD on the day of the exam and the results are statistically significant. The only meaningful difference is that hotter temperatures do not significantly affect the pass rate for ELA.

The most notable result of Figure 2.4.4 are the insignificant effects of temperature on the days prior to and following exams. This provides evidence that cumulative

[25] We calculate CDDs as the absolute deviation in the maximum temperature above 17.5°C and set the value equal to zero when maximum temperature is below the threshold. HDDs are defined similarly as the absolute deviation below 17.5°C.

[26] Specifically, we estimate the model

$$Y_{it} = \sum_{\tau=-1}^{1} \theta_{\tau}^{CDD} CDD_{\tau} + \theta_{\tau}^{HDD} HDD_{\tau} + X_{it}\beta + \alpha_i + \delta_t + \varepsilon_{it} \tag{2.3}$$

where CDD_{τ} and HDD_{τ} correspond to the CDDs and HDDs in period τ. $\tau = -1$ represents the 3 days prior to the exam, $\tau = 0$ represents the day of the exam, and $\tau = 1$ represents the 3 days following to the exam. The rest of the specification is defined in the same manner as (2.1).

effects of temperature in the days leading up to the exam are negligible. Further, the fact that temperatures following the exam have an insignificant effect is encouraging for the research design. We do not report HDDs since all coefficients were insignificant from zero, consistent with the results of the semi-parametric response function analysis.

2.4.3 Effects by Income Quartile

In order to understand the mechanisms driving temperature responses, we examine heterogeneity in the temperature effect by the income of the area surrounding the school. Income is likely to to have been correlated with many factors that affected temperature exposure. For example, schools in higher income areas may have been better equipped with climate control technologies. Alternatively, students in low income neighborhoods may have been more poorly prepared for the test and thus required greater cognitive effort to overcome the burden of extreme temperatures.

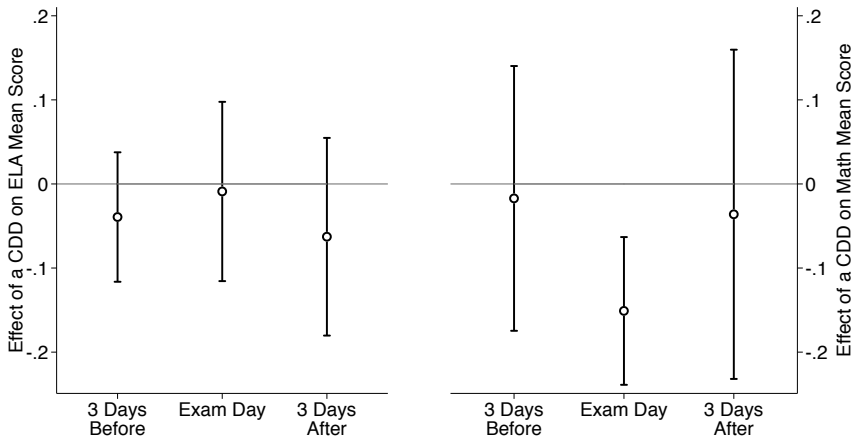
Using the ACS income data, we partition the sample into income quartiles by yearly median household income for each school's ZIP code. This roughly proxies for the income of the students and correlates strongly with the percentage of economically disadvantaged students at the school.[27] On average, schools in the bottom quartile had median household incomes of $35,122, while the top quartile was nearly 2.5 times higher at $88,945.

Tables 2.4.2 and 2.4.3 report the results for regressing mean scale score and cohort pass rates using the same econometric specification as the temperature response functions reported in the previous section. The results are reported from the lowest quartile on the left to the highest quartile on the right. The size of the subsamples and ranges of the income quartiles are reported at the bottom of the table. The number of observations is roughly the same, but the income distribution is right-skewed so the higher quartiles span larger ranges.

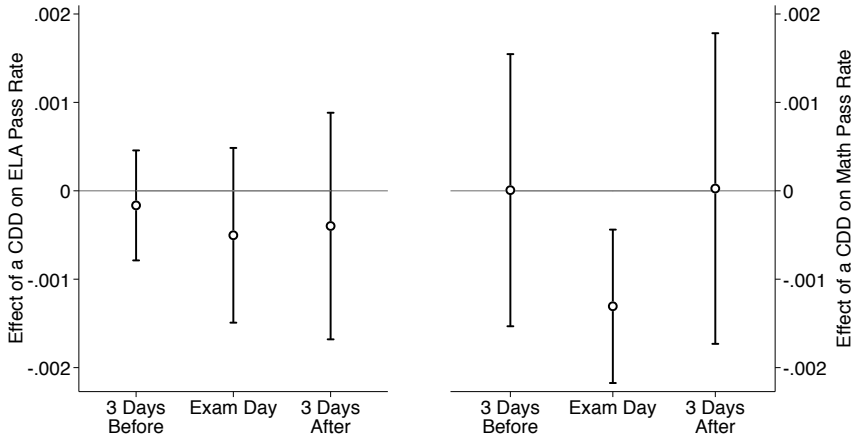
While the smaller sample sizes decrease the precision of our estimates, the results show larger estimates for the effect of high temperatures in the lowest income quartile. For scores and pass rates, the impact of temperatures above 27.5°C on students in the lowest income areas are about twice the size of those in the highest income areas. Testing the null hypotheses that the point estimates for the hottest temperatures are the same for the top and bottom income quartiles have p-values of 0.002 and 0.001 for math pass rates and scores, respectively. For ELA, the analogous p-values are borderline significant at 0.109 and 0.0946 for pass rates and scores, respectively.

[27]Regressing the average percentage of economically disadvantaged on our household income indicator gives a statistically significant association with a t-statistics of -31.

Figure 2.4.4: Effect of Temperature on Days Prior to and After Exam



(a) Scale Scores



(b) Pass Rates

These figures plot causal estimates for the effect of maximum temperature on days before and after the exam on performance. The point estimates represent the effect of an additional cooling degree day (CDD) on a given performance metric and are reported for mean scale scores (top panels) and pass rates (bottom panels), and administrations of ELA (left panels) and math (right panels) sections. Each panel's estimates come from a separate regression estimated with weighting, school and test date fixed effects, and demographic controls. Vertical bars indicate 95 percent confidence intervals constructed using a covariance matrix two-way clustered by school and county by year.

The point estimates are also monotonic for the bottom three income quartiles, but adjacent income quartiles are generally not significantly different. Interestingly, Table 2.4.3 shows statistically significant effects for the top income quartile, suggesting the negative impacts are not limited to schools in low income areas.

Table 2.4.2: Temperature Effect on Scores by Neighborhood Income Quartile

	1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Temperature Bin								
$< 12.5°$	0.43	0.77	-0.84^{*}	-0.67	-0.13	-0.26	-1.10^{*}	-0.64
	(0.73)	(0.69)	(0.50)	(0.46)	(0.52)	(0.52)	(0.57)	(0.45)
12.5°-15°	-0.18	0.04	-0.28	-0.35	-0.13	-0.11	0.30	-0.10
	(0.52)	(0.50)	(0.39)	(0.42)	(0.39)	(0.32)	(0.28)	(0.27)
15°-17.5°	0.81^{**}	0.79^{**}	0.65^{**}	0.26	-0.14	-0.10	0.19	-0.04
	(0.39)	(0.36)	(0.33)	(0.23)	(0.29)	(0.23)	(0.22)	(0.22)
17.5°-20°	-	-	-	-	-	-	-	-
20°-20.5°	0.16	-0.34	-0.65^{**}	-0.23	-0.52^{**}	-0.26	-0.33	-0.04
	(0.38)	(0.28)	(0.28)	(0.21)	(0.26)	(0.24)	(0.22)	(0.18)
22°-25°	-0.43	-0.42	-1.06^{***}	-0.48^{*}	-0.62^{*}	0.04	-0.76^{***}	0.39^{*}
	(0.46)	(0.35)	(0.38)	(0.27)	(0.33)	(0.23)	(0.27)	(0.21)
25°-27.5°	-1.30^{**}	-0.50	-1.02^{**}	0.00	-0.82^{**}	-0.36	-0.99^{***}	-0.12
	(0.51)	(0.36)	(0.40)	(0.26)	(0.41)	(0.30)	(0.30)	(0.27)
$> 27.5°$	-2.66^{***}	-0.79	-1.52^{***}	-0.43	-1.09^{**}	-0.14	-1.36^{***}	-0.26
	(0.70)	(0.49)	(0.46)	(0.38)	(0.48)	(0.39)	(0.36)	(0.34)
Default Mean	368.0	367.4	371.7	370.7	377.9	375.7	388.6	386.5
Obs.	3,714	3,714	3,982	3,982	4,102	4,102	4,480	4,480
Inc. Range	\$13,504 – \$43,000		\$43,004 – \$54,633		\$54,712 – \$72,218		\$72,264 – \$161,567	

Standard errors in parentheses

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

The table reports estimates from regressing mean scale score on temperature bins, demographic controls and school and test date fixed effects after partitioning the data by income quartile. Income is measured by the median household income for the ZIP code of the school's address. Math and ELA sections are estimated separately. Standard errors reported in parenthesis are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years. Below the estimates I report the mean in the omitted default bin and the range of incomes by quartile.

Table 2.4.3: Temperature Effect on Pass Rates by Neighborhood Income Quartile

	1st Quartile		2nd Quartile		3rd Quartile		4th Quartile	
	Math	ELA	Math	ELA	Math	ELA	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Temperature Bin								
$< 12.5°$	0.005	0.002	-0.004	-0.003	-0.004	-0.004	-0.009**	-0.003
	(0.007)	(0.007)	(0.005)	(0.004)	(0.005)	(0.004)	(0.004)	(0.003)
12.5°-15°	0.001	-0.004	0.001	-0.004	-0.003	-0.003	0.000	0.002
	(0.006)	(0.005)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)
15°-17.5°	0.005	0.003	0.005	0.001	0.000	-0.003	-0.000	0.000
	(0.004)	(0.004)	(0.003)	(0.002)	(0.003)	(0.002)	(0.002)	(0.002)
17.5°-20°	-	-	-	-	-	-	-	-
20°-20.5°	0.004	-0.006**	-0.003	-0.002	-0.000	-0.002	-0.003*	-0.001
	(0.004)	(0.003)	(0.003)	(0.002)	(0.003)	(0.002)	(0.001)	(0.001)
22°-25°	-0.005	-0.004	-0.004	-0.004	-0.002	-0.003	-0.004**	0.000
	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)	(0.002)	(0.002)	(0.002)
25°-27.5°	-0.010*	-0.006	-0.004	0.000	-0.003	-0.006**	-0.005**	-0.002
	(0.005)	(0.004)	(0.004)	(0.003)	(0.004)	(0.003)	(0.002)	(0.002)
$> 27.5°$	-0.026***	-0.010**	-0.011**	-0.006	-0.006	-0.004	-0.009***	-0.005*
	(0.007)	(0.005)	(0.005)	(0.004)	(0.004)	(0.004)	(0.003)	(0.003)
Default Mean	0.663	0.696	0.699	0.718	0.754	0.761	0.821	0.831
Obs.	3,714	3,714	3,982	3,982	4,102	4,102	4,480	4,480
Inc. Range	$13,504 – $43,000		$43,004 – $54,633		$54,712 – $72,218		$72,264 – $161,567	

Standard errors in parentheses

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

The table reports estimates from regressing pass rates on temperature bins, demographic controls and school and test date fixed effects after partitioning the data by income quartile. Income is measured by the median household income for the ZIP code of the school's address. Math and ELA sections are estimated separately. Standard errors reported in parenthesis are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years. Below the estimates I report the mean in the omitted default bin and the range of incomes by quartile.

2.4.4 Discussion

To test the robustness of our results, we estimate our models using a variety of control specifications and on a balanced panel of schools with data for all 12 years and find qualitatively similar results.[28] We also test the robustness of the temperature bin specification by running alternative configurations that shift bin boundaries by 0.5°C.[29] The interpretation of the coefficients changes with each specification, but the point estimates for the highest temperature bins generally remain significant. We also estimate our temperature response functions with average temperature and minimum temperature rather than maximum temperature and find the results are not markedly different.[30]

One concern for identification we also considered was the possibility of selection among the student population that is correlated with temperature. For instance, if students were more likely to be absent on hot days, the results would be confounded by differences in student ability. The sign of this bias is not particularly intuitive, but if we think that lower-performing or at-risk students would be more sensitive to environmental factors, this would imply a negative bias. To rule out this concern, we ran our temperature response regression with the number of students tested as the dependent variable. These results show positive and insignificant coefficients for bins above the 17.5-20°C default. This pattern suggests selective attrition is not a major concern for days with higher temperatures.

Taken together, the temperature response functions and the event-study analysis provides strong evidence that contemporaneous higher outdoor temperatures negatively impact student performance on high-stakes exams. The difference in findings for math versus ELA are consistent with the results presented in Graff Zivin, Hsiang, and Neidell (2018). However, these differences should be interpreted with caution. While they could be evidence of a fundamental way thermal stress affects different tasks, they also could be the result of differences in the tested cohorts. As noted in Table 2.2.1, there are minor levels of attrition within cohorts between the ELA and the math dates. Since the ELA section is always administered first, this means the population that takes the mathematics section may be different than the population taking the ELA section. Alternatively, the ordering of the exams could cause fatigue, leading students to be more sensitive to environmental factors on the second date that assesses math. Since we were unable to control for individual student ability across cohorts, we cannot rule out these concerns.

[28] We report the results for all of these exercises in the Appendix.

[29] This yields four alternative different bin specifications, but the results are generally unchanged.

[30] This is likely because the average temperature metric is calculated as the midpoint between the daily maximum and minimum.

Our results also show the effect is more negative for schools in lower income areas. Given the contemporaneous nature of the effects, a strong candidate is the environmental quality of the educational setting, but we cannot rule out other factors correlated with income that affected exam performance. However, all of the results we estimate are agnostic to the presence of any climate control technology at the school. This means that schools with varying levels of heating and AC installation are pooled in the estimate for the effect of temperature. If these technologies negate the effect of extremes, then our estimates provide a lower bound on the effect of temperature. One explanation could be that wealthier schools provided testing environments that were more resilient to extreme temperatures. Data on AC penetration in schools is difficult to come by so we are pursing information on which schools had AC installed using satellite images and electricity consumption data in ongoing work.

2.5 Conclusion

This paper provides quasi-experimental evidence that extreme outdoor temperature on the day of an exam adversely affects student test performance. Our results indicate that student pass rates on the mathematics section were as much as 2 percent lower when the maximum temperature for the day rose above 27.5°C (81°F). The effect of hot temperatures on ELA pass rates was smaller, but also statistically significant.

Our estimates provide a lower bound of the effect of temperature since we cannot explicitly control for the penetration of climate control technologies, but the effect we document appears to be more pronounced for schools in low-income areas. We also find that extreme temperatures prior to the exam had an insignificant effect, suggesting testing environments may have been important. Access to climate-control technologies is a possible factor, but we cannot rule out other mechanisms likely to be correlated with temperature such as social distractions.

The nonlinear relationship shown in both the mean scale scores and the cohort pass rates on the mathematics section is similar to findings in the laboratory and recent papers that also study other standardized tests. Importantly, our results are derived from 48 testing days over a 12 year period with over 3 million student outcomes. This supports a broader interpretation on the effects based on population heterogeneity, but we are unable to comment on the effect of temperature on longer-run educational outcomes. Further research in this area is strongly encouraged as there are likely to be distinct long-run effects of temperature on human capital development.

While the CAHSEE has been discontinued as a requirement for receiving a high school diploma in California, our results raise important questions on the interaction

between environmental factors, income, and high-stakes standardized tests which continue to be a significant feature of the U.S. education system.

Chapter 3

The Economics of Targeting: Theory and Evidence from a Field Experiment[1]

3.1 Introduction

The proliferation of rigorous evaluation methodologies continues to provide a valuable stream of information for policymakers, academics, and decision-makers in industry on the effectiveness of their programs.[2] One recurring theme in this broad literature is to characterize treatment effect heterogeneity using observable characteristics and comment on how the cost effectiveness of program impacts can be maximized through targeting. However, there remains relatively little research testing how well these recommendations perform in the field. This paper addresses this question theoretically and empirically through the use of a field experiment conducted within the product infrastructure of a technology company.

The ability of a given evaluation to predict impacts in other contexts is termed external validity. Yet this term blankets many underlying mechanisms responsible for differences in impacts between settings. For example, heterogeneity in treatment effects between multiple evaluations could due to differences in sample composition, program design, methodology, or even the macroeconomic environment at the time.

[1]The material in this chapter is from a working paper coauthored with Maximilian Balandat and Datong Zhou.

[2]Within economics, the fields of development, labor, industrial organization, and energy economics are increasingly using randomized field experiments to evaluate programs. Further, the world's largest technology companies such as Facebook, Amazon, and Google utilize randomized experiments to evaluate different strategies.

While macroeconomic factors pose challenges to the generalizability of any microeconomic study, there are efforts being made to address the others. Characterizing heterogeneity in treatment effects provides traction for addressing composition and design challenges by estimating fundamental parameters and the standardization of methods can be potentially addressed ex-post as is often done in meta-analyses. Understanding which of these factors dominate is thus useful in predicting the effectiveness of targeting strategies that leverage estimate heterogeneity. In this paper, we provide a theoretical framework for how these factors affect targeting and how experimental evaluation can be used to parse competing mechanisms.

Our theoretical framework contributes to the literature by distinguishing between group-level heterogeneity as defined by covariates and subject-level effects we call individual treatment effects (ITEs).[3] The ITEs we define can only be gleaned through observing program participation using panel data, but provide an additional heterogeneity within the group-level effects. Estimates of ITEs provide criteria on which to target interventions and are increasingly feasible given the number of repeated observations per individual in technology settings.

We partnered with a technology company in order to understand the impact of using ITEs to target in the field. Following the experiment evaluated in Gillan (2017), we cross randomize assignment to a targeting strategy intended to reduce the implementer's costs. The setting we study is the product of a demand response provider (DRP) in California that pays households to reduce their energy consumption and then is compensated for aggregate reductions via the wholesale market. We observe hourly consumption data for 2,844 participants and the specifics of the company's intervention which allows us to estimate ITEs using random forest methods from machine learning (ML). We use the first 90 days of program enrollment to estimate ITEs and then evaluate our targeting strategy for the same households during the following 90 days by assigning half to the targeting and the other half to a control group. The incentives offered the targeted group are non-random and conditional on our ITE estimates while the control group continued to receive incentives in the same manner as the first 90 days.

By randomly assigning households to targeting, we are able to evaluate its causal impact on various outcomes of interest. We find our targeting strategy reduces the costs of the DRP by 52 percent and the results are highly significant. The strategy also reduces revenue by 24 percent, indicating an overall increase in profit on the order of 28 percent. We also examine the persistence of the effects and find the cost savings begin to diminish only 60 days after deployment of the targeting

[3]These are similar in spirit to the random effects models commonly estimated in the labor literature on teacher value added.

strategy. These findings suggest significant potential for reducing the cost of the program as designed in the first 90 days, but only in the short-term. Importantly, the experimental evaluation allows us to understand its performance without having to rely on the common practice of conducting ex-post simulations.

This paper contributes to several strands of literature. More broadly, the paper relates to how heterogeneity in agents affects optimal public policy design.[4] The most relevant is work from Dubé and Misra (2017) which use an experiment to evaluate targeting in the form of third-degree price discrimination for an employment recruiter. They cast their demand estimation as a heterogeneous treatment effects problem with price as the intervention and find their price targeting increases profits captured by the firm significantly. This paper differs from their methodology in that we use target amongst the same population of households to refine the cost-effectiveness of the program as compared to targeting a new sample of participants. This distinction is important for programs that are continued over a long time or expose participants to multiple interventions.

Within the program evaluation literature, there is relatively little empirical evidence on the impact of targeting and none that use an experiments for evaluation (Athey and Imbens 2017). Allcott (2015) provides the most compelling empirical evidence to date on how effective extrapolation performs in predicting program efficiency at different sites at later interventions. He finds using heterogeneity in estimates for 10 early sites to predict the cost effectiveness at 101 future interventions systematically over-predict. Hotz, Imbens, and Mortimer (2005) develop a framework for evaluating external validity in employment training and perform an exercise that precedes the work in Allcott (2015) for 5 sites. However, both of these studies perform out-of-sample predictions ex-post for different sets of subjects.

The paper proceeds as follows: Section 3.2 develops the theoretical framework for targeting and introduces our definition of ITEs, Section 3.3 reports the experimental design and setting, Section 3.4 reports the empirical results of the experiment, and Section 3.5 concludes.

[4]Examples of recent work in this space are Rees-Jones and Taubinsky (2016), that consider how attention heterogeneity among agents affects their tax burden, and Sallee (2018), that deals with the implications of imperfect prediction on transfers necessary for Pareto improvement – a fundamental efficiency criterion in many economic theorems.

3.2 Theory

This section provides a theoretical framework to fix ideas and establish the type of targeting documented in this paper. The section also shows how it compares to other factors policymakers and firms consider in optimizing program efficiency.

3.2.1 Model Setup

Consider a model for evaluating program outcomes that have the following data-generating process (DGP):

$$y_i = \theta_i D_i + \varepsilon_i \tag{3.1}$$

where $D_i \in \{0, 1\}$ represents an indicator whether subject i has received the treatment intervention and y_i represents the objective metric stated in dollars that the policymaker or firm seeks to optimize.[5] Without loss of generality, we assume the effect of no intervention is zero on average. While treatment effects do not necessarily map directly to cost-effectiveness, we combine the two here for ease of exposition. $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is a mean-zero stochastic error term which we assume is uncorrelated with program effects.

Assumption 1. *$\theta_i \sim F_\theta$ with mean μ_θ and variance σ_θ^2.*

The parameter θ_i represents a subject-level effect we will call an individual treatment effect (ITE). θ_i can also be thought of as an agent *type* within a formal economic model where positive (negative) values are associated with types whose marginal participation increases (decreases) efficiency.[6] Importantly, θ_i represents heterogeneity in the cost effectiveness which is separate from residual variation which we assume to be normally distributed.[7]

Assumption 2. *$\forall i, j \in S,\ E[\varepsilon_i | D_j] = 0$ (random assignment and SUTVA)*

[5] For example, an energy efficiency retrofit program would be concerned with the reduction in energy expenditures net of marginal costs. In development economics this could be the health savings net the cost of deploying bed nets. In digital marketing this could be the revenue generated per ad clicks less the amount spent on each individual in the campaign. Numerous other examples exist with the basic structure being marginal return less marginal cost in dollars.

[6] This also generally captures imperfect take-up issues where D_i is assignment to a program encouragement and the θ_i maps to the return conditional on encouragement. For those who take-up the parameter captures the program savings net of cost and those who don't it only captures the costs of encouragement.

[7] This framework is analogous to a random-effects model in panel settings.

We assume the program is evaluated via random assignment of interventions and the stable unit treatment value assumption (SUTVA) holds in the set of experimental sample of individuals, denoted S. Together with Assumption 1, this implies an evaluator can obtain an unbiased estimate $\hat{\theta}$ of $\theta \equiv \mu_\theta = \mathrm{E}[y_i|D_i = 1] - \mathrm{E}[y_i|D_i = 0]$ by taking the difference in mean outcomes for the treatment and control groups or estimating the OLS regression in (3.1).

θ corresponds to the average treatment effect (ATE) for marginal cost effectiveness of the intervention in the sample and is often the focus of program evaluation since it indicates whether the program is potentially efficient. Generally, the experimental program is deemed cost effective if $(\hat{\theta} \cdot |S| - F > 0)$ where F represents the fixed cost of the program. This then prompts discussion of whether there is external validity between the experimental sample and the broader population so that estimates may be constructed for whether a scaling up of the program is worthwhile. If $\hat{\theta} < 0$, we generally think the program is not worth scaling, but performing inference on only θ provides information whether the program should be continued via random assignment and stops short of understanding whether exploiting treatment effect heterogeneity to strategically target future interventions can improve efficiency.

3.2.2 Targeting Interventions to Improve Efficiency

The term "targeting" is used in various ways throughout the literature. In this paper, we define targeting as non-randomly assigning interventions with the goal of improving cost-effectiveness. This framework allows us to derive the testable implications for targeting in our experiment and also provides a concrete foundation for how the process of targeting can be thought of more broadly by policymakers and evaluators.

We first consider the assignment of program interventions to individuals. Let $\hat{\theta}_i$ denote an ITE estimate on the efficiency return of subject i We define a *targeting strategy* as a policy function $g : \mathbb{R} \to \{0, 1\}$ mapping an ITE *estimate* of a subject's efficiency to intervention assignment. Here the model makes no assumption on how $\hat{\theta}_i$ is formulated and $g \in \mathcal{G}$ exists in the function space of all possible strategies. This distinction separates the challenges in estimating θ_i from the multiple assignment rules based on given a set of estimates. Estimates and assignment rules would likely be linked in practice, but it is worth distinguishing between the two dimensions over which program efficiency would be improved in the model. The framework also allows for ad-hoc estimates and rules which target populations selected by a policymaker without any statistical estimation of participant efficiency.

For example, the model accommodates an expert elicited score for the effect each individual will have on the program serving as $\hat{\theta}_i$.[8] The targeting strategy would then be the rule which stated scores above some threshold would be assigned to interventions. Alternatively, an evaluator could assign interventions to individuals from demographic groups that have shown positive returns in the past.

Definition 1 states when we can claim one strategy is more efficient than another, conditional on a set of individual estimates of efficiency returns.

Definition 1 (Policy Design Channel)**.** *For a set of ITE estimates $\{\hat{\theta}_i\}_{i\in S}$ and two distinct targeting strategies $g_1, g_2 \in \mathcal{G}$ where $g_1 \neq g_2$, g_1 (weakly) dominates g_2 if*

$$E[y_i|D_i = g_1(\hat{\theta}_i)] \geq E[y_i|D_i = g_2(\hat{\theta}_i)]. \tag{3.2}$$

In (3.2) we do not require that D_i be randomly assigned because we are no longer seeking an estimate of μ_θ, the mean of the distribution in the sample. Instead our goal is to determine if the mean in treatment effects conditional on assignment is greater among one population or the other.[9] Definition 2 states an analogous condition to Definition 1 for two sets of ITEs under a given assignment rule:

Definition 2 (Information Channel)**.** *For a given targeting strategy $g \in \mathcal{G}$ and two distinct sets of ITE estimates $\Theta_1 = \{\hat{\theta}_i^1\}_{i\in S}$ and $\Theta_2 = \{\hat{\theta}_i^2\}_{i\in S}$, estimates Θ_1 yield a more efficient program than Θ_2 if and only if*

$$E[y_i|D_i = g(\hat{\theta}_i^1)] > E[y_i|D_i = g(\hat{\theta}_i^2)]. \tag{3.3}$$

Distinguishing between the policy design or assignment rule and the information set is important because improving the ITE estimates is likely to come at a cost. That is, improving characterization of heterogeneity might require additional expertise on how to refine estimates. If this information is costly, a change in the policy design may be more appealing.

Lastly, we establish the condition which are necessary for there to be any potential in improving efficiency. Proposition 1 states that there must be treatment effect

[8] In energy efficiency programs this could come from an engineering model. For labor training programs, this could come as an evaluation from an expert over a candidate's capability of retraining.

[9] To see this formally, let $G_1, G_2 \subseteq S$ denote the set of individuals targeted using two distinct strategies g_1 and g_2. The LHS of (3.2) estimates $E[\theta_i|i \in G_1]$ the RHS is the quantity $E[\theta_i|i \in G_2]$. If the ITEs are on average higher in G_1 than G_2, then strategy g_1 improves efficiency more than g_2.

heterogeneity in the sample for there to be any advantage to targeting individuals differently.

Proposition 1. *There must be heterogeneity in treatment effects, i.e., $\sigma_\theta > 0$, in order for targeting to improve program efficiency. Proof: See Appendix*

Important to note in Proposition 1 is that heterogeneity in the efficiency return per unit cost is separate from variation in the error terms σ_ε^2, which cannot be characterized using variation in program assignment.

3.2.3 Individual versus Group Targeting

In this section we discuss the practice for estimating $\hat{\theta}_i$ and the conditions under which targeting designs that seek to characterize individual effects (the strategy evaluated in our experiment) may be preferable to those that use demographic data to group individuals.

Due to the fundamental challenge of causal inference, obtaining unbiased estimates of each *individual* θ_i is infeasible using cross-sectional data. Thus, researchers often rely on estimates of the ATE for a *group* of individuals with the same covariate values. This is the most common practice in the literature and recent work has provided substantial innovation in systematically determining strategies for which covariates best explain effect heterogeneity.

However, this paper contrasts with the extant literature that seeks to capture group effects for the purpose of out-of-sample prediction to individuals in other groups. Instead, we focus on ITEs that can only be gleaned from program participation and consider the exercise of targeting interventions among existing users to maximize efficiency.

To clarify the distinction consider modifying the DGP in (3.1) to the following:

$$y_i = \xi_i D_i + D_i X_i \delta + \varepsilon_i \tag{3.4}$$

where X_i is a vector of k covariates for individual i and δ is a vector of treatment effects for individuals by covariate. X_i is not included as a control in (3.4) for parsimony since the treatment effects are of interest, but including them would only increase precision if treatment is randomized. The treatment effect heterogeneity in (3.1) can be mapped to the ITE from (3.1) as $\theta_i = \xi_i + X_i\delta$.

For simplicity, assume X_i is composed of indicator variables for various groups. Then for each covariate group, the average effect of the program $\delta_k = \mathrm{E}[\theta_i | X_{i,k} = 1]$ and an unbiased estimate can be obtained by taking the difference in means between

treated and untreated individuals inside the group

$$\hat{\delta}_k \equiv \mathrm{E}[y_i|D_i = 1, X_{i,k} = 1] - \mathrm{E}[y_i|D_i = 0, X_{i,k} = 1]. \tag{3.5}$$

By iterating through the covariate space, a program evaluator could discover which groups were estimated to have positive returns and which did not. This information could be used then to construct $\hat{\theta}_i$ estimates for targeting consistent with the strategies defined in Section 3.2.2.[10] However, these strategies are inherently limited by the set of covariates available to the researcher.

It is possible that group-level ATEs leave substantial heterogeneity unexploited and in this case, subject-level ITEs may be valuable. Within this framework, individual targeting will only improve efficiency over group targeting if $Var(\xi_i) > 0$. This condition is analogous to Proposition 1 for the residual treatment effect variation not captured by the group treatment effect. Since there is no feasible way to estimate ξ_i using cross-sectional data we consider a panel data setting where you observe individuals pre- and post-treatment. These settings are becoming increasingly common in digital environments, electricity consumption, and are likely to increase with the growing number of individual-level data being collected.

3.3 Research Design and Setting

This section describes the experimental setting and design used to evaluate targeting and program efficiency.

3.3.1 Experimental Setting

The setting we study is the product environment of Ohmconnect, a third-party residential demand response provider (DRP). Demand response is a class of programs created by electricity market regulators to more dynamically price consumption with the goal of improving overall system efficiency. Since prices are typically static and costs are highly variable, sending higher prices to elastic consumers during high cost periods has the potential to improve allocative efficiency.[11]

Households that decide to enroll with the DRP program receive one-hour incentive events 1-3 times per week that financially incentivize them to reduce their electricity

[10] For example, let $\hat{\theta}_i = \theta_x$ for $\{i|X_i = x\}$ and define the set of groups with positive ATE estimates as $T = \{x \in \mathcal{X}|\hat{\theta}_x > 0\}$ where $\mathcal{X}$ is the set of all possible group permutations. The targeting function could then be defined as $g(\hat{\theta}_i) = \{1 \text{ if } X_i \in T, 0 \text{ if } X_i \notin T\}$.

[11] See Borenstein and Holland (2005), Joskow and Wolfram (2012), and Gillan (2017) for a deeper discussion on the social value of dynamic electricity pricing.

consumption. Households are given short notice – up to 1 hour ahead, but on average 5 minute – and are rewarded (penalized) points if their consumption is less (greater) than an individual forecast. The payout to household i over a set of event periods T is thus given by the following formula:

$$Payout_i(p_i, T) = \sum_{t \in T} p_{it} \cdot (\hat{q}_{it} - q_{it}(p_i; \eta_i)). \tag{3.6}$$

The incentive each period t is given by p_{it} and is offered in points per kWh. The input p_i denotes the vector of prices shown to individual i over events T. The forecast $\hat{q}_{it}$ and consumption $q_{it}(p)$ are measured in kWh. Points can be cashed out electronically at a rate of \$1 per 100 points.[12] For example, if $p = 50$ and a household's forecast is 1.2 kWh, they would gain 10 points if they consumed 1 kWh and they would lose 15 points if they consumed 1.5 kWh. The forecast is estimated using past consumption observations which creates the potential for "gaming" in favor of payment, but the short notice and unpredictability of when events are called suggests this is a minor issue.[13]

Similar to the formula in (3.6), the DRP generates revenue from the wholesale market based on the difference between forecast and actual aggregate consumption multiplied by the clearing price for that period.[14] The return for the same set of event periods is thus given by

$$Return(T) = \underbrace{\sum_{t \in T} p_t^w \cdot (\hat{Q}_t - Q_t)}_{Revenue} - \underbrace{\sum_{i \in S} Payout_i(p_i, T) - c(S, T)}_{Costs} \tag{3.7}$$

where p_t^w is the wholesale clearing price in period t, $Q_t \equiv \sum_{i \in S} q_{it}$ is aggregate consumption for the sample of households S, and $\hat{Q}_t$ is the aggregate forecast. Individual and aggregate forecasts are estimated using the same formula of lagged consumption. The final term in (3.7) denotes the cost of the event schedule T for a sample set of households S and we assume there is a constant marginal cost per household-event such that $c(S, T) = k|S||T|$. We assume the wholesale price is exogenous and leave

[12] Survey evidence from a non-random subset of the sample suggests over 70% of respondents understand the monetary value of points.

[13] Similar programs that use day-ahead notice have been shown to suffer from this issue (Wolak 2006), but Gillan (2017) shows there is no evidence of systematic change in consumption in non-event periods.

[14] The DRP also can generate revenue from the market for demand response capacity which features similar accounting for performance, but a payment conditional on provision. See Bushnell, Flagg, and Mansur (2017) for a deeper description of how capacity markets function in the setting.

the impact the prices have on system efficiency to other work. Instead, we focus on minimizing payouts conditional on an event being called in order to maximize the return as written (3.7).

We evaluate the impact of a specific targeting strategy on return using a natural experiment rolled out during the first 180 days a household enrolled in the program.[15] The experiment had two phases as shown in Figure 3.3.1a. During Phase 1, the first 90 days of enrollment, all customers received the same incentive levels: 5, 25, 50, 100, or 300 points per kWh with equal probability.[16] Phase 2 occurred for days 91-180 of enrollment, during which half of the households were randomly assigned to a treatment "Targeted Group" and the remaining half to a "Control Group".[17] Control households continued to receive the same incentive distribution for the remainder of the experiment. Households in the Targeted group received incentives based on the targeting strategy described below.

3.3.2 Targeting Strategy

Due to constraints in the implementer's business model and product infrastructure, we were unable to unregister customers from the program to evaluate the optimal targeting strategy. While this means we cannot comment empirically on the theoretical first-best, our strategy was designed to reduce costs and the randomization of individuals into targeting allows us to measure the causal impact of our feasible targeting strategy.[18]

Our strategy involved partitioning the incentives individuals received by their estimated ITE of receiving an event message. Gillan (2018) finds that individuals are extremely insensitive to the level of the incentive they are presented during the first 90 days. That is, $q_{it}(p_{it}; \eta_i) \approx \eta_i \cdot 1(p_{it} > 0)$. Optimally, the DRP would send all individuals infinitesimally small incentives to minimize payouts, but due to constraints within the program we were only able to re-assign the incentives while keeping approximately the same average incentive level across all individuals in an event. Thus, we identify the "most responsive" (MR) individuals and send them

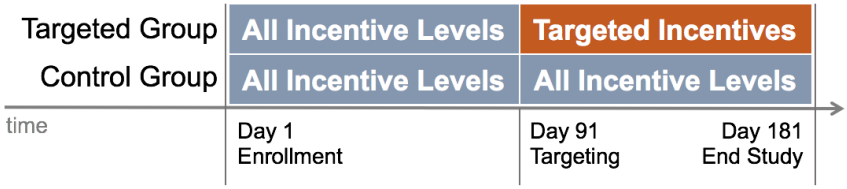
[15] Following the language of Harrison and List (2004), natural experiment refers to the fact that participants experienced the product without being notified that they were in a study.

[16] Specifically each household's incentive was randomly drawn from the five stated levels with equal probability each time an event was called.

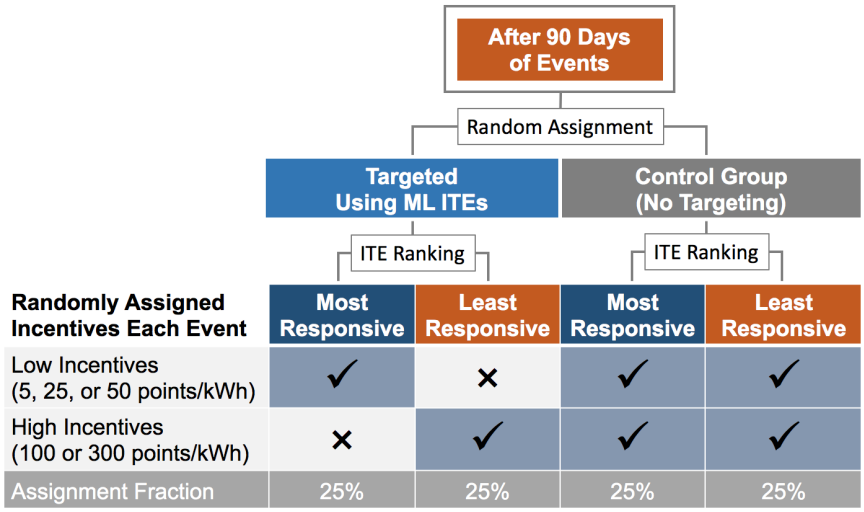
[17] The analysis was performed weekly and our assignments were sent over every Wednesday for users that would reach the 91 day mark over the following week. The next day (Thursday) the DRP would update its databases of our assignments and participants would automatically be rolled into Phase 2 on their respective day 91. This means there was some minor variation in the number of Phase 1 days used, ranging from 83-90 days.

[18] We characterize the optimal strategy in the Appendix.

Figure 3.3.1: Experimental Design



(a) Timeline



(b) Treatment Assignment

Panel (a) shows the timeline of assignment for the first 180 days that a household enrolls with the implementer. Panel (b) shows the assignment into treatment interventions for phase 2, days 91-180 after a household enrolls.

lower incentives (5, 25, and 50 points per kWh) to minimize their payout and send the "least responsive" (LR) individuals higher incentives (100 and 300 points per kWh).[19] Within the Control Group individuals continue to receive all five incentive levels with equal probability. Figure 3.3.1b shows how assignment was performed and the dimensions across which comparison reveals compositional differences as compared with differences due to variation in the incentive levels.

We estimate the ITEs using each households consumption and temperature data for the first 90 days of program interventions. The model for each household is

[19]Specifically, in the Targeted Group those ranked MR receive each levels with $\frac{1}{3}$ probability and those ranked LR receive each level with $\frac{1}{2}$ probability.

an auto-regressive specification of consumption and temperature during the 5 hours leading up to an event along with individual by hour-of-day by month-of-year effects. The model is estimated using random-forests and is trained on days where no events were called. This allows us to estimate a counterfactual for the event period. We average the differences between each household's actual consumption and our counterfactual estimates for all events during the first 90 days. This metric provides the ITE estimate that we use to rank households.[20]

3.4 Empirical Results

3.4.1 Data and Sample

We measure the impact of the targeting strategy using a panel with data provided by the DRP that includes customers electricity usage during events and the hours in between. We also observe the individual forecasts and incentives received each event.[21] We use this data to construct the household-level payout per event during the targeting period consistent with (3.6). We also merge in hourly temperature data based on the ZIP code of the household.[22]

We evaluate the targeting strategy for 52 events that occurred between June 30, 2017 and November 30, 2017. The primary sample consists of 2,843 households who experienced the strategy for at least one week. The average number of events received per household was 18 with a minimum of 1 and a maximum of 30. Some households did not receive the full 90 days of targeting because the experiment concluded on a fixed date before they reached the 180 day point from when they signed up. We also analyze a subsample of 486 households who received the complete 90 days of targeting to explore the persistence of the effects. The impact of targeting is causally identified for both samples, but the ATEs reported for each will have different compositions and interpretations. We report summary statistics and tests on the balance of random assignment in the Table 3.4.1.

[20] We report more details on the targeting algorithm in the Appendix.

[21] Gillan (2018) describes the data acquisition and recruitment process for households into the sample in more detail. The energy data comes from Advanced Metering Infrastructure meters which it gathers from each household's respective utility. The eligible population was any customer living in the three major California investor-owned utilities – Pacific Gas & Electric, Southern California Edison, and San Diego Gas & Electric.

[22] The temperature data comes from the Dark Sky API which combines data from multiple government sources and uses a meteorological model to predict weather for a given geocode. Information on the sources used in the Dark Sky API model are summarized here http://darksky.net/dev/docs/sources.

Table 3.4.1: Pre-period Balance and Summary Stats

	Full Sample			90-day Sample			Full vs.
	Control	Targeted	T − C p-value	Control	Targeted	T − C p-value	90-day p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Payout (points)	15.8	16.6	0.597	7.54	9.63	0.559	0.000
	(138.1)	(139.5)		(141.7)	(131.2)		
Reduction (kWh)	0.153	0.162	0.500	0.069	0.080	0.733	0.000
	(0.971)	(0.973)		(1.005)	(0.918)		
Consumption (kWh)	0.872	0.866	0.861	1.011	0.768	0.010	0.620
	(1.38)	(1.37)		(1.49)	(1.37)		
Incentive	96.7	96.2	0.456	97.4	98.2	0.654	0.113
(points/kWh)	(107.2)	(106.8)		(108.3)	(107.5)		
CDH	7.34	7.11	0.102	7.16	6.53	0.100	0.027
	(0.10)	(0.10)		(6.10)	(5.93)		
HDH	0.222	0.256	0.043	0.421	0.520	0.142	0.000
	(0.875)	(0.949)		(1.20)	(1.34)		
1(Automated)	0.081	0.093	0.247	0.078	0.116	0.165	0.419
	(0.273)	(0.291)		(0.269)	(0.320)		
Hour-of-Day	15.2	15.2	0.217	15.4	15.4	0.919	0.000
	(2.35)	(2.35)		(2.38)	(2.37)		
Observations	39,716	40,964		6,844	6,734		
Households	1,442	1,401		245	241		

Standard deviations in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The table reports summary statistics for the pre-period consisting of days 1-90 after enrollment. Columns (1)-(3) report results for the full sample and (4)-(6) for the sample of households that complete 90 days of targeting. Columns (1) and (4) report means in the control group and (2) and (5) for the targeted group. Columns (3) and (6) report the p-value on testing the null hypothesis that the mean for the control and targeted groups in each respective sample is the sample. Parentheses report standard deviations of each statistic.

3.4.2 The Effect of Targeting

We estimate the effect of targeting on returns using the following econometric model:

$$Y_{it} = \delta_0 + \delta_1 D_i + \varepsilon_{it} \tag{3.8}$$

where Y_{it} the outcome for individual i in event period t, D_i is an indicator that individual i was assigned the targeting strategy, and ε_i is the error term. The coefficient of interest is δ_1 which captures the ATE of the targeting strategy per event per individual. The coefficient δ_0 captures the mean outcome in the control group. We leave the data in panel form to allow for the addition of event-level fixed effects and temperature controls. These variables are not needed for identification, but improve the precision of the estimates. We cluster errors by individual to account for arbitrary within-individual correlation in the residuals.

Table 3.4.2 reports the impact of the targeting strategy on various outcomes of interest. Columns (1), (2), and (3) show the results of running the regression specification in (3.8) with payouts in points, measured reductions in kWh, and usage in kWh as the dependent variable, respectively. Column (1) indicates that on average the DRP pays individuals 7.86 points per user per event in the control group, but the payouts are 4.12 points lower for households that receive the targeted incentives. The effect is significant with a p-value less than 0.001 and indicates a 52 percent decrease in payouts. This effect is large and suggests costs can be dramatically reduced with the strategy.

In order to quantify the effect on the benefits of the targeting strategy to the DRP we estimate its impact on measured reductions and consumption. The measured reduction is simply the forecast less the consumption. We report the consumption to show the effect on behavior absent the forecast. Column (2) shows the effect of the targeting strategy on measured reductions. On average, household reductions are 0.086 kWh per event without the targeting strategy and about 0.020 kWh lower, a 24 percent decrease.[23] The effect is borderline significant with a p-value of 0.04. This suggests revenues may be lower under the targeting strategy. Column (3) shows the effect on consumption is also negative, but the estimate is insignificant suggesting the effect on reductions comes from the forecast.[24]

Columns (4)-(6) of Table 3.4.2 report estimates on the heterogeneity of the treatment effect by our ITE ranking of responsiveness to MR or LR. The results are

[23] This value is roughly the same as the ATE reported from the first 90 days in Gillan (2018)

[24] Regressing the treatment indicator on the forecast yields an ATE estimate of -0.023, but this is also insignificant. The significance of the measured reduction results from the fact that the combination of the two outcomes has a lower variance.

Table 3.4.2: Effect of Targeting Strategy

	Average Treatment Effects			Heterogeneity by ITE Rank		
	Payout (Points)	Reduction (kWh)	Usage (kWh)	Payout (Points)	Reduction (kWh)	Usage (kWh)
	(1)	(2)	(3)	(4)	(5)	(6)
Intercept (Control)	7.86***	0.086***	0.658***			
	(0.82)	(0.008)	(0.017)			
$1(Most\ Resp.)_i$				10.87***	0.125***	0.685***
				(1.33)	(0.013)	(0.025)
$1(Least\ Resp.)_i$				4.55***	0.042***	0.631***
				(0.85)	(0.006)	(0.022)
D_i (Targeted)	−4.12***	−0.020**	−0.003			
	(1.05)	(0.010)	(0.023)			
$D_i \times 1(Most\ Resp.)_i$				−8.25***	−0.023	−0.003
				(0.023)	(0.017)	(0.033)
$D_i \times 1(Least\ Resp.)_i$				0.34	0.015*	−0.003
				(0.023)	(0.008)	(0.032)
Observations	52,323	52,323	52,323	52,323	52,323	52,323

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

This table reports estimates on the ATE of targeting for the full sample of 2,844 households. Columns (1)-(3) report the ATE estimates for all households and columns (4)-(6) report ATE estimates broken out by ITE rank. Columns (1) and (4) use individual payouts in points as the dependent variable, columns (2) and (5) use measured reductions in kWh, and columns (3) and (6) use usage/consumption in kWh. Standard errors are clustered by household and reported in parentheses.

obtained by interacting the intercept and treatment indicator with indicators for a household's ITE rank. The treatment estimates can be interpreted as causal since we know the hypothetical assignment of individuals in the control group. Columns (4), (5), and (6) report the results from separate regressions with payout, reduction, and usage as the dependent variable, respectively. Rows 2 and 3 of Table 3.4.2 show differences in outcomes for the control group. MR households received more than twice as much in payouts each event, but had measured reductions nearly three times as large. Interestingly, MR households also had higher usage during events, but this is due to the fact that they also tended to consume more in any given hour.

Rows 5 and 6 show how targeting affected households by ITE rank. The estimates show almost the entire reduction in payouts comes from MR households, which re-

ceived 76 percent less in payouts due to the targeting strategy. This was due, in part, to the fact that they no longer received 100 or 300 point per kWh incentive levels. MR households also reduced less, but the difference is insignificant without adding controls to improve precision. The effect on LR households insignificant for payout and usage, but there is a borderline significant effect on reductions. LR households were consistently sent 100 and 300 point per kWh incentives and this outcome would be consistent with them reducing more due to the higher incentives.

3.4.3 Persistence

In this section we examine the persistence of the effect of the targeting strategy over the 90 day period. For this exercise, we limit the balanced panel of 484 households that completed the full 90 day period and include fixed effects and controls to improve the precision of our estimates. Specifically we estimate the following event-study specification:

$$Y_{ith} = \sum_{\tau=-3}^{2} \delta_\tau D_i \times \mathbf{1}(t \in P_\tau^{30}) + X_{it}'\beta + \alpha_{ih} + \gamma_t + \varepsilon_{it}. \tag{3.9}$$

where Y_{it} and D_i are the outcome and treatment indicators as defined in (3.8) and $1(t \in P_\tau^{30})$ are a set of indicators that the event in hour t is occurring τ periods relative to the date of targeting assignment. Period $\tau = 0$ denotes the first 30 days of targeting, or days 91-120 after enrolling with the DRP. We omit period -1 which maps to the 30 days prior to targeting assignment so estimates can be interpreted as changes relative to that period. We include household by hour of day fixed effects α_{ih}, hour-of-sample fixed effects γ_t, and temperature controls X_{it} to improve precision. ε_{it} denotes the error term.

Figures 3.4.2a and 3.4.2b reports the ATE results of estimating (3.9) for payouts and measured reductions, respectively. The gray dashed line indicates the break between the pre-period before the targeting strategy was deployed. Vertical bars plot 95 percent confidence intervals constructed using standard errors clustered by household. The results show reductions in payouts are significantly different for the targeted group for the first 60 days following the strategy's deployment, but the payouts increase in the final 90 day period.[25] The balanced panel point estimates are about twice as larger than those for the full sample.[26] This is likely a result of the targeting period for these households occurring during hotter summer months when

[25]The fact that the pre-period estimates are insignificant and generally closer to zero is encouraging evidence that the research design is intact.

[26]This is shown by the analogous results to Table 3.4.2, which I report in the Appendix.

treatment effects and measured reductions were generally larger due to cooling load. Point estimates on reductions are also negative, but insignificant for each period. The pooled effect is -0.073 with a p-value of less than 0.01. The insignificance is due in part to the smaller sample size and the power of estimating effects for each 30 day period.

Figures 3.4.2c and 3.4.2d show the same event study results broken out by ITE rank. Again these results can be interpreted causally for each group. The results show the targeting strategy statistically significantly decreased payouts among MR households for the first 60 days of targeting, but then were indistinguishable from payouts in the pre-period during the final 30 days. LR household payouts follow a similar pattern to MR households, but the estimates insignificant from pre-period levels for all 30 day periods. Notably the effect on reductions remains low and persistent even during the final 90 day period. If measured reductions are correlated with actions, this suggests the negative impact on revenue could be persistent while cost decrease is short-lived.

3.4.4 Discussion

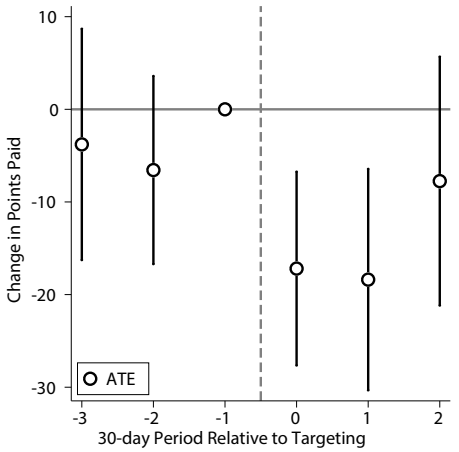
The estimates presented above show the targeting strategy we implemented significantly reduced costs as predicted. The 52 percent reduction in costs is likely to be smaller than the negative impact on revenue, which is likely to be around 25 percent. Together these results imply the targeting strategy increased the profit by around 25 percent. Given the fact that the strategy we tested was not the theoretical optimal, the size of these findings suggest substantial gains to the firm to implementing targeting.

However, the event-study results indicate the persistence of the effects may be short-lived. The reduction in costs appears to shrink only 60 days after rollout and is concentrated in the households who were most responsive. This would be consistent with the intuition that sending households consistently lower incentives compared with what they have come to expect reduces overall participation as measured through reductions. If these behavior changes are persistent, the longer-term benefits of the targeting strategy may be reduced.

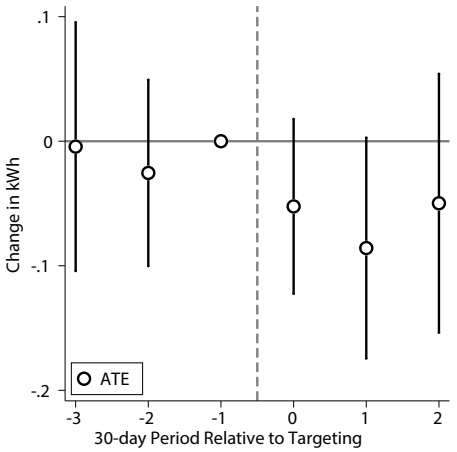
3.5 Conclusion

This paper provides some of the first experimental evidence on the effect of a targeting strategy on program outcomes. We show the efficiency of the program studied here could be significantly increased by lowering the incentives offered to the most

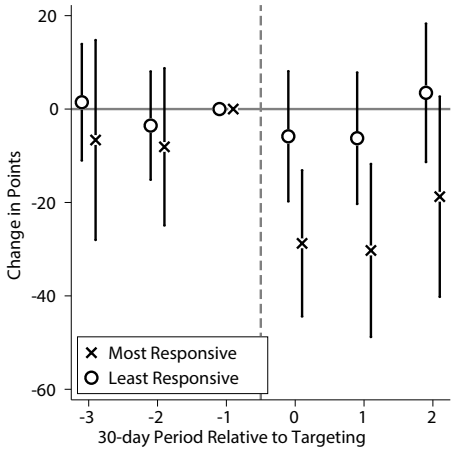
Figure 3.4.2: Persistence of Targeting Effect



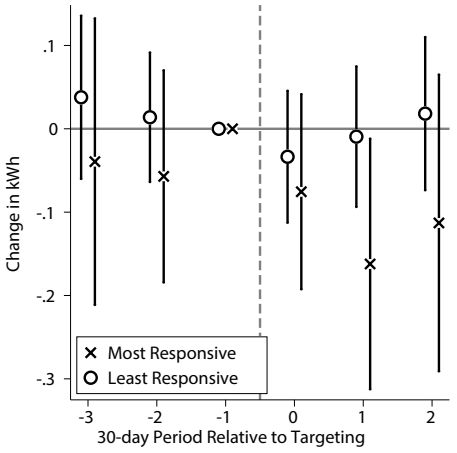
(a) Effect on Payouts



(b) Effect on Reductions



(c) Effect on Payouts by ITE Rank



(d) Effect on Reductions by ITE Rank

These figures plot event-study estimates on the effect of the targeting strategy over time for the balanced panel. The time intervals are 30 day periods 90 days prior to and following the start of the targeting strategy. The gray dashed line indicates the division between pre- and post-periods. Panels (a) and (c) represent estimates for payouts and panels (b) and (d) on measured reductions. Panels (a) and (b) show the ATE estimates for all households and panels (c) and (d) break out the effect by ITE rank. Vertical bars indicate 95 percent confidence intervals estimated using standard errors clustered by household. The omitted category is period -1 meaning the estimates represent changes relative to the period 30 days prior to rollout.

responsive participants, but that the gains could be short-lived. The targeting strategy we implement reduces costs by more than twice the amount it reduces revenue, but the improvements in cost-effectiveness decrease substantially only 60 days after deployment. We also introduce a theoretical framework to fix ideas over how and when targeting can improve program efficiency. Ongoing related work will seek to better understand the channels through which targeting operates with simulation until a broader set of experiments is feasible.

While the financial value of the targeting strategy here may be limited, our paper demonstrates the research value of evaluating targeting based on treatment effect heterogeneity using experiments. Given that this is a growing strategy among the worlds largest technology companies, this provides a unique opportunity to study how individual data can be leveraged to improve program outcomes. These opportunities will likely continue to grow as firms and policymakers make increasing use of big data.

References

Allcott, Hunt. 2011a. "Rethinking real-time electricity pricing." *Resource and Energy Economics* 33 (4): 820–842.

———. 2011b. "Social norms and energy conservation." *Journal of Public Economics* 95:1082–1095.

———. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117–1165.

Allcott, Hunt, and Dmitry Taubinsky. 2015. "Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market." *American Economic Review* 105 (8): 2501–2538.

Athey, Susan, and Guido W Imbens. 2017. "The State of Applied Econometrics: Causality and Policy Evaluation." *Journal of Economic Perspectives* 31 (2): 3–32.

Balandat, Maximilian, James Gillan, and Datong Zhou. 2018. "The Economics of Targeting: Theory and Evidence from a Field Experiment."

Barreca, Alan, Karen Clay, Olivier Deschênes, Michael Greenstone, and Joseph S. Shapiro. 2013. "Adapting To Climate Change: The Remarkable Decline in the U.S. Temperature-Mortality Relationship Over the 20th Century."

Bartoš, Vojtech, Michal Bauer, Julie Chytilová, and Filip Matějka. 2016. "Attention Discrimination: Theory and Field Experiments with Monitoring Information Acquisition." *American Economic Review* 106 (6): 1437–1475.

Bernheim, B. Douglas, and Antonio Rangel. 2009. "Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics." *Quarterly Journal of Economics,* no. February: 51–104.

Bollinger, Bryan K., and Wesley R Hartmann. 2017. "Information versus Automation and Implications for Dynamic Pricing."

Borenstein, Severin, and Stephen Holland. 2005. "On the efficiency of competitive electricity markets with time-invariant retail prices." *RAND Journal of Economics* 36 (3): 469–493.

Borenstein, Severin, and Garth Saloner. 2001. "Economics and Electronic Commerce." *Journal of Economic Perspectives* 15 (1): 3–12.

Bushnell, James, Michaela Flagg, and Erin Mansur. 2017. "Capacity Markets at a Crossroads."

Carleton, Tamma A., and Solomon M. Hsiang. 2016. "Social and economic impacts of climate." *Science* 353 (6304).

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte C. Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions." *Quarterly Journal of Economics* 124 (4): 1639–1674.

Chetty, Raj. 2012. "Bounds on Elasticities With Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply." *Econometrica* 80 (3): 969–1018.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review* 99 (4): 1145–1177.

Cramer, Judd, and Alan B. Krueger. 2016. "Disruptive Change in the Taxi Business: The Case of Uber." *American Economic Review: Papers & Proceedings* 106 (5): 177–182.

Cullen, Julie Berry, Steven D. Levitt, Erin Robertson, and Sally Sadoff. 2013. "What Can Be Done To Improve Struggling High Schools?" *Journal of Economic Perspectives* 27, no. 2 (February): 133–152.

Daly, Christopher, and Kirk Bryant. 2013. *The PRISM Climate and Weather System - An Introduction.* Technical report June. PRISM Climate Group.

DellaVigna, Stefano. 2009. "Psychology and Economics: Evidence from the Field." *Journal of Economic Literature* 47, no. 2 (May): 315–372.

Deschênes, Olivier. 2012. "Temperature, Human Health, and Adaptation: A Review of the Empicial Literature."

Deschênes, Olivier, and Michael Greenstone. 2011. "Climate Change, Mortality, and Adaptation: Evidence from Annual Fluctuations in Weather in the US." *American Economic Journal: Applied Economics* 3 (1): 152–185.

Dubé, Jean-pierre, and Sanjog Misra. 2017. "Scalable Price Targeting."

Faruqui, Ahmad, and Sanem Sergici. 2010. *Household response to dynamic pricing of electricity: A survey of 15 experiments,* 38:193–225. 2.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. n.d. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *Quarterly Journal of Economics.*

Fowlie, Meredith, Catherine Wolfram, C. Anna Spurlock, Annika Todd, Patrick Baylis, and Peter Cappers. 2017. "Default Effects and Follow-On Behavior: Evidence from an Electricity Pricing Program."

Gillan, James M. 2017. "Dynamic Pricing, Attention, and Automation: Evidence from a Field Experiment in Electricity Consumption."

Graff Zivin, Joshua, Solomon M. Hsiang, and Matthew Neidell. 2018. "Temperature and Human Capital in the Short and Long Run." *Journal of the Association of Environmental and Resource Economists* 5 (1): 77–105.

Graff Zivin, Joshua, and Matthew Neidell. 2014. "Temperature and the Allocation of Time: Implications for Climate Change." *Journal of Labor Economics* 32 (1): 1–26.

Ham, John C., Jacqueline S. Zweig, and Edward Avol. 2014. "Pollution, Test Scores and Distribution of Academic Achievement: Evidence from California Schools 2002-2008."

Hancock, P. A., Jennifer M. Ross, and James L. Szalma. 2007. "A Meta-Analysis of Performance Response Under Thermal Stressors." *Human Factors* 49 (5): 851–877.

Hancock, P. A., and I. Vasmatzidis. 2003. "Effects of heat stress on cognitive performance: the current state of knowledge." *International Journal of Hyperthermia* 19 (3): 355–372.

Harding, Matthew, and Steven Sexton. 2017. "Household Response to Time-Varying Electricity Prices." *Annual Review of Resource Economics* 9:337–359.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature* XLII (December): 1009–1055.

Hocking, C, R B Silberstein, W M Lau, C Stough, and W Roberts. 2001. "Evaluation of cognitive performance in the heat by functional brain imaging and psychometric testing." *Comparative biochemistry and physiology. Part A, Molecular & integrative physiology* 128, no. 4 (April): 719–34.

Hossain, Tanjim, and John Morgan. 2006. "...Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay." *Advances in Economic Analysis & Policy* 6 (2).

Hotz, V. Joseph, Guido W. Imbens, and Julie H. Mortimer. 2005. "Predicting the efficacy of future training programs using past experiences at other locations." *Journal of Econometrics* 125 (1-2 SPEC. ISS.): 241–270.

Houde, Sebastien. n.d. "How Consumers Respond to Product Certification and the Value of Energy Information." *RAND Journal of Economics.*

Hughes, Jonathan E., and Molly Podolefsky. 2015. "Getting Green with Solar Subsidies: Evidence from the California Solar Initiative." *Journal of the Association of Environmental and Resource Economists* 2 (2): 235–275.

Ito, Koichiro. 2014. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *American Economic Review* 104, no. 2 (February): 537–563.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2017. "Moral Suasion and Economic Incentives : Field Experimental Evidence from Energy Demand."

Jessoe, Katrina, and David Rapson. 2014. "Knowledge is (less) power: Experimental evidence from residential energy use." *American Economic Review* 104 (4): 1417–1438.

Joskow, Paul L. 2012. "Creating a Smarter U.S. Electricity Grid." *Journal of Economic Perspectives* 26 (1): 29–48.

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

Kahneman, Daniel. 2003. "Maps of Bounded Rationality: Psychology for Behavioral Economics." *American Economic Review* 93 (5): 1449–1475.

Kahneman, Daniel, Ilana Ritov, and David Schkade. 1999. "Economic Preferences or Attitude Expressions?: An Analysis of Dollar Responses to Public Issues." *Journal of Risk and Uncertainty* 19 (1-3): 203–235.

Lavy, Victor, Avraham Ebenstein, and Sefi Roth. 2012. "Ambient Air Pollution, Cognitive Performance and Long Term Consequences for Human Capital Formation."

Papay, John P., Richard J. Murnane, and John B. Willett. 2010. "The Consequences of High School Exit Examinations for Low-Performing Urban Students: Evidence From Massachusetts." *Educational Evaluation and Policy Analysis* 32, no. 1 (February): 5–23.

Park, Jisung. 2018. "Hot Temperature and High Stakes Exams: Evidence from New York City Public Schools."

Ransom, Michael R., and C. Arden Pope III. 1992. "Elementary School Absences and PM10 Pollution in Utah Valley." *Environmental Research* 58 (1): 204–219.

Reardon, Sean F., Nicole Arshan, Allison Atteberry, and Michal Kurlaender. 2010. "Effects of Failing a High School Exit Exam on Course Taking, Achievement, Persistence, and Graduation." *Educational Evaluation and Policy Analysis* 32, no. 4 (November): 498–520.

Rees-Jones, Alex, and Dmitry Taubinsky. 2016. "Heuristic Perceptions of the Income Tax: Evidence and Implications for Debiasing."

Sallee, James M. 2011. "The Surprising Incidence of Tax Credits for the Toyota Prius." *American Economic Journal: Economic Policy* 3 (2): 189–219.

———. 2014. "Rational Inattention and Energy Efficiency." *Journal of Law and Economics* 57 (August): 781–820.

Schlenker, Wolfram, and Michael J. Roberts. 2009. "Nonlinear temperature effects indicate severe damages to U.S. crop yields under climate change." *PNAS* 106, no. 37 (September): 15594–8.

Seppänen, Olli, William J Fisk, and Q H Lei. 2006. "Effect of Temperature on Task Performance in Office Environment."

Sims, Christopher A. 2003. "Implications of rational inattention." *Journal of Monetary Economics* 50:665–690.

Staal, Mark A. 2004. *Stress, Cognition, and Human Performance: A Literature Review and Conceptual Framework.* Technical report August. NASA.

Wolak, Frank A. 2006. "Residential Customer Response to Real-Time Pricing: The Anaheim Critical-Peak Pricing Experiment."

———. 2010. "An Experimental Comparison of Critical Peak and Hourly Pricing: The PowerCentsDC Program." *2010 POWER Conference:* 1–46.

Wolak, Frank A. 2011. "Do residential customers respond to hourly prices? Evidence from a dynamic pricing experiment." *American Economic Review* 101 (3): 83–87.

Appendix A

Are Households Attentive to Time-varying Marginal Incentives? Evidence from a Field Experiment – Appendix

A.1 Proof of Proposition 1

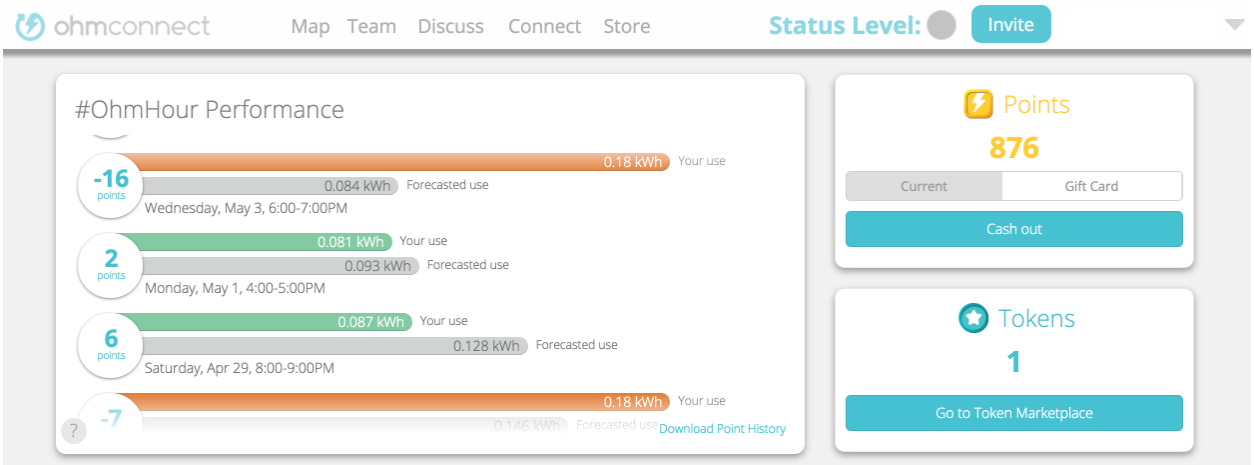
Proposition 3. *If* $\sigma < 1$ *and* $p_b \neq p_0$*, then* $\lim_{p_c \to p_0} \Delta Y(p_c, p_b) \neq 0$.

Proof: Suppose $\sigma < 1$ and $p_b \neq p_0$. Note that $\lim_{p_c \to p_0} \sigma \Delta y^*(p_c) = 0$, but $\Delta y^*(p_b) = y^*(p_b) - y^*(p_0) \neq 0$ for all p_c because $y^*(p) = p^\eta$ is strictly monotonic for $\eta < 0$. Thus, $\lim_{p_c \to p_0} \Delta Y(p_c, p_b) = \lim_{p_c \to p_0} (1 - \sigma) \Delta y^*(p_b) \neq 0$. □

A.2 Event Feedback

Figure A.2.1 shows the web-portal shown to households when they log into the account with the DRP. This page is the first thing the household sees when they log into the implementer's website. The left box reports the forecast, usage and reward for each event. The upper right box shows the household's point balance. The yellow boxes cover the company name and the company's proprietary term for events which I have redacted for the purpose of confidentiality.

Figure A.2.1: Web-Portal Homepage with Event Results



The figure shows a screenshot of the web-portal homepage for the household's account.

A.3 Supplementary Summary Statistics

Table A.3.2 reports summary statistics of pre-enrollment data for the study sample and the eligible population. Columns (1)-(6) report statistics calculated from household-level aggregates. Column (7) reports the population mean for households living in the area eligible for enrollment. Column (8) reports the difference between the population and sample means in (2) and (7). Column (9) reports p-values estimated the difference. I bootstrap standard errors on the difference by drawing 1,000 population-weighted random samples of 5,491 households from the eligible households and empirically constructing the sampling distribution. Except for percent households with low income, the sample is observably different from the eligible population. This could reflect the voluntary selection into signing up or the targeted marketing used by the DRP. In particular, the sample appears to be in areas with fewer standalone houses, more renters, and more electric heating. These differences generally reflect more urban areas. Data on the energy consumption of the eligible population will become available in November 2017.

Table A.3.1: Additional Sample Summary Statistics

	Households (1)	Mean (2)	Std. Dev (3)	Min (4)	Median (5)	Max (6)
Energy & Weather						
Daily Consumption (kWh)	5,491	16	11	0	14	144
Max Consumption (kWh)	5,491	5	3	0	5	44
Hourly Outdoor Temp. (°C)	5,491	17	3	1	17	34
Mean Daily CDHs	5,491	34	32	0	28	335
Mean Daily HDHs	5,491	114	42	0	119	449
Pre-Enrollment Obs.	5,491	8,710	4,372	336	8,735	20,084
Census Block Group Demographics						
% HH Income < $25K	4,664	19	14	0	16	89
% Population Age 21-39	4,664	30	12	0	28	100
% Family HHs	4,664	66	19	0	70	100
% Population w/ Bachelors	4,664	22	13	0	21	67
Median Year Built	4,603	1974	17	1939	1975	2011
% Renters	4,664	49	27	0	48	100
% HHs w/ 3+ bedrooms	4,664	52	30	0	53	100
% Detached Units	4,664	54	33	0	56	100
% Electric Heating	4,664	28	18	0	25	100
Median Monthly Rent ($)	4,366	1330	503	246	1,267	3,501

The table reports summary statistics for study participant using household-level data. Column (1) reports the number of households used to calculate each statistic. The energy, temperature, and pre-enrollment observations are calculated by averaging within a household using pre-enrollment energy and weather data. The demographic statistics from row (7) and below are calculated using a subset of the study participants for whom census block group designations could be made. The census variables are constructed from block-group-level data from the 2015 American Community Survey 5-year estimates.

Table A.3.2: Sample vs. Eligible Population

	HHs (1)	Mean (2)	Mean (μ_{pop}) (7)	Difference (8)	$\mu_{sample} = \mu_{pop}$ (9)
Energy & Weather					
Daily Consump. (kWh)	5,491	16	-	-	-
Max Consump. (kWh)	5,491	5	-	-	-
Hourly Temp. (°C)	5,491	17	-	-	-
Mean Daily CDHs	5,491	34	-	-	-
Mean Daily HDHs	5,491	114	-	-	-
Pre-Enrollment Obs.	5,491	8,710	-	-	-
Demographics					
% HH Inc. < $25K	4,664	19	19	0.07	0.71
% Pop. Age 21-39	4,664	30	26	3.40	0.00
% Family HHs	4,664	66	69	-2.85	0.00
% Pop. w/ Bachelors	4,664	22	21	1.00	0.00
Median Year Built	4,603	1974	1973	1.16	0.00
% Renters	4,664	49	43	5.75	0.00
% HHs w/ 3+ beds	4,664	52	56	-4.35	0.00
% Detached Units	4,664	54	60	-6.23	0.00
% Electric Heating	4,664	28	25	2.96	0.00
Median Rent ($)	4,366	1330	1360	-30	0.00

The table reports summary statistics for study participant using household-level data in columns (2)-(6). Column (1) reports the number of households used to calculate each statistic. The energy, temperature, and pre-enrollment observations are calculated by averaging within a household using pre-enrollment energy and weather data. The demographic statistics from row (7) and below are calculated using a subset of the study participants for whom census block group designations could be made. The census variables are constructed from block-group-level data from the 2015 American Community Survey 5-year estimates. Column (7) reports the population mean for eligible households using the same ACS data and column (8) reports the difference between the sample and population means. Column (9) reports empirical p-values testing the two-sided null that the mean for the sample equals the mean for the population. I estimate the sampling distribution of the mean by drawing 1,000 random samples from the population.

A.4 Attention Dynamics

The model developed in the paper assumes household prior beliefs over the distribution of possible price changes are exogenous and homogeneous. In this section, I relax these assumptions to make the model generalizable and more applicable to the sending. I proceed by extending the model to consider multiple price changes and show how information costs affects consumption behavior.

To gain intuition, consider the simplified example with choices between two information sets illustrated above with multiple sequential price changes over many time periods. Suppose that each time period, the two period framework ensues in isolation and that some periods may not have price changes. Formally, this can be achieved by assuming households are myopic to future changes or that they completely discount any future consumption. Let y_t denote consumption in time period t which has a regular price of p_t. In each period, $p_{d,t} \sim F_p$ if there is a price change and $p_{d,t} = p_t$ if there is no price change. Let m_t denote the signal about the price change with $m_t = 0$ if there is no change.[1] Let $\widehat{F}_{p,t}$ denote the household priors in period t with $\widehat{F}_{p,0}$ denoting initial priors, which may be biased.

For all $t > 0$, define $\widehat{F}_{p,t} \equiv b(\widehat{F}_{p,t-1}, S_{t-1})$ where b is a correspondence that determines current household priors based on past priors, signals and information sets chosen. I assume b captures the behavior of a Bayesian updater when the complete information set is chosen in every period.[2] That is, as $t \to \infty$, $\widehat{F}_{p,t} \to F_p$ if $S_t = S_{c,t}$ for all periods t where a price change occurred. Further, I assume using the bounded information set provides no additional information such that $\widehat{F}_{p,t} = b(\widehat{F}_{p,t}, S_{b,t})$. Thus, households *learn* from using the complete information and do not from using the bounded.

Lastly, assume $\kappa_t \equiv \kappa + \xi_t$ where $\kappa \sim F_\kappa$ is varies across the population of households but not over time. The term $\xi_t \sim \mathcal{N}(0, \sigma_\xi)$ is a stochastic shock to the information costs for period t. This structure allows for heterogeneity across households and over time that may drive attentiveness. In this simplified case, the shock is common across all households, but the model is easily extended to having the shocks vary across households within a period.

The model nests standard models of Bayesian updating if there are no information costs such that $\kappa = 0$ and $\xi_t = 0$ for all households in all periods. In each period, the household chooses the complete information set and variation in p_d traces out demand with $\varepsilon_t = \eta$ because $\sigma_t = 1$ for all t, where σ_t is the fraction of attentive households

[1] There is a slight abuse of notation here since m_t is technically an ordered set of information.

[2] Formally, this can be developed by defining b a contraction mapping over the parameter space, but this structure is unnecessary for the model.

in period t. In this scenario, regardless of any bias in priors, all household's priors will converge to the true distribution given enough time. If information is costly, households that would otherwise behave as Bayesian updaters behave substantively differently because priors can become rigid if information costs are sufficiently large.

I provide a simple example to guide intuition. Let $c_t \equiv \mathrm{E}[\tilde{V}(p_{d,t}, Z)|\widehat{F}_{p,t}] - \tilde{V}(\mathrm{E}[p_{d,t}|\widehat{F}_{p,t}], Z)$ denote the threshold cost where households are indifferent to the two sets of information. First, consider the case where information costs are homogeneous and do not vary over time ($\kappa_t = k, \forall t$). The households can be described by a single representative household in this case. For $k > c_0$, households choose the bounded information set initially and thus never update their priors. Demand shifts but is otherwise inelastic to the size of changes every time a price change occurs.

For $k \leq c_0$, households choose the complete information set and update their priors for the second price change. However, the equilibria are more complicated here. If the variance of the prior belief distribution is sufficiently small and information is sufficiently costly, a household will rationally choose the bounded information for all future price changes. There is no guarantee that households converge to the true distribution over time. Thus, $\sigma_t > 0$ for some period, but households may choose to be rationally inattentive in equilibrium such that $\sigma_t = 0$ beyond some threshold period where learning the true price is not worth the potential benefit.

Introducing heterogeneity across households further complicates the equilibrium description, but the primary takeaway is that even small information costs can drive inattention in equilibrium where households are Bayesian updaters. This framework could explain why even after significant exposure to dynamic pricing, households may be boundedly rational and use simplified decision rules.

A.5 Sample Attrition

Table A.5.3 reports the estimates of probability of being enrolled or in the sample conditional on treatment assignment. Each column reports coefficients from separate regressions estimated using OLS. Standard errors are estimated assuming *iid* errors. Column (1) reports estimates for probability of completing enrollment (defined as connecting a utility account) for all recruited households (defined as having entered an email into the website). The constant term captures the estimated probability for the control group and the coefficients on the indicators for Standard and Encouraged assignment capture estimates on the difference in probability of completing enrollment for those groups. The fourth row reports the p-value on the test that the two treatment group coefficients are the same. Column (2) reports estimates for the probability of being in the study sample (defined as not having on-site generation or

erroneous meter data) for all households that completed enrollment conditional on treatment assignment.

Table A.5.3: Estimates of Sample Attrition

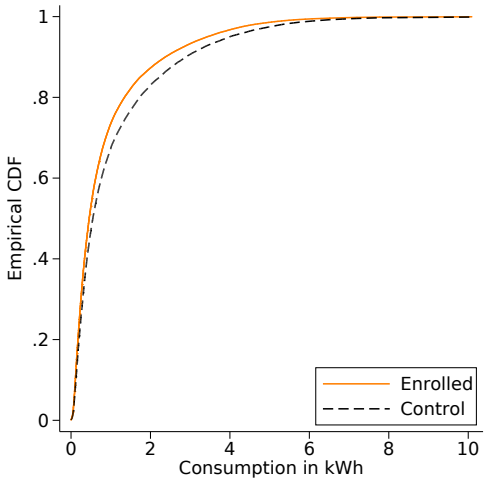
Dependent Variable:	Pr(Enrolled if Recruited)	Pr(In Sample if Enrolled)
	(1)	(2)
Constant (Assigned Control)	0.417***	0.899***
	(0.009)	(0.009)
1(Assigned Standard)	0.043***	-0.005
	(0.012)	(0.011)
1(Assigned Encouraged)	0.034***	0.002
	(0.012)	(0.011)
p-value for $H_0 : S = E$	0.39	0.44
Observations (Households)	13,782	6,169

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

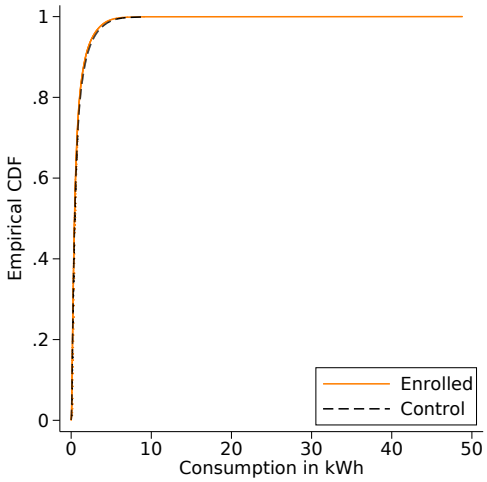
A.6 Raw Variation

This section reports on the raw experimental variation in two ways. First, I plot empirical CDFs for event periods and non-event periods to show how the distribution of consumption changed in Figure A.6.2. The figure plots the empirical CDFs for enrolled and control households in orange and black dashed lines, respectively, broken out by event and non-event periods and in kWh and log(kWh). Panels (a) and (c) plot distributions for event periods and (b) and (d) for non-event periods. Panels (a) and (b) plot distributions for raw kWh and (c) and (d) for non-event periods.

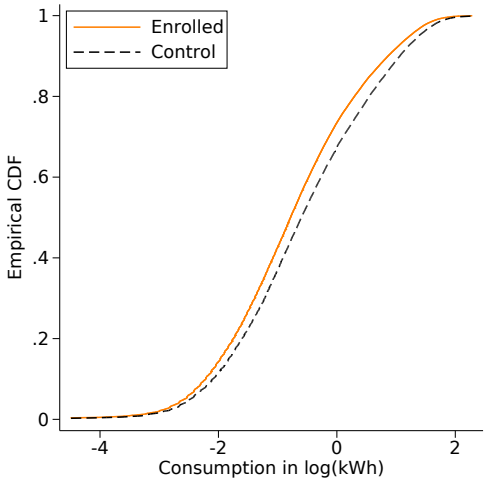
Figure A.6.2: Distribution of Post-Enrollment Consumption



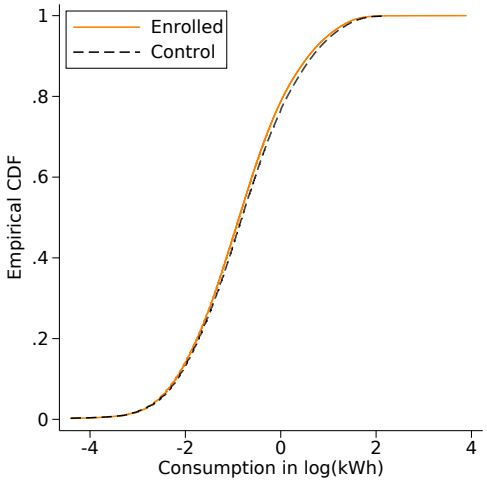
(a) Event Periods



(b) Non-Event Periods



(c) Event Periods



(d) Non-Event Periods

Second, I shows regressions for the ITT effect estimated in Table 3 that do not include fixed effects or controls. These provide information on how the estimates change as a result of adding controls and fixed effects. Table A.6.4 reports the estimates using only post-enrollment data. Columns (1)-(3) report estimates with kWh as the dependent variable and columns (4)-(6) report estimates for log(kWh). Columns (1) and (4) report estimates of the ITT using only event periods, columns (2) and (5) using only non-event periods, and columns (3) and (6) using the full post-enrollment sample and including temperature controls and hour-of-sample fixed effects. All regressions use OLS to estimate the treatment effects and standard errors reported in parentheses are clustered by household and hour-of-sample.

The effect sizes are larger than the DD specification, but the difference between event and non-event hours is quantitatively similar. The DD specification with fixed effects is preferred because it is more robust to the concern that households with systematically different consumption were assigned treatment and control groups by chance. This concern is legitimate because the design did not allow for stratification. Notably the precision of the estimates improves substantially when household fixed effects are included (for the paper estimates).

Table A.6.4: Estimates of Pricing Event Effect using Post Data Only

	kWh	kWh	kWh	log(kWh)	log(kWh)	log(kWh)
	OLS	OLS	OLS	OLS	OLS	OLS
	(1)	(2)	(3)	(4)	(5)	(6)
$D_{it} \times H_t$	-0.185^{***}		-0.165^{***}	-0.185^{***}		-0.165^{***}
	(0.036)		(0.029)	(0.032)		(0.028)
$D_{it} \times N_t$		-0.059^{**}	-0.045^{***}		-0.065^{**}	-0.049^{**}
		(0.022)	(0.018)		(0.027)	(0.023)
Temp. Controls	No	No	Yes	Yes	Yes	Yes
Hour-of-Sample FEs	No	No	Yes	Yes	Yes	Yes
HHs	5,491	5,491	5,491	5,491	5,491	5,491
Obs.	155,403	11,540,163	11,695,566	155,403	11,540,163	11,695,566

Standard errors clustered by household and hour-of-sample in parentheses

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

A.7 Results with kWh Consumed as Dependent Variable

This section provides results for the main empirical exercises of the paper using kWh as the dependent variable instead of log(kWh). Table A.7.5 reports the regression results for the ITT and LATE estimates analogous to Table 3 in the paper. Table A.7.6 reports the regression results for the moral suasion treatments that estimate the effect of moral suasion and price analogous to Table 4 in the paper. Table A.7.7 reports the regression results for the OLS, reduced form, and IV estimates of the LATE of automation analogous to Table 6 in the paper. Figure A.7.3 shows the estimates for the pricing hour event study regressions analogous to Figure 4 in the paper. Figure A.7.4 shows the estimates for the effect of marginal price analogous to Figure 5 in the paper. Figure A.7.5 shows the estimates for the persistence of the price response analogous to Figure 6 in the paper.

Table A.7.5: Effect of Pricing Events on kWh Consumed

	OLS (1)	OLS (2)	IV (3)
Panel A: Effect Estimates			
Enrolled × Event ($D_{it} \times H_t$)	-0.107^{***}		
	(0.012)		
Called × Event ($C_{it} \times H_t$)		-0.120^{***}	-0.120^{***}
		(0.013)	(0.014)
Enrolled × Non-event ($D_{it} \times N_t$)	-0.018^{***}	-0.018^{***}	-0.018^{***}
	(0.005)	(0.005)	(0.005)
Panel B: 1st Stage Estimates			
Enrolled × Event ($D_{it} \times H_t$)			0.889^{***}
			(0.005)
Households	5,491	5,491	5,491
Observations	22,926,631	22,926,631	22,926,631

Standard errors clustered by household and hour-of-sample in parentheses

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

Panel A reports estimates of the effect of enrollment during pricing event periods and non-event periods on hourly energy consumption measured in kWh. Each column report coefficients estimated from separate regressions. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. Column (1) and (2) reports the raw experimental variation using the sample of event periods and post-enrollment non-event periods, respectively. Column (3) combines event and non-event post-enrollment periods and estimates the model with hour-of-sample fixed effects and temperature controls. Column (4) and (6) report the preferred estimates. Column (4) reports estimates for the DD specification that includes pre-enrollment data and household by hour-of-day fixed effects. Column (5) reports estimates for the LATE for called households using OLS. Column (6) reports the 2SLS estimates where called being called is instrumented with enrollment. The first stage coefficient on the instrument reported in Panel B.

Table A.7.6: Effect of Price vs. Moral Suasion (No Price) on kWh Consumed

	(1)	(2)	
Panel A: Effect Estimates			
Moral Suasion Only (δ_M)	-0.039^{***}		
	(0.016)		
× Non-Automated		-0.006	
		(0.16)	
× Automated		-0.305^{***}	
		(0.078)	
Price Only (δ_P)	-0.079^{***}		
	(0.019)		
× Non-Automated		-0.052^{***}	
		(0.019)	
× Automated		-0.294^{***}	
		(0.066)	
Moral Suasion + Price (δ_B)	-0.060^{***}		
	(0.021)		
× Non-Automated		-0.021	
		(0.021)	
× Automated		-0.365^{***}	
		(0.078)	
Panel B: Coefficient Tests	*Pooled*	*Non-automated*	*Automated*
H_0: $\delta_M = \delta_P$	0.005	0.002	0.865
H_0: $\delta_M = \delta_B$	0.171	0.328	0.247
H_0: $\delta_P = \delta_B$	0.192	0.056	0.144
H_0: $\delta_M + \delta_P = \delta_B$	0.004	0.091	0.003
Households	1,255	1,255	
Observations	2,579,728	2,579,728	

Standard errors clustered by household and hour-of-sample in parentheses

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

Panel A reports estimates of the casual effect of message content from the *non-price treatments* on hourly energy consumption measured in kWh. Each column report coefficients estimated from separate regressions. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. Column (1) reports the model estimated using the DD specification that includes pre-enrollment data, household by hour-of-day fixed effects, temperature controls, and hour-of-sample fixed effects. Column (2) reports estimates for the same model interacting each message type indicator with automation status. Estimates comparing within automation type can be interpreted causally. Panel B reports p-values on tests of whether treatments are equivalent and additive. δ_M representing the parameter for moral suasion only, δ_P representing the parameter for price only, and δ_B representing the parameter for both interventions together. Column (1) reports the p-values for the coefficients that pool automated and non-automated. Column (2) breaks out the p-values by automated and non-automated.

Table A.7.7: Effect of Automation on kWh Consumed

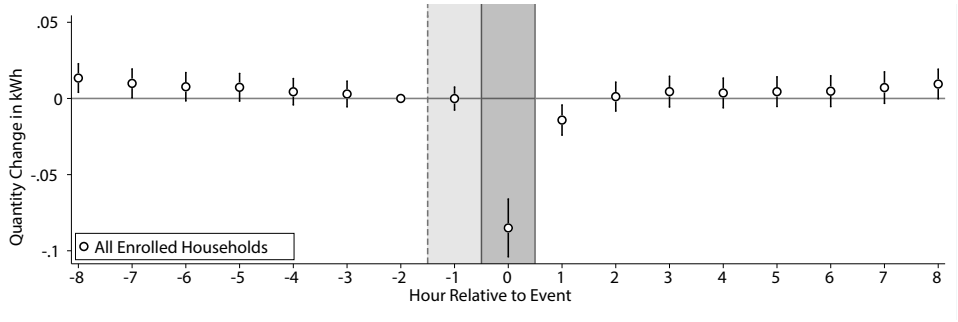
	OLS	OLS	IV
	(1)	(2)	(3)
Panel A: Effect Estimates			
Adopted × Event ($A_{it} \times H_t$)	-0.190^{***}		-0.567^{***}
	(0.028)		(0.165)
Adopted × Non-Event ($A_{it} \times N_t$)	0.021		-0.139
	(0.013)		(0.106)
Encouraged × Event ($Z_{it} \times H_t$)		-0.040^{***}	
		(0.011)	
Encouraged × Non-Event ($Z_{it} \times N_t$)		-0.009	
		(0.007)	
Panel B: Selected First Stage Estimates			
Encouraged × Event ($Z_{it} \times H_t$)			0.075^{***}
			(0.009)
Encouraged × Non-Event ($Z_{it} \times N_t$)			0.074^{***}
			(0.007)
Households	4,448	4,448	4,448
Observations	18,559,930	18,559,930	18,559,930

Standard errors clustered by household and hour-of-sample in parentheses

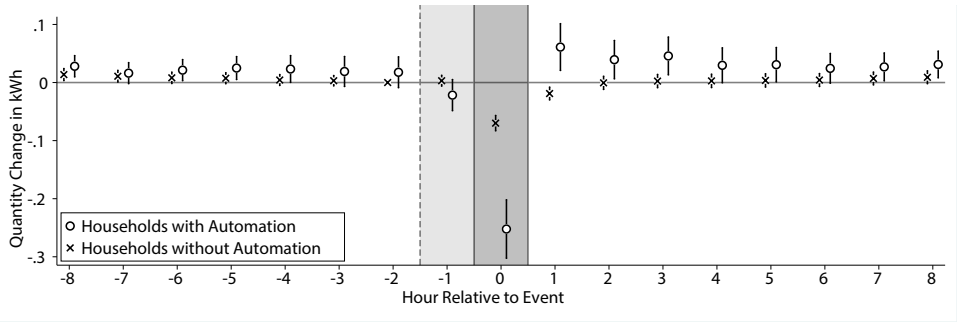
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Panel A reports estimates of the effect of automation on energy consumption in kWh for the subsample of enrolled households (excluding controls). Columns report coefficients estimated from separate regressions. Standard errors are reported in parentheses and estimated using a covariance matrix two-way clustered by household and hour-of-sample. All regressions include household by hour-of-day and hour-of-sample fixed effects and temperature controls. Column (1) reports the OLS estimate for the effect of automation on consumption. Column (2) reports the reduced form specification regressing consumption on the encouragement indicator. Columns (3) reports the 2SLS estimates instrumenting the automation indicator with the encouragement. Panel B reports results of the first-stage regressions for the coefficient on the instrument corresponding to the endogenous variable.

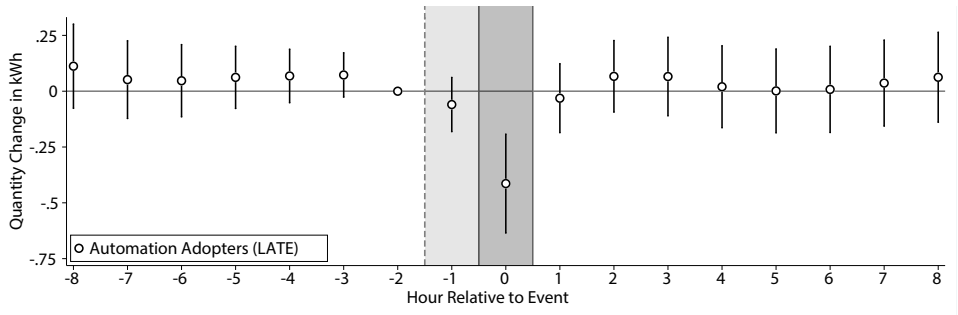
Figure A.7.3: Pricing Hour Event Studies in kWh Consumed



(a) Enrolled vs. Control Households



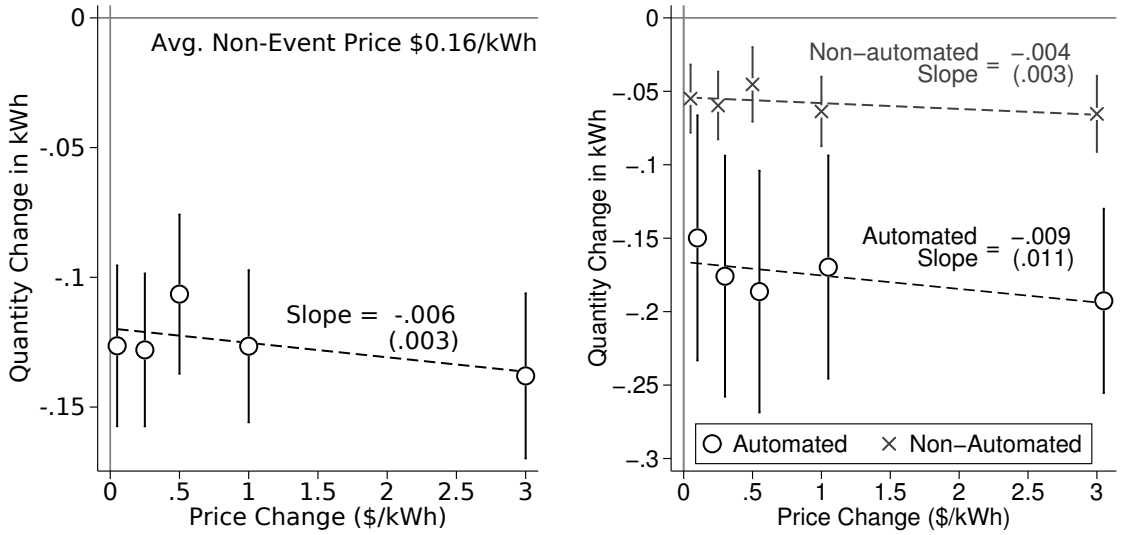
(b) Automated vs. Non-Automated Households



(c) Additional Effect of Automation for Enrolled Households

Panel (a) plots event-study point estimates of the ITT effect that compares the consumption of enrolled households to control households in the hours leading up to and following a pricing event (shaded dark gray). Panel (b) plots the non-causal decomposition of the ITT estimates in (a) by whether households automated their decisions. I add a gray horizontal line indicating the level of consumption for households with automation in period -2 to assist in visual interpretation. Panel (c) plots event study estimates of the LATE of adopting automation on consumption during pricing periods. Households are contacted during the hour prior to the event (shaded light gray). I use the period two hours prior as the reference category so the vertical axis is the quantity change relative to period -2 in kWh. For (b) the reference category is period -2 for non-automated households. Vertical bars indicate 95 percent confidence intervals clustered by household and hour-of-sample.

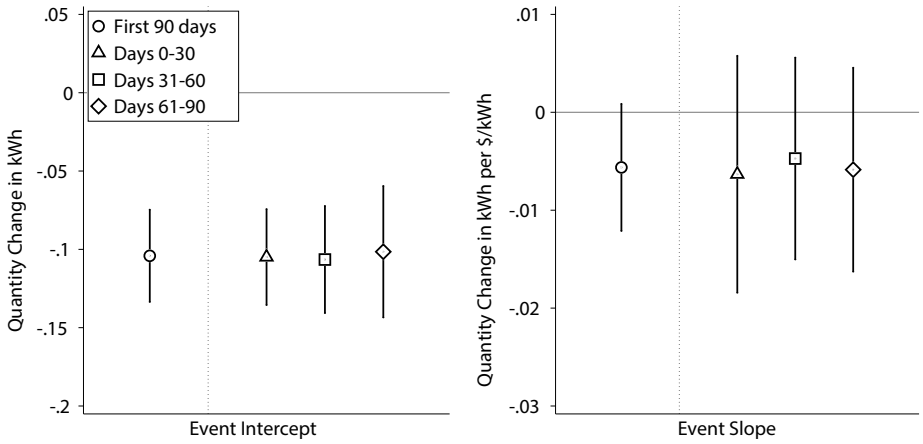
Figure A.7.4: Effect of Marginal Price in kWh Consumed



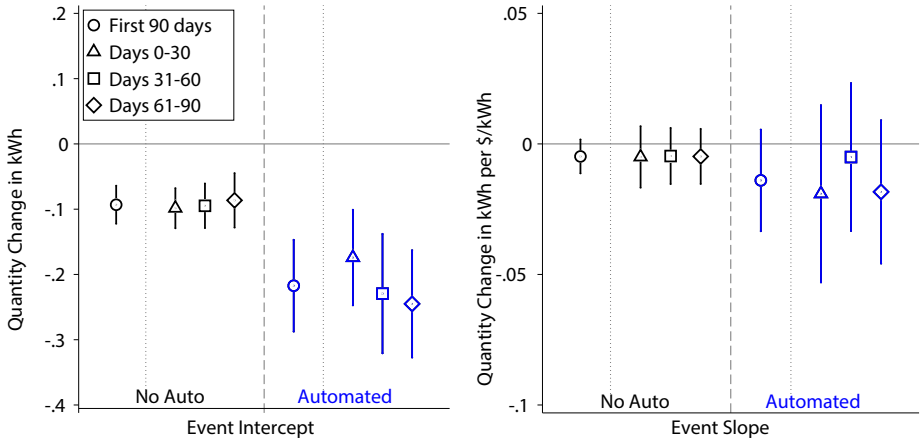
(a) Effect of Price Change in kWh (b) Automated vs. Non-Automated in kWh

The figure plots the point estimates for the change in the conditional mean of consumption as a result of changes in the effective price of electricity in dollars per kWh. Note the difference in the scales of the axes between the left and right panels. The left panel shows the effect with consumption in kWh on the vertical axis and the price change on the horizontal. The right panel shows the treatment effect heterogeneity by households with and without automation. Both panels show confidence interval around estimates using two-way clustered standard errors. The dashed line plots the linear parametric estimation of the response as a function of the price change. The figure also reports the estimates for the slope and intercept of the parametric estimation with two-way clustered standard errors in parentheses.

Figure A.7.5: Persistence of Price Response in kWh Consumed



(a) Event Price Response Effect



(b) Automated vs. Non-automated Households (non-causal)

This figure shows how *price treatment* effects change as a function of time in the program. The estimates are broken out by 30 day periods during the 90 days for which *price treatments* occurred. Triangles represent days 0-30, squares days 31-60, and diamonds days 61-90. Circles represent the estimate for the full 90 days as reference. Vertical bars show 95 percent confidence intervals constructed from standard errors two-way clustered by household and hour-of-sample. Each panel represents estimates from a separate regression. Panel (a) plots causal estimates of the event price response and non-event enrollment response for all enrolled households regardless of automation choice. Panel (b) plots the same parameters non-causally decomposed to show differences between non-automated (black) and automated (blue) households. Estimates labeled "Event Intercept" and "Event Slope" represent the parametric estimation of the price response of consumption for called households.

A.8 Examples of Household Consumption Distributions

This section provides evidence demonstrating that the effect is not likely to be driven by discrete changes in the demand function. To show this, I provide information on the within-household distribution of electricity consumption in kWh during one hour of the day of 4pm-5pm (the hour when the most pricing events occurred) during the three months prior to enrollment. I choose four households at random from the sample.

A.9 Point Balances Over Time

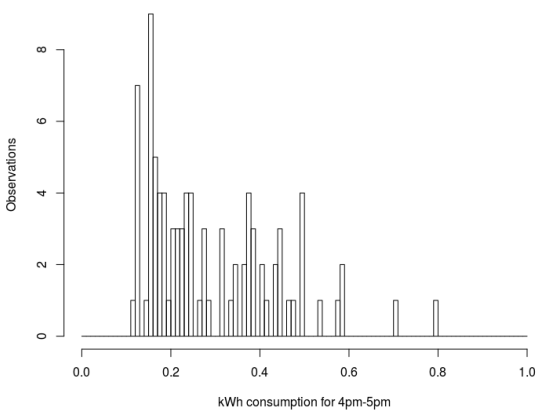
This section provides information on the distribution of point balances over time. Figure A.9.7 reports the distribution of point balances for households as a function of their time enrolled in the program. The blue line reports the average, which generally rises during the first 18 weeks enrolled in the program. The solid black line with points reports the median, which does not change generally during the program, suggesting the change in average is driven largely by households on the upper tail of the distribution. The dashed (dotted) lines report the 40th/60th (25th/75th) percentiles of the distribution of balances by week of enrollment. The figure also clearly shows a large fraction of households have a zero balance for the majority of the program, which averages around 30 percent.

Figure A.9.7 averages over all enrollment periods and masks heterogeneity in the initial enrollment balance. Figure A.9.8 plots the average balance after two weeks enrolled to demonstrate that the initial enrollment bonus changed in late May from 500 points ($5) to 0 points ($0). The figure captures time series variation in the initial balance which is largely a function of the initial sign-up bonus. Since enrollment sometimes took longer than a week I average the point balance for households after two weeks. The variation around these two values is due to other bonus incentives offered for referrals and from event periods that occurred during the first two weeks.

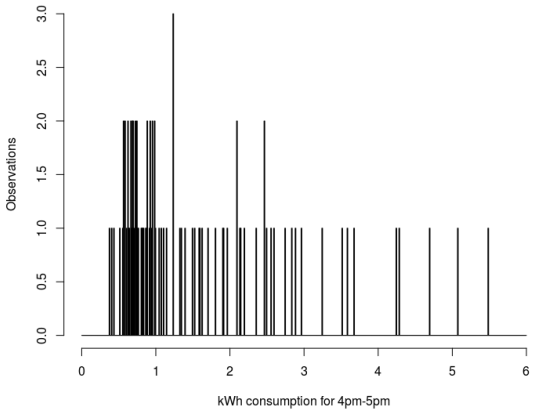
A.10 Variation by Point Balance

This section reports estimates of the price response heterogeneity by the households point balance at the time the event was called. Table A.10.8 reports the results for DD regressions with log consumption as the dependent variable, interacting the terms with six bin indicators for point balance at the time of the event. Column

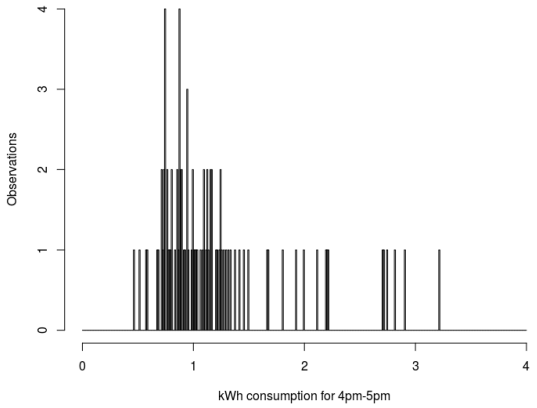
Figure A.8.6: Example Distributions of Individual Household Consumption



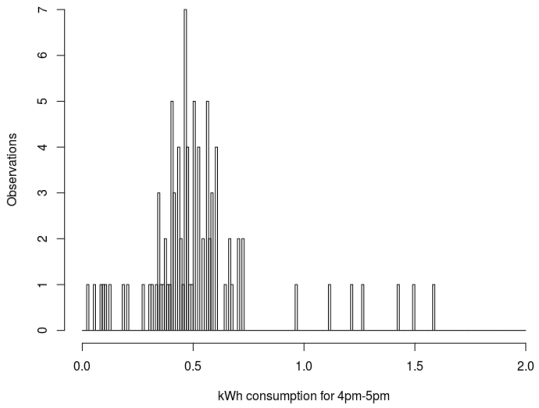
(a) Example Household #1



(b) Example Household #2

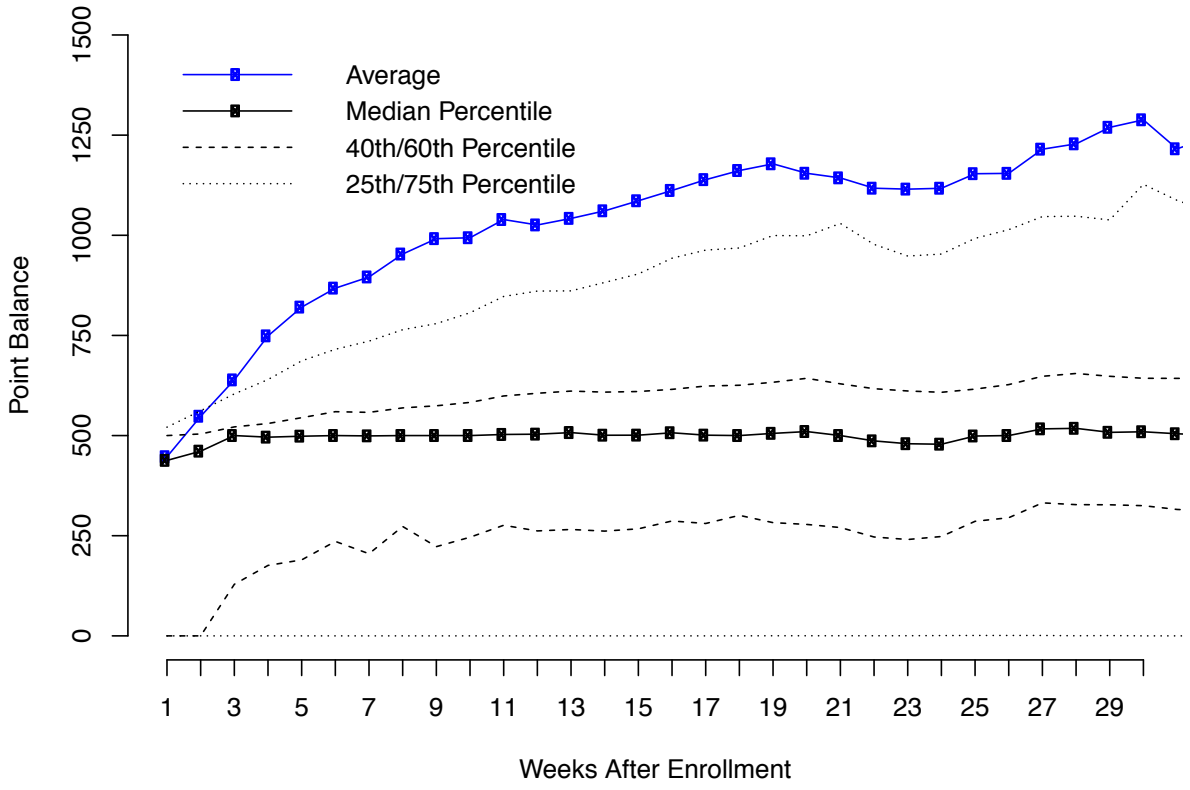


(c) Example Household #3



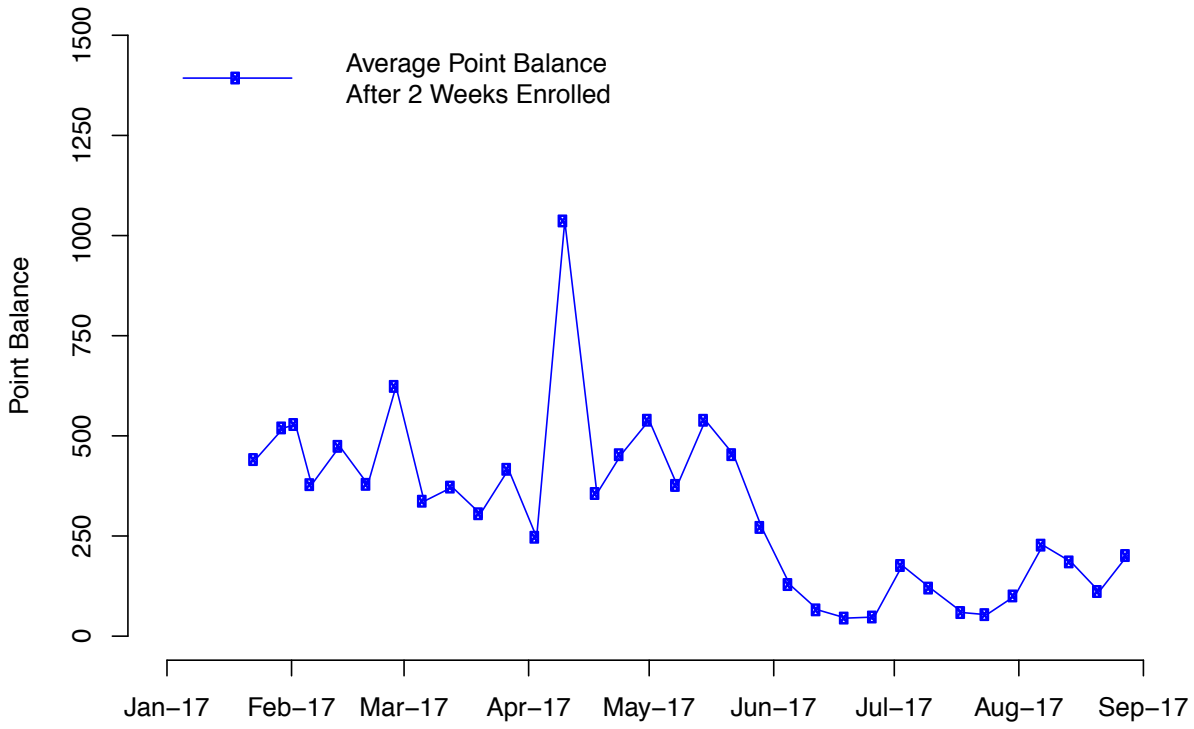
(d) Example Household #4

Figure A.9.7: Point Balances by Week After Enrollment



(1) reports the estimates from the full sample for reference. Columns (2)-(7) report coefficients from a single regression where each point balance dummy is interacted with a slope and intercept variable for price response. The first row reports estimates for the intercept, the second row reports the slope which is in log points reduced as a result of a \$1/kWh price increase. Two-way clustered standard errors are reported in parentheses. Importantly, none of the estimates can be interpreted causally because there is selection that determines a household's point balance and still the pattern is revealing. Individuals with a zero balance consume similar to the control group and have a near zero price sensitivity. The households with greater point balances have steeper price response slopes. While column (6) shows the steepest it is also worth

Figure A.9.8: Initial Point Balance Variation



noting the number of event observations in each point balance bin, shown below the estimates. The treatment effect is a weighted average of this heterogeneity.

Table A.10.8: Price Response by Point Balance at Time of Event

Regression:	Sample	Price Response × Point Balance					
Balance Range:	-	Zero	1 -499	500 -999	1,000 -1,499	1,500 -2,000	> 2,000
Cashout Eligible?		No	No	No	Yes	Yes	Yes
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	-0.12^{***}	-0.03^{***}	-0.15^{***}	-0.19^{***}	-0.27^{***}	-0.29^{***}	-0.29^{***}
	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.04)	(0.03)
Slope	-0.008^{***}	-0.001	0.01	-0.03^{***}	-0.05^{***}	-0.08^{***}	-0.04^{***}
	(0.003)	(0.003)	(0.01)	(0.01)	(0.01)	(0.03)	(0.01)
Event Obs.	155,403	61,835	35,065	18,933	3,194	1,024	5,656
Balance Stats							
Mean	333	0	145	637	1,171	1,730	4,429
(SD)	(1,187)	(0)	(150)	(132)	(134)	(151)	(4,299)
HHs	5,491	5,491					
Obs.	22,929,631	22,929,631					

SEs clustered by household and hour-of-sample in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.11 Results for Candidate Heuristics

This section uses an encompassing test to see if the empirical results are driven by exogenous variation in previous price changes and heuristics constructed from them. I construct three candidate heuristics. Differences between incentive levels in these heuristics can be interpreted causally because the incentive level was randomly assigned each event between users.

1) **Immediate Prior** – This heuristic is defined for each individual as the incentive they received on the previous event. For instance, households with limited memory may make decisions based on the most recent experience because it is more readily accessible.

2) **Limited Bayesian prior of past three incentives** – This heuristic is defined for each individual as the average of past three incentive levels received. If households are making decisions based on what they expect the incentive level to be and Bayesian updating, this would be reflected in the variation captured by this heuristic.

3) **Full Bayesian prior of all previous events** – This heuristic is defined similar to the limited Bayesian except that it uses all past incentive levels received.[3]

Table A.11.9 reports results for the parametric estimation of the price response for each heuristic. Column (1) reports the parametric estimates based on the actual incentive as reported in the paper, Column (2) for the immediate prior, Column (3) for the limited Bayesian prior, and Column (4) for the full Bayesian prior. Panel A reports parametric estimates and Panel B reports non-parametric estimates. For the non-parametric estimates, the heuristic is divided into quintiles. Note that estimates for different heuristics have different support because the quintiles become more narrow as the priors average over more events. Thus, Panel C reports the quintile ranges for each of the incentive levels with Column (1) showing the quintiles for the actual incentives.

[3]I assume the expected incentive level as the prior for the first event, but dropping these events from the estimation does not substantively affect the results.

Table A.11.9: Effect of Heuristics on log(kWh)

Heuristic:	Actual Incentive (1)	Immediate Prior (2)	Avg. Last 3 events (3)	Bayesian Prior (4)
Panel A: Parametric Estimates				
Intercept	-0.124^{***}	-0.1206^{***}	-0.118^{***}	-0.117^{***}
	(0.010)	(0.009)	(0.009)	(0.011)
Incentive Slope	-0.0072^{***}	-0.0072^{***}	-0.0072^{***}	-0.0072^{***}
	(0.002)	(0.0022)	(0.0022)	(0.0021)
Heuristic Slope		-0.0033	-0.0063	-0.0067
		(0.0022)	(0.0047)	(0.0086)
Panel B: Non-parametric Estimates ($C_{it} \times H_t \times$)				
$1(\Delta MP_{it} = 0.05)$	-0.123^{***}	-0.112^{***}	-0.105^{***}	-0.116^{***}
	(0.010)	(0.011)	(0.013)	(0.014)
$1(\Delta MP_{it} = 0.25)$	-0.131^{***}	-0.120^{***}	-0.113^{***}	-0.124^{***}
	(0.010)	(0.011)	(0.014)	(0.015)
$1(\Delta MP_{it} = 0.50)$	-0.116^{***}	-0.105^{***}	-0.098^{***}	-0.109^{***}
	(0.010)	(0.011)	(0.011)	(0.015)
$1(\Delta MP_{it} = 1.00)$	-0.139^{***}	-0.129^{***}	-0.121^{***}	-0.133^{***}
	(0.010)	(0.011)	(0.014)	(0.015)
$1(\Delta MP_{it} = 3.00)$	-0.144^{***}	-0.133^{***}	-0.126^{***}	-0.137^{***}
	(0.011)	(0.011)	(0.014)	(0.015)
$1(\Delta PRIOR_{it} \in Q^{1st})$		-0.007	-0.012	-0.012
		(0.013)	(0.012)	(0.012)
$1(\Delta PRIOR_{it} \in Q^{2nd})$		-0.010	-0.013	0.002
		(0.013)	(0.015)	(0.013)
$1(\Delta PRIOR_{it} \in Q^{3rd})$		-0.010	-0.024	-0.004
		(0.013)	(0.016)	(0.013)
$1(\Delta PRIOR_{it} \in Q^{4th})$		-0.018	-0.027^{*}	-0.006
		(0.013)	(0.015)	(0.013)
$1(\Delta PRIOR_{it} \in Q^{5th})$		-0.016	-0.025	-0.019
		(0.013)	(0.015)	(0.013)
Panel C: Incentive Distribution Statistics				
Min	5	5	5	5
q = 0.20	25	25	37	67
q = 0.40	50	50	58	86
q = 0.60	100	100	117	101
q = 0.80	300	300	142	120
Max	300	300	300	300
[0.1em] Households	5,491	5,491	5,491	5,491
Observations	22,574,289	22,574,289	22,574,289	22,574,289

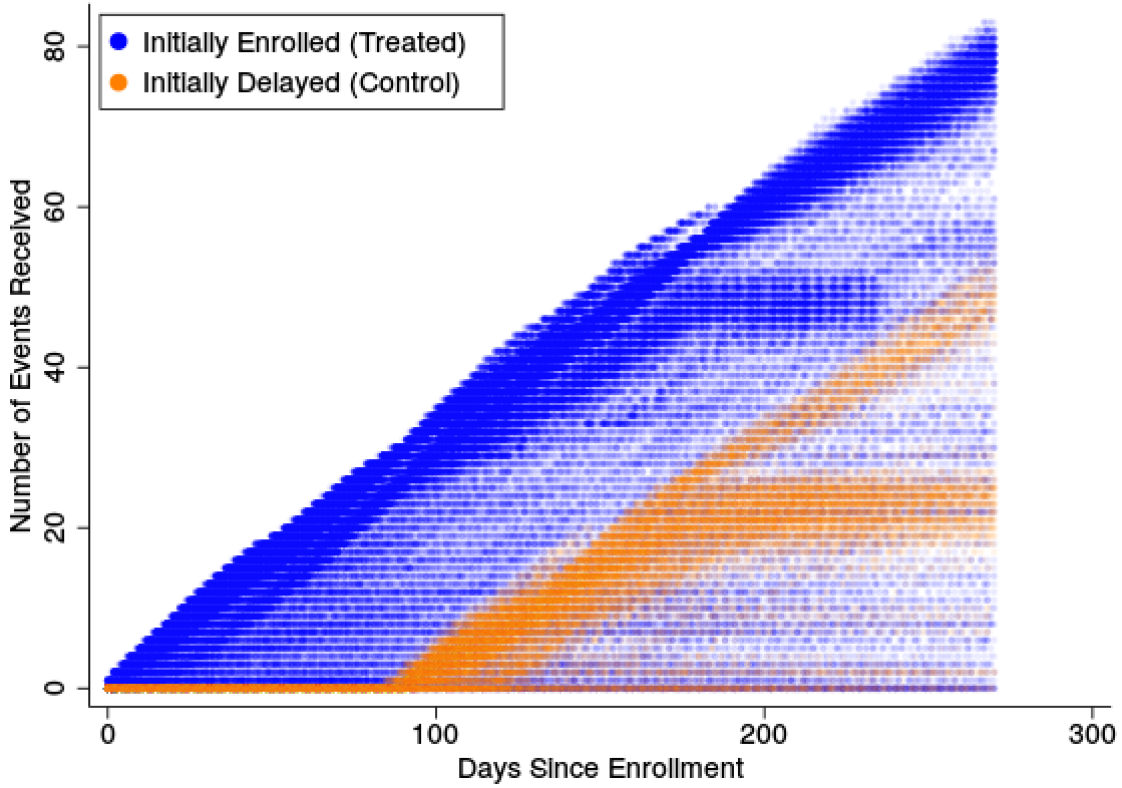
Standard errors clustered by household and hour-of-sample in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

A.12 Event Calling

This section provides information on how the pricing events were called during the program. Since the research team did not have control over the dispatch of events, I provide figures showing the distribution of events called as a function of days enrolled in the program. This information is meant to assist in interpreting the treatment effect since it averages across households regardless of the number of events they received. It also provides insight into the degree to which certain households were called while others were not.

Figure A.12.9 shows the distribution of number of events called per household as a function of their tenure in the program. Each data point corresponds to an event call for a household. The horizontal axis is the days since enrollment and the vertical axis is the number of events they have received to that point. Blue points plot households that were initially enrolled and orange points plot the initially delayed households. The points are transparent so that darker colors can be interpreted as a denser distribution of households. The figure shows how control households did not receive any events for the first 90 days due to the product delay. After that point, households were sent messages at a rate consistent with the initially enrolled, The dense cluster of points sloping up correspond to the frequency of events being called about every 3 days. However, comparing the density within a given vertical segment shows that some households with the same tenure in the program received less than the every three days. The density grows more sparse as tenure increases due to the rolling enrollment and the fixed end of the study. More households were enrolled for the first 90 days than for 250.

Figure A.12.9: Distribution of Event Calls Over Time



A.13 Survey

A.13.1 Survey Recruitment

Households were recruited to the survey via email within two weeks of the 180th day they enrolled (90 days after the price treatments had concluded). The email read:

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

A.13.2 Summary Stats and Comparison to Non-Responders

This section reports tests for observable differences between survey responders and non-responders. I compare the responders to the full sample, which includes individuals who may have not had the opportunity to respond yet. This means the differences may suffer from some classical measurement error. The table will be revised again at the end of the study. Table A.13.10 reports the results. Energy consumption in the two groups appears to be largely similar, but households who responded tended to be in cooler places with more heading degree days. They also tended to have shorter panel lengths which may proxy for more movers. For the census variables, respondents tended to come from areas with a slightly lower population 21-39 and areas with more 3+ bedroom houses.

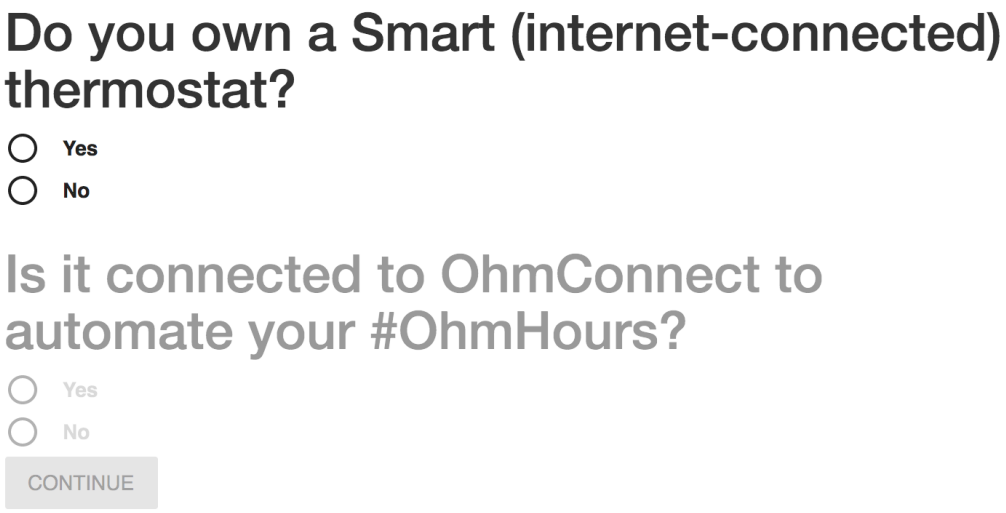
Table A.13.10: Difference in Means between Survey Responders and Non-Responders

	No Response	Survey Respondent	p-value
Daily Consumption (kWh)	16.8	15.8	0.04
	(12.6)	(11.2)	
Max Consumption (kWh)	4.1	4.0	0.25
	(2.4)	(2.2)	
Mean Daily CDHs	42.7	21.5	0.00
	(50.5)	(36.2)	
Mean Daily HDHs	107.2	154.2	0.00
	(68.6)	(75.1)	
Pre-Enrollment Obs.	2045.9	2041.3	0.75
	(356.9)	(353.0)	
% HH Income < $25K	19.6	19.2	0.57
	(13.9)	(13.4)	
% Population Age 21-39	30.0	28.9	0.06
	(12.1)	(11.3)	
% Family HHs	66.3	67.4	0.18
	(18.7)	(18.9)	
% Population w/ Bachelors	21.8	21.8	0.88
	(12.5)	(12.5)	
Median Year Built	1974.2	1975.1	0.21
	(17.0)	(16.5)	
% Renters	49.3	46.8	0.04
	(26.6)	(26.2)	
% Detached Units	53.3	55.7	0.12
	(33.0)	(33.3)	
% Electric Heating	28.1	27.2	0.23
	(17.5)	(17.6)	
Households/Observations	4,819	672	

A.13.3 Survey Questions

This section contains the wording and content of the survey. Sets of questions were presented as separate webpages designed to look consistent with the company's website. Figure A.13.10 shows the first question. The remaining questions are written out and the format of the webpage is described in the following manner. A box indicates a clickable box for the user to answer the question. [IF YES] indicates a conditional question that should be presented to the user if the question immediately above it in the list is answered "Yes". If the answer to a question is a number or integer indicated by "(integer)" a box should be presented to the user to enter a number. Completion Conditions follow each question and must be met before the "Continue" button is activated for each page.

Figure A.13.10: Screenshot of First Page of Survey



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

Completion Condition: No or Yes + Yes/No

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

Completion Condition: No or Yes + Yes/No

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

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

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

Completion Condition: No or Yes + Yes/No

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

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

Completion Condition: At least one box checked

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

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

Completion Condition: At least one box checked

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

Completion Condition: One box checked

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

Completion Condition: One box checked

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

☐ $1.00
☐ $10.00

Completion Condition: One box checked

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

Completion Condition: One box checked

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

Completion Condition: One box checked

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

Completion Condition: An integer entered into the field

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

Completion Condition: An integer entered into the field

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

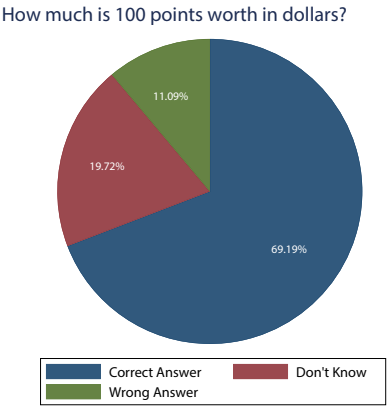
Completion Condition: One box checked

A.13.4 Selected Survey Results

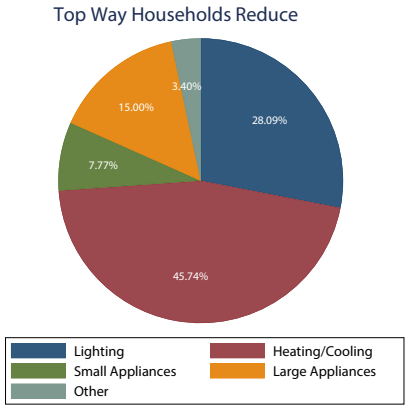
This section reports results for selected survey questions. Figure A.13.11 reports results for questions from the survey on (a) what their top reason for joining was, (b) whether they understand the monetary value of points, (c) the top ranked way they reduced consumption, and (d) what they plug into their smart plugs if owned.

Figure A.13.11: Results for Survey Questions

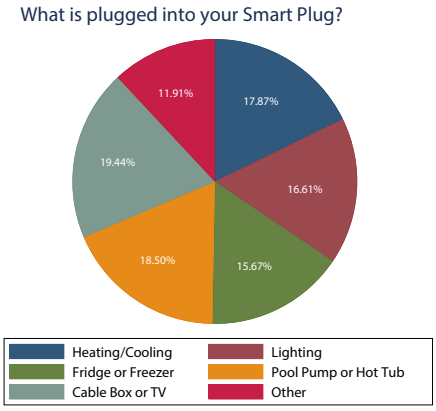
(a) Reason for Joining



(b) Understanding of Point Value



(c) Top Ranked Way of Reducing Consumption



(d) What is plugged into you smart plugs?

Appendix B

The Effect of Temperature on Student Test Performance: Evidence from High School Exit Exams – Appendix

B.1 Main Results with Alternative Control Specifications

This section reports additional regression results for characterizing temperature responses using alternative specifications with different fixed effects. Table B.1.1 and B.1.2 report results for alternative specifications. Columns (3) and (4) report the preferred specification reported in the paper. Columns (5) and (6) report results that add district by year fixed effects to soak up district trends over time.

Table B.1.1: Temperature Effect on Scores with Varying Controls

	Math	ELA	Math	ELA	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature Bin						
$< 12.5°$	0.41	0.31	-0.79*	-0.22	-0.04	-1.12
	(0.70)	(0.48)	(0.46)	(0.46)	(1.09)	(0.83)
12.5°-15°	-0.02	-0.17	-0.08	0.08	1.01*	-0.16
	(0.39)	(0.34)	(0.32)	(0.27)	(0.58)	(0.49)
15°-17.5°	0.38	-0.20	0.38	0.11	0.70*	0.14
	(0.35)	(0.26)	(0.25)	(0.17)	(0.39)	(0.24)
17.5°-20°	-	-	-	-	-	-
20°-20.5°	-0.12	-0.17	-0.46**	-0.31*	1.25***	-0.10
	(0.28)	(0.20)	(0.23)	(0.18)	(0.39)	(0.34)
22°-25°	-0.31	0.11	-0.77***	-0.31	0.39	-0.21
	(0.36)	(0.26)	(0.29)	(0.22)	(0.48)	(0.41)
25°-27.5°	-0.28	-0.12	-1.18***	-0.49*	-0.59	-0.23
	(0.43)	(0.31)	(0.32)	(0.26)	(0.60)	(0.48)
$> 27.5°$	-0.79*	0.04	-1.87***	-0.49	-1.31**	-0.01
	(0.46)	(0.39)	(0.37)	(0.30)	(0.64)	(0.63)
School & Date FEs	✓	✓	✓	✓	✓	✓
Demo. Controls			✓	✓	✓	✓
District × Year FEs					✓	✓
Observations	13,466	13,466	11,493	11,493	11,493	11,493

Standard errors in parentheses

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

The table reports estimates from regressing mean scale score on various fixed effects and controls. Standard errors reported in parentheses are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years.

Table B.1.2: Temperature Effect on Pass Rates with Varying Controls

	Math (1)	ELA (2)	Math (3)	ELA (4)	Math (5)	ELA (6)
Temperature Bin						
$< 12.5°$	0.005	0.004	-0.006	-0.000	0.008	-0.002
	(0.007)	(0.005)	(0.004)	(0.004)	(0.011)	(0.007)
12.5°-15°	-0.000	-0.005	-0.001	-0.001	0.014*	0.001
	(0.004)	(0.003)	(0.003)	(0.003)	(0.008)	(0.004)
15°-17.5°	0.001	-0.005**	0.002	-0.001	0.006	0.002
	(0.004)	(0.003)	(0.003)	(0.002)	(0.004)	(0.002)
17.5°-20°	-	-	-	-	-	-
20°-20.5°	0.000	-0.002	-0.003	-0.004***	0.013***	-0.000
	(0.003)	(0.002)	(0.002)	(0.002)	(0.004)	(0.003)
22°-25°	0.003	0.002	-0.004	-0.005**	0.005	-0.003
	(0.004)	(0.003)	(0.003)	(0.002)	(0.005)	(0.004)
25°-27.5°	0.003	-0.001	-0.007**	-0.006**	-0.003	-0.004
	(0.005)	(0.003)	(0.003)	(0.003)	(0.006)	(0.004)
$> 27.5°$	-0.002	0.001	-0.014***	-0.008**	-0.015***	-0.005
	(0.005)	(0.004)	(0.004)	(0.003)	(0.006)	(0.006)
Observations	13,466	13,466	11,493	11,493	11,493	11,493
School & Date FEs	✓	✓	✓	✓	✓	✓
Demo. Controls			✓	✓	✓	✓
District × Year FEs					✓	✓
Observations	13,466	13,466	11,493	11,493	11,493	11,493

Standard errors in parentheses

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

The table reports estimates from regressing pass rates on various fixed effects and controls. Standard errors reported in parentheses are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years.

B.2 Main Results for Balanced Panel

This section reports additional regression results for characterizing temperature responses using alternative specifications with different fixed effects for the subset of schools that have data for all 12 years. Table B.2.3 and B.2.4 report results for alternative specifications. Columns (3) and (4) report the preferred specification reported in the paper. Columns (5) and (6) report results that add district by year fixed effects to soak up district trends over time. Figures B.2.1a and B.2.1b show the main results for the balanced panel. Figures B.2.2a and B.2.2b show the event study figures for the balanced panel.

Table B.2.3: Temperature Effect on Scores (Balanced Panel)

	Math	ELA	Math	ELA	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature Bin						
$< 12.5°$	0.40	0.32	-1.10**	-0.43	-0.84	-1.32
	(0.91)	(0.58)	(0.55)	(0.52)	(1.40)	(1.13)
12.5°-15°	-0.09	-0.13	-0.07	0.07	0.79	-0.13
	(0.45)	(0.39)	(0.34)	(0.31)	(0.78)	(0.69)
15°-17.5°	0.44	-0.21	0.45*	0.16	0.39	0.30
	(0.40)	(0.33)	(0.25)	(0.21)	(0.54)	(0.35)
17.5°-20°	-	-	-	-	-	-
20°-20.5°	-0.04	-0.20	-0.47*	-0.26	1.39***	0.31
	(0.33)	(0.24)	(0.28)	(0.20)	(0.53)	(0.38)
22°-25°	-0.31	-0.05	-0.94***	-0.54**	0.46	-0.35
	(0.40)	(0.34)	(0.35)	(0.27)	(0.64)	(0.48)
25°-27.5°	-0.19	-0.10	-1.22***	-0.47	-0.16	0.10
	(0.52)	(0.36)	(0.35)	(0.31)	(0.79)	(0.59)
$> 27.5°$	-0.67	-0.01	-1.91***	-0.52	-0.67	0.18
	(0.58)	(0.46)	(0.45)	(0.36)	(1.02)	(0.83)
Observations	6240	6240	5200	5200	5200	5200
School & Date FEs	✓	✓	✓	✓	✓	✓
Demo. Controls			✓	✓	✓	✓
District × Year FEs					✓	✓
Observations	13,466	13,466	11,493	11,493	11,493	11,493

Standard errors in parentheses

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

The table reports estimates from regressing pass rates on various fixed effects and controls for the panel of schools with 12 years of data. Standard errors reported in parentheses are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years.

Table B.2.4: Temperature Effect on Pass Rates (Balanced Panel)

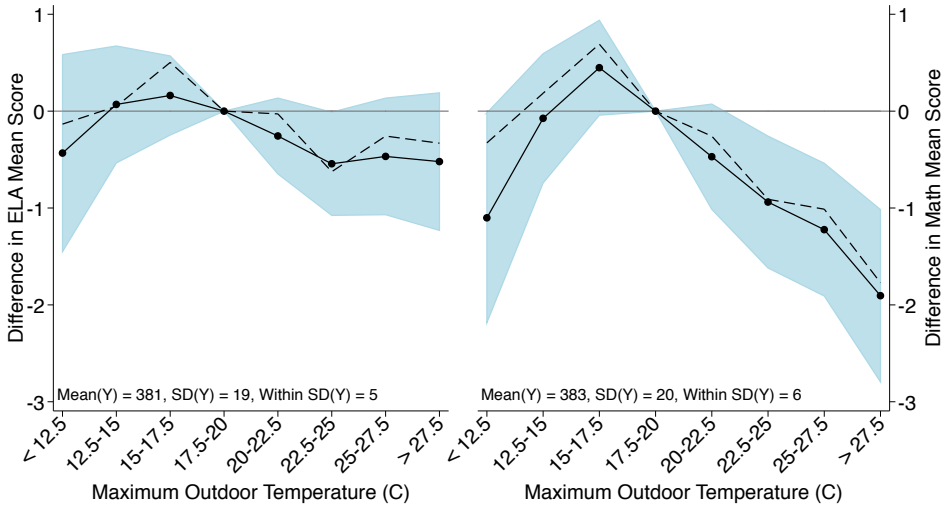
	Math	ELA	Math	ELA	Math	ELA
	(1)	(2)	(3)	(4)	(5)	(6)
Temperature Bin						
$< 12.5°$	0.005	0.006	-0.007	0.001	0.013	-0.004
	(0.010)	(0.006)	(0.004)	(0.004)	(0.011)	(0.009)
12.5°-15°	-0.000	-0.003	-0.000	0.002	0.018**	0.000
	(0.005)	(0.004)	(0.003)	(0.003)	(0.008)	(0.005)
15°-17.5°	0.000	-0.005	0.002	0.001	0.005	0.003
	(0.004)	(0.003)	(0.003)	(0.002)	(0.005)	(0.003)
17.5°-20°	-	-	-	-	-	-
20°-20.5°	-0.000	-0.003	-0.004	-0.005***	0.017***	0.003
	(0.004)	(0.003)	(0.003)	(0.002)	(0.005)	(0.004)
22°-25°	0.002	-0.000	-0.007**	-0.008***	0.005	-0.003
	(0.004)	(0.003)	(0.003)	(0.003)	(0.006)	(0.004)
25°-27.5°	0.003	-0.001	-0.009***	-0.008**	-0.001	-0.001
	(0.006)	(0.004)	(0.003)	(0.003)	(0.007)	(0.005)
$> 27.5°$	-0.001	0.001	-0.016***	-0.008**	-0.012	-0.004
	(0.006)	(0.005)	(0.004)	(0.004)	(0.007)	(0.008)
Observations	6,240	6,240	5,200	5,200	5,200	5,200
School & Date FEs	✓	✓	✓	✓	✓	✓
Demo. Controls			✓	✓	✓	✓
District × Year FEs					✓	✓
Observations	13,466	13,466	11,493	11,493	11,493	11,493

Standard errors in parentheses

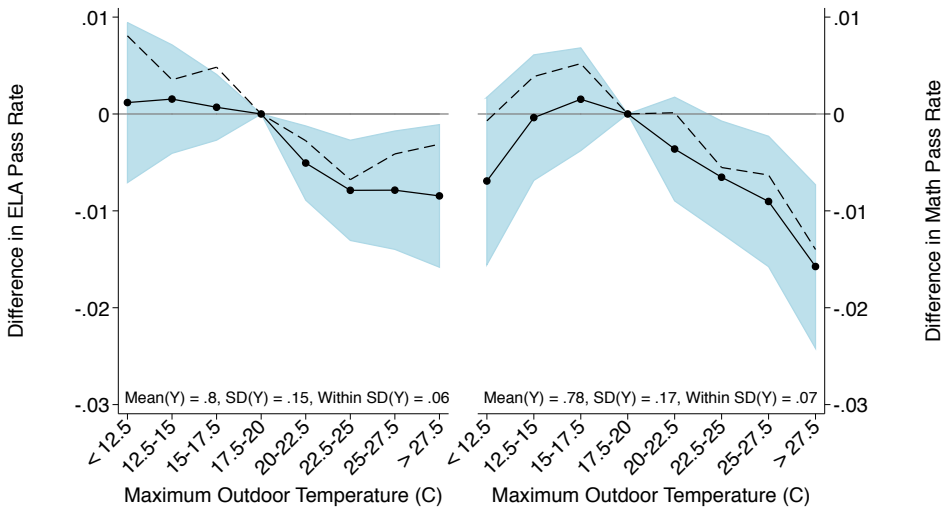
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

The table reports estimates from regressing mean scale scores on various fixed effects and controls for the balanced panel of schools with 12 years of data. Standard errors reported in parentheses are two-way clustered by school and county-year. Observations are weighted by the average number of students for a school over all available years.

Figure B.2.1: Temperature Effect (Balanced Panel)



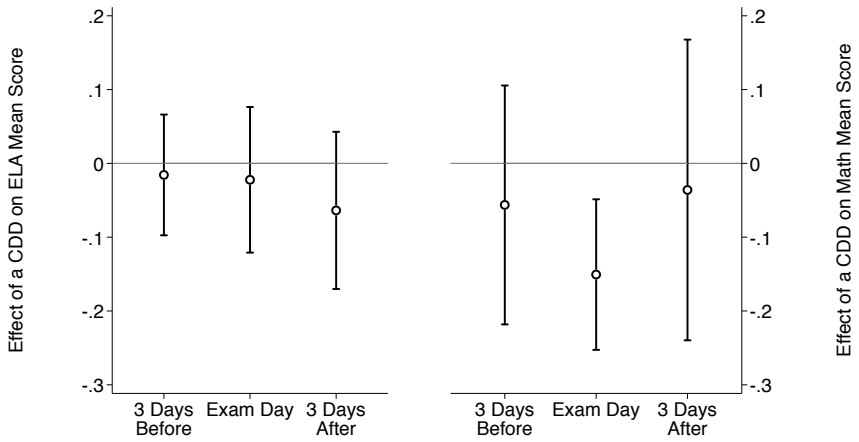
(a) Scale Scores



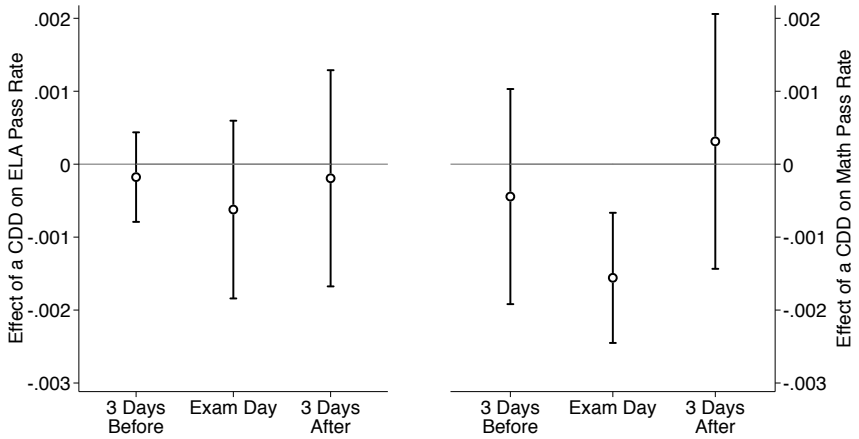
(b) Pass Rates

These figures plot causal estimates on the effect of the outdoor temperature on exam performance for the balanced panel of schools with 12 years of data. Response functions are reported for mean scale scores (top panels) and pass rates (bottom panels), and administrations of ELA (left panels) and math (right panels) sections. Each panel's estimates come from a separate regression estimated with weighting, school and test date fixed effects, and demographic controls, omitting the bin spanning 17.5-20°C. Shaded areas indicate 95 percent confidence intervals constructed using a covariance matrix two-way clustered by school and county by year. The dashed line indicates the unweighted OLS point estimates.

Figure B.2.2: Temperature Effect on Days Prior to and After Exam (Balanced Panel)



(a) Scale Scores



(b) Pass Rates

These figures plot causal estimates on the effect of the temperature on days before and after the exam on performance. The point estimates represent the effect of an additional cooling degree day (CDD) on a given performance metric and are reported for mean scale scores (top panels) and pass rates (bottom panels), and administrations of ELA (left panels) and math (right panels) sections. Each panel's estimates come from a separate regression estimated with weighting, school and test date fixed effects, and demographic controls. Vertical bars indicate 95 percent confidence intervals constructed using a covariance matrix two-way clustered by school and county by year.

Appendix C

The Economics of Targeting: Theory and Evidence from a Field Experiment – Appendix

C.1 Proof of Proposition 1

Proof of Proposition 1: We construct a proof by contradiction. Suppose $\sigma_\theta = 0$ and there exists a subsample $A \subset S$ of households for which the policy is more cost effective. Since $\sigma_\theta = 0$, $y_i = \theta D_i + \varepsilon_i$ for all households $i \in S$. Consider an arbitrary $B \subset S$ such that $A \neq B$. $E[y_i | i \in B] = \theta = E[y_i | i \in A]$. Since the choice of B was arbitrary, this contradicts the initial assumption so no A must exist. □

C.2 ITE Ranking

We estimate each ITE with the following model:

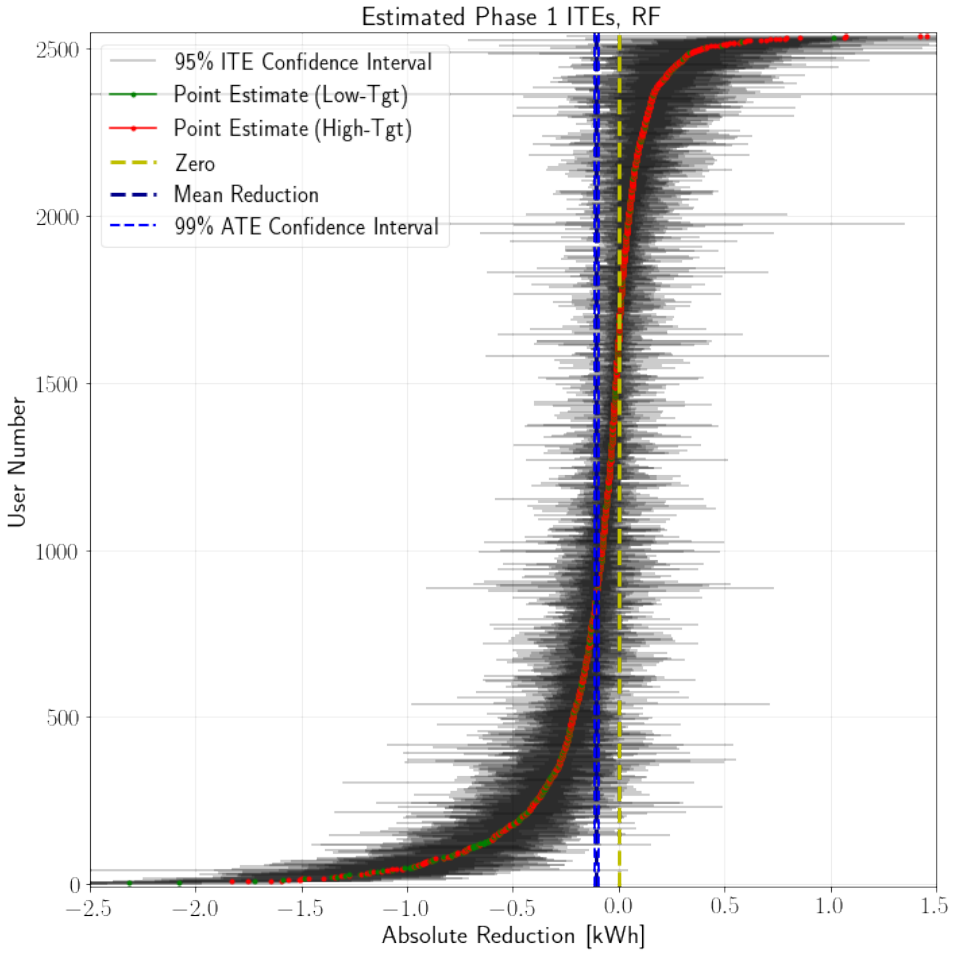
$$\widehat{\theta}_i = \frac{1}{|\mathcal{H}_i|} \sum_{t \in \mathcal{H}_i} y_{it} - f(X_{i,t-5}, \ldots, X_{i,t-1}; \beta_i) \tag{C.1}$$

where y_{it} is the kWh consumed and X_{it} is a vector of outdoor temperature and hour-of-day by month-of-year fixed effects for household i in hour-of-sample t. $f(\cdot)$ represents the counterfactual for that period as estimated using a regression model fitted with random forests for household i using data from non-event periods that occurred at least eight hours prior to or following an event. The set of periods during which events for household i were called is denoted $\mathcal{H}_i$. Thus, (C.1) shows

the ITE estimate $\hat{\theta}_i$ is the average of the difference between our random forest model counterfactual and actual consumption during events.

Figure C.2.1 shows the distribution of ITE estimates we use to rank individuals for both targeted and control households. MR households are shown as green points and LR as red points. These are estimated each week so the ranking is not monotonic over all since the median ITE that separates MR from LR changes. We show confidence intervals for our estimates, but these were not used in designating ITE rankings.

Figure C.2.1: ITE Distribution



The figure plots the distribution of ITE estimates we use to rank responsiveness. The green and red dots indicate the point estimate in terms of absolute reduction in kWh (on the horizontal axis). The green dots indicate those designated most responsive and red least responsive. The vertical yellow line indicates zero and the blue lines the average over all ITEs to produce an estimate of the ATE. We bootstrap 95 percent confidence intervals for each estimate which are shown as horizontal black lines.

C.3 Supplemental Empirical Results

This section reports the ATE results for the balanced panel in Table C.3.1 which is analogous to the Table 3.4.2 in the main body of the paper. We also report supplemental empirical results for the effect of targeting on consumption and the incentive level offered. Since we did not implement event messaging ourselves, the latter serves to validate that incentives levels were sent in accordance with design described. Figure C.3.2 reports the results the same event-study analogous to Figure 3.4.2 in the main body of the paper.

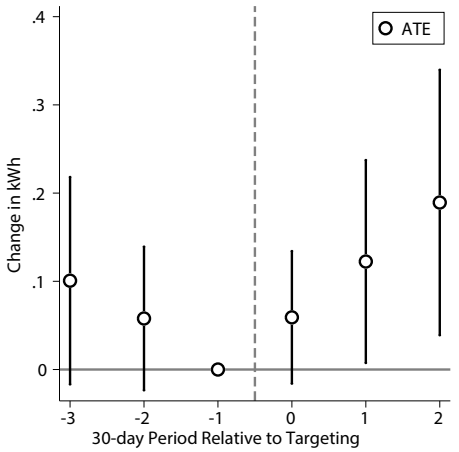
Table C.3.1: Effect of Targeting Strategy (Balanced Panel)

	Average Treatment Effects			Heterogeneity by ITE Rank		
	Payout (Points) (1)	Reduction (kWh) (2)	Usage (kWh) (3)	Payout (Points) (4)	Reduction (kWh) (5)	Usage (kWh) (6)
Intercept (Control)	17.33***	0.189***	0.824***			
	(2.29)	(0.023)	(0.051)			
$1(Most\ Resp.)_i$				24.15***	0.286***	0.835***
				(3.59)	(0.038)	(0.077)
$1(Least\ Resp.)_i$				9.37***	0.074***	0.823***
				(2.26)	(0.015)	(0.066)
D_i (Targeted)	−11.63***	−0.073***	−0.113*			
	(2.64)	(0.028)	(0.067)			
$D_i \times 1(Most\ Resp.)_i$				−19.42***	−0.089*	−0.144
				(3.60)	(0.047)	(0.101)
$D_i \times 1(Least\ Resp.)_i$				−2.77	−0.035*	−0.088
				(3.35)	(0.019)	(0.091)
Observations	13,107	13,107	13,107	13,107	13,107	13,107

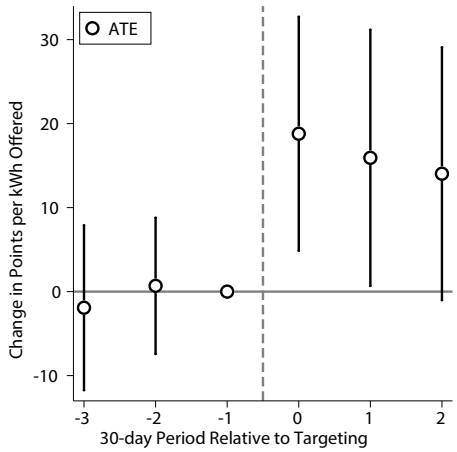
Standard errors are reported in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

This table reports estimates on the ATE of targeting for the full sample of 486 households. Columns (1)-(3) report the ATE estimates for all households and columns (4)-(6) report ATE estimates broken out by ITE rank. Columns (1) and (4) use individual payouts in points as the dependent variable, columns (2) and (5) use measured reductions in kWh, and columns (3) and (6) use usage/consumption in kWh. Standard errors are clustered by household and reported in parentheses.

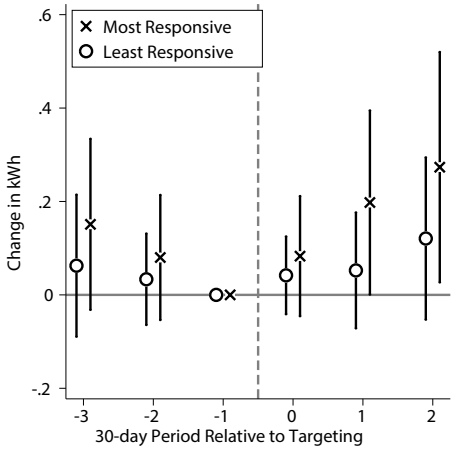
Figure C.3.2: Persistence of Targeting Effect on Incentives and Consumption



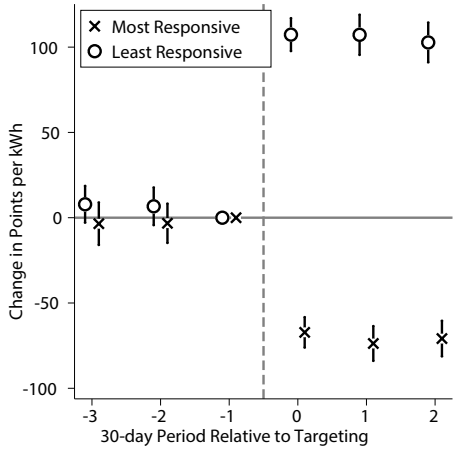
(a) Effect on Consumption



(b) Effect on Incentive Levels



(c) Effect on Consumption by ITE Rank



(d) Effect on Incentive Levels by ITE Rank

These figures plot event-study estimates on the effect of the targeting strategy over time for the balanced panel. The time intervals are 30 day periods 90 days prior to and following the start of the targeting strategy. The gray dashed line indicates the division between pre- and post-periods. Panels (a) and (c) represent estimates for consumption and panels (b) and (d) on incentive levels. Panels (a) and (b) show the ATE estimates for all households and panels (c) and (d) break out the effect by ITE rank. Vertical bars indicate 95 percent confidence intervals estimated using standard errors clustered by household. The omitted category is period -1 meaning the estimates represent changes relative to the period 30 days prior to rollout.

C.3.1 The Implementer's Problem and Optimal Targeting

In this section we describe the optimal strategy for the implementer operating the program and how the targeting strategy implemented in the experiment increases efficiency.

The firm generates revenue from the wholesale market for electricity and shares a fraction of the revenue with customers as follows. It takes the wholesale price and the fraction of revenue shared as exogenous. It's only dimension for optimization is choosing which households participate in the program and when to call events. For simplicity, we assume the events are called during periods when there is social value to reducing electricity consumption since the intricacies of the social return are secondary to the methodological contribution.[1] The firm is required to register participating households for several months at a time in order to claim revenue so it commits to an expected return on each customer.

Consider the profit generated by registering a sample of N households where consumption of household i in period t is given by

$$q_{it}(p) = q_i^0 + f(p;\eta_i) + \varepsilon_{it} \tag{C.2}$$

where $f(p;\eta_i) \equiv E[q_{it}|p] - E[q_{it}|p=0]$ represents a treatment effect function generated by the difference between the conditional expectation function (CEF) for household i of as a function of the incentive offered p and the level of consumption given an incentive of $p = 0$. ε_{it} is a mean-zero stochastic error term so it follows that $q_i^0 = E[q_{it}|p=0]$ represents what the household consumption would be in the event no incentive was offered. η_i is an individual preference parameter that allows the CEF to differ between households.[2]

For a period t, the revenue is given by the market clearing price p times the difference between the aggregate counterfactual forecast $\hat{Q}_t$ and consumption level $Q_t(p) = \sum_{i=1}^{N} q_{it}(p)$. The payouts are calculated as the fraction $\sigma \in (0,1)$ of the revenue generated on an individual basis following (**??**) with the incentive level being σp. Thus, for a set of N individuals the firm's profit in an event period t with wholesale price p_t is:

$$\pi_t(p_t) = \underbrace{p_t(\hat{Q}_t - Q_t(\sigma p_t))}_{wholesale\ revenue} - \underbrace{\sigma p_t \left(\sum_{i=1}^{N} \hat{q}_{it} - q_{it}(\sigma p_t)\right)}_{payouts} - \underbrace{cN}_{costs}. \tag{C.3}$$

[1] The implementer's objective function aligns with that of the social planner when the wholesale price is higher than the retail price and the deadweight loss reduced offsets the costs of the program operation. Gillan (2018b) explores whether these assumptions are likely to hold in the institutional setting.

[2] For example, if $f(p;\eta_i)$ were linear, η_i could represent the slope of demand. Alternatively if demand had a constant own price elasticity, η_i would govern the elasticity.

We also assume the firm experiences a constant marginal cost c for each household's participation.

We make two assumptions to simplify technical details related to the forecast. First, the methodology for forecasting aggregate consumption in the market settlement and the individual payout follows the same formula so we assume the individual forecasts sum to the aggregate in expectation so that $E_t[\hat{Q}_t] = \sum_{i=1}^{N} E_t[\hat{q}_{it}]$.[3] Further, we assume the forecast provides an unbiased prediction of counterfactual consumption if no incentive is offered so that $E_t[\hat{q}_{it}] = E_t[q_{it}(0)] = q_i^0$ and together with the previous assumption we have $E_t[\hat{Q}_t] = \sum_i q_i^0$. Lastly, we assume assume $p_t = p > 0$ for notational brevity and to clarify the dimension of optimization across individuals. Plugging these relations into (C.3), taking expectations over all events called during a registration period, and dividing by N shows the marginal trade-offs in terms of the expected return per household each event to be:

$$\frac{E_t[\pi_t(p)]}{N} = -(1-\sigma)\sigma p E[f(\sigma p; \eta_i)] - c \tag{C.4}$$

where $E[f(p;\eta_i)] = \int f(p;\eta) dF(\eta)$ is the expected response among the N participants given their distribution of preference parameters. Note $f(p;\eta_i) < 0$ if households reduce consumption when presented with a positive incentive $p > 0$ and we can see more elastic households to generate greater returns. If we partition households into those who generate a positive return $(-f(p;\eta_i) > c/(p\sigma(1-\sigma))))$ and those who generate a negative return $(-f(p;\eta_i) < c/(p\sigma(1-\sigma)))$, profits can be increased if participants with negative returns can be unregistered from the program. Conversely, when deciding whether to register new users, the firm could choose not to register those with negative returns.

The challenge is that the firm does not observe information about η_i explicitly. Thus, the firm's problem is analogous to a social planner who must identify and target which households for whom the social efficiency return is positive. To map this back to the notation introduced in Section 3.2, the relevant metric for efficiency would be defined

$$\theta_i = -f(p;\eta_i)p\sigma(1-\sigma) - c \tag{C.5}$$

where D_i from the earlier section refers to being registered in the program. Since p, σ, and c are exogenous to the firm's optimization, we focus on estimation of η_i.

[3]This assumption is reasonable since aggregation should not affect the first moment of the forecast distributions.