Who heeds the call to conserve in an energy emergency? Evidence from smart thermostat data

Dylan Brewer & Jim Crozier*

January 29, 2023

Abstract

In 2019, a fire at a natural gas plant and historically low temperatures caused an emergency shortage of natural gas in Michigan. To avoid an outage, the Governor issued a request via statewide text alert to turn thermostats down to 65°F. We analyze the effectiveness of this request using high-frequency smart-thermostat data from Michigan and four neighboring states. Using a difference-in-differences research design, we find that Michigan households reduced thermostat settings by 1.1 degrees on average following the Governor's request. Households that were previously above 65°F responded strongly, while households that were below did not respond at all or increased their thermostat settings. Meanwhile, households in districts that voted for the Governor in 2018 were more likely to comply. Our results suggest that unrealistic compliance goals and political polarization reduce the effectiveness of emergency calls to conserve energy.

Keywords: behavioral economics, nudges, moral suasion, energy use, natural gas, natural disasters, reference points, natural field experiments

*School of Economics, Georgia Institute of Technology, 221 Bobby Dodd Way, Atlanta, Georgia 30332, brewer@gatech.edu and jim.crozier@gatech.edu. We gratefully acknowledge the support and contribution of Ecobee and Ecobee customers to this research. This material is based upon work supported by the Google Cloud Research Credits program with the award GCP19980904. We thank Soren Anderson, Laura Taylor, A. Justin Kirkpatrick, Prabhat Barnwal, Alecia Cassidy, Casey Wichman, Nathan Chan, Elise Breshears, Matt Oliver, Cody Orr, and seminar participants at Virginia Tech, Ohio University, the Southern Economics Association Meetings, the Midwest Economics Association Meetings, the UCLA Climate Adaptation Research Symposium, and the AERE Summer Conference for useful discussion and feedback. Thank you to Graham Lewis for research assistance. Employees at Consumers Energy provided valuable perspective and total gas consumption data, for which we are grateful.

1 Introduction

During emergencies, government officials often make appeals to citizens to contribute effort to a common goal. These appeals have taken the form of requests to contribute effort to a public good (such as buying war bonds), to voluntarily ration consumption of scarce goods (such as reducing consumption of water during a drought), or to comply with safety protocols (such as taking certain precautions during a pandemic). Requests have become sophisticated over time as the communication medium has evolved from print materials, to radio and television announcements, and now to digital alerts with the ability to target specific individuals in real time. The strategy of requests has also evolved with advances in psychological research in how to influence behavior (Cialdini, 2006) and the use of the "nudge" framework to reduce the costs of compliance (Thaler and Sunstein, 2008). Modern requests for pro-social behavior often take the form of nudge reminders, informational treatments (Allcott and Taubinsky, 2015), appeals to morality or "moral suasion" (Ito et al., 2018), appeals to expert authority (Breza et al., 2021), or social comparison to peer behavior to induce compliance (Ferraro et al., 2011; Allcott and Rogers, 2014).

This paper analyzes the efficacy of an emergency request by the Michigan Governor for households to reduce thermostat settings during a natural gas shortage caused by a fire at a natural gas facility. During the cold wave of the 2019 polar vortex, outdoor temperatures were extremely low, causing high demand for natural gas for space heating. At 10:30 am on January 30, 2019, a fire broke out at Consumers Energy's largest natural gas storage facility. Consumers Energy is a gas and electric utility that serves roughly half of Michigan's households, and 75% of Michigan households rely on natural gas for heating. By 1:00 pm, officials at the utility had recognized that demand for natural gas might exceed supply, with the potential to cause the system to fail. At 2:30 pm, the utility requested via emails, social media, and news media that all households reduce natural gas consumption. At 10:00 pm as pipeline pressures continued to drop, the Michigan Governor tweeted a request to conserve natural gas and at 10:30 pm followed up with an emergency alert directly to cell phones in Michigan requesting that households reduce thermostat settings to 65°F or below. The utility

communications ensured that at least some households were aware of the request, but the cell phone alert went out to all households within the lower peninsula of Michigan. The next day at 4:30 pm, the utility issued an "all clear" time of midnight to its customers—thanks to voluntary reductions in demand by households and industrial consumers, the system did not fail and natural gas outages were avoided.

To measure household responses to the requests, we use smart thermostat data provided by Ecobee's Donate-Your-Data program. The data include thermostat setting and furnace fan run time at 5-minute intervals. This high-frequency, household-level data allows us to observe and measure each household's response to the requests as the emergency unfolded. Our empirical strategy uses households in the surrounding states of Ohio, Indiana, Illinois, and Wisconsin as control units for a difference-in-differences approach. We find that mean thermostat settings, the proportion of homes with a thermostat setting at or below 65°F, and furnace fan run time exhibit parallel pre-trends across the treatment and control units, which support our interpretation of our estimates as a causal effect of the emergency request on changes in thermostat settings.

Using a difference-in-differences strategy with four control states, we estimate the average treatment effect of the emergency request. We find that on average, households lowered their thermostats by 1.1°F, roughly 25% of the size of the typical variation in the average thermostat setting. The request increased the proportion of household thermostat settings at or below 65°F by 10 percentage points, a 45% increase relative to the proportion of households whose thermostat settings are normally 65°F or less. Finally, we examine the effect of the request on furnace fan run time, which is our best available proxy for household natural gas consumption. We find evidence that the emergency request reduced the furnace run time by 1.5 minutes per hour, a 6% reduction relative to the predicted run time for Michigan households during the emergency if there had been no reduction. These results are robust across a number of specifications and checks for spillover treatment to border counties, and a placebo test with an earlier cold wave shows that this behavior is not driven by outside temperature alone. An event-study analysis reveals that the Governor's amplification of the

utility's earlier request was critical for achieving complian ce. Prior to the Governor's alert, only 0.4% of additional households reduced thermostat settings to 65°F or less; after the Governor's alert, the additional compliance rate was as high as 20%. We interpret this as evidence that the Governor's authority was essential to increasing the salience of the appeal.

The utility and Governor framed the emergency request with a clear reference point of 65°F that affected household responses. We develop a theoretical framework that models a nudge with a compliance target as a moral tax on behavior that exceeds the reference point and as a moral subsidy on behavior that is below the reference point, extending the work of Levitt and List (2007). A nudge with a compliance goal creates two types of reference point heterogeneity: households that normally set the thermostat below this point were essentially exempted from the emergency request and given license to increase the thermostat setting, and households that normally set the thermostat significantly higher were asked to deviate more from their typical consumption patterns. Our model predicts that households with baseline thermostat settings below the 65°F target will increase their thermostat settings, and households with thermostat settings near the target will exactly comply with the request and therefore have a smaller response to the emergency request relative to households whose baseline thermostat settings are far from the compliance target. Finally, the model implies that a nudge with a compliance target imposes unequal marginal incentives on households, which violates the equimarginal principle and suggests that a nudge with a compliance target does not achieve the least-cost behavior change.

Empirically, we observe strong perverse framing effects in the data consistent with our model (Tversky and Kahneman, 1981). Using an estimate of each household's expected baseline temperature, we find that households that typically set the thermostat below 65°F were unresponsive to the emergency request on average. Households that are typically the coldest (far below 65°F) increased the thermostat after the emergency request. For baseline thermostat settings above the reference point, the higher the baseline thermostat setting, the less likely a household was to meet the compliance target. At first, the average thermostat reduction increases with distance from the compliance target, but for baseline settings 73°F

and above the average thermostat reduction decreases as the compliance rate effect begins to dominate. The results suggest that setting a more aggressive reference point trades off an increased treatment effect for individuals near the reference point with decreased compliance from discouraged individuals far from the reference point.

We scrutinize the role of political polarization as a factor in determining compliance with the request. The Governor, Gretchen Whitmer, assumed office in January 2019, less than a month before the polar vortex. We hypothesize that households that did not approve of the Governor may have been less likely to comply with the request. Our analysis studies the differential compliance of households in counties that supported the Governor's 2018 election bid, using data on county-level election returns. We show that compliance rates and the average reduction in thermostat setting are increasing in the Governor's vote share. Households in counties where the Governor's vote share was above 70% reduced thermostat settings by about twice as much relative to households in counties where the Governor's vote share was below 40%.

The results of this study are important for policymakers studying compliance with emergency requests in a broad range of fields. For instance, during the COVID-19 pandemic, local, national, and international governmental bodies sought to coordinate behavior to reduce the spread of the virus through a combination of compulsory policies and requests for voluntary compliance. Pandemic-related policies and requests were met with mixed compliance and even open defiance, and a growing literature seeks to understand the effects of political affiliation on cooperation with requests for social distancing and stay-at-home orders (e.g., Allcott et al., 2020; Barrios and Hochberg, 2020). Related to compliance with energy and environmental policy, economists have studied firm strategic avoidance of air quality monitoring (Zou, 2021), imperfect enforcement of emissions caps (Sigman and Chang, 2011), voluntary reductions of emissions (Foster et al., 2009; Foster and Gutierrez, 2013), compliance with the US acid rain program (Montero, 1999), and the use of regulatory loopholes to avoid compliance with fuel efficiency regulations (Anderson and Sallee, 2011). Beatty et al. (2019) study household emergency preparedness for hurricanes, finding that household

behavior is highly influenced by recent hurricane events and that households generally do not follow government preparedness recommendations. Other work shows that an increased perception of risk and confidence in government institutions increases compliance with hurricane evacuation orders (Whitehead et al., 2000; Kim and Oh, 2015). In another context, Wichman et al. (2016) find that households in North Carolina reduced consumption of water during a drought when both voluntary and mandatory non-price mechanisms were implemented to restrict water use. During or after energy emergencies, utilities and governments often resort to emergency appeals for conservation. For example, Luyben (1982) studies a 1977 request by President Carter for US households to reduce thermostat settings to 65°F or below, finding that compliance was low overall (27%) and that self-reported compliance was higher than recorded compliance. After the 2000 and 2001 California energy crisis, energy conservation campaigns were successful in reducing electricity consumption when electricity prices were capped (Reiss and White, 2008). Our paper contributes to these literatures by providing what we believe is the most granular data on household compliance with emergency requests. In addition, the unexpected nature of the emergency and its isolation to one state creates a credible natural experiment that allows us to pursue an identification strategy that takes advantage of plausibly exogenous time and cross-sectional variation, which is uncommon for this literature. The potential for political polarization in our setting is particularly salient, given the proximity of the emergency to the Governor's election in a politically divided state.[1]

Our work also contributes to the empirical literature studying reference points and economic behavior. Research in this area examines labor supply behavior relative to earnings expectations (Thakral and Tô, 2021; Farber, 2008; Camerer et al., 1997), retirement decisions relative to age reference points (Seibold, 2021), and loss aversion in tax filing (Engström et al., 2015). Other work focuses on the use of social comparison as a reference point to influence behavior and is often applied to energy conservation (e.g., Allcott (2011), Ferraro et al. (2011), Ferraro and Price (2013), Brent et al. (2015), and Hallsworth et al. (2017)). In the charitable

[1]Also of note is that Governor Whitmer was later the target of a politically-motivated abduction plot in response to her COVID-19 lockdown policies.

giving literature, suggested donation amounts increase voluntary contributions and anchor donations to the suggested amount (Edwards and List, 2014). Harding and Hsiaw (2014) study how non-binding goal setting for energy conservation leads to behavior consistent with reference-dependent preferences. Brown et al. (2013) find that factory-default thermostat settings substantially impacted subsequent thermostat levels chosen in the workplace. Our paper contributes to this literature by studying a novel reference point created within the phrasing of a governmental emergency request. Our results suggest that for the policymaker, setting a reference point more aggressively trades off an increase in the effect of meeting the reference point with the cost of meeting the reference point. In our context, further lowering the requested thermostat setting would have reduced compliance from those with high baseline settings but would have increased the effort from those with medium and low baseline settings. These findings imply that policies may be designed so as to have effect-maximizing reference point levels. Finally, our theoretical framework extends the moral payoff model of Levitt and List (2007) to nudges with a compliance target, providing new predictions for heterogeneous behavior based on a person's pre-treatment distance from the reference point. This is particularly important because we show that compliance targets can cause households to respond perversely when they are already in compliance, undermining the goal of the policy.

In addition, this paper is relevant to the literature in environmental and energy economics analyzing the use of non-price mechanisms to conserve household consumption of water, natural gas, and electricity.[2] Given political constraints on raising prices of these goods, regulators and suppliers have sought to curb consumption via mandatory restrictions and voluntary requests, which have seen varying levels of success. In Ito et al. (2018), the authors conduct a field experiment that provided Japanese households with voluntary appeals or price incentives to reduce electricity consumption. Relative to a control group, voluntary appeals resulted in a short term reduction in electricity consumption of 8% while price incentives resulted in a reduction in electricity consumption of 17% that was sustained over

[2]See Carlsson et al. (2021) for a recent overview of papers analyzing nudges and non-price mechanisms.

a longer period. In the United States, Holladay et al. (2015) find that utility and media appeals for peak-hour conservation can perversely lead to increases in energy consumption in anticipation of an outage, Burkhardt et al. (2019) find very little responsiveness to voluntary appeals to conserve during peak hours relative to price mechanisms, while Brandon et al. (2018) find that utility-led requests resulted in a 4% average reduction in consumption during peak hours. In our setting, we document that households were unresponsive to utility and media appeals, but the intervention of the Governor via the wireless alert system resulted in compliance of a similar magnitude to the field experimental results in Ito et al. (2018) and Brandon et al. (2018). Our paper contributes novel evidence that reach and authority of the messenger can substantially affect the salience of emergency appeals for conservation, and that political context substantially impacts the effectiveness of non-price mechanisms. Allcott and Rogers (2014) find evidence that households reduce electricity consumption in response to home energy reports and that repeated treatments induce additional reductions and enforce habits, while Costa and Gerard (2021) find that nine-month quotas on energy consumption provided reductions in energy use up to nine years after the quotas ended. In contrast, we find that without repeated treatments that household compliance wanes by the end of the request, and while some conservation persists after the all-clear, the effect is modest.

Our paper proceeds by describing the polar vortex and natural gas fire events in greater detail. We develop a theoretical model of thermostat setting choice in the presence of a nudge with a compliance target, which generates hypotheses for household behavior. We then introduce the smart thermostat data used in the paper. Next, we present our empirical strategy and analysis, which we subdivide into a section estimating the average treatment effect of the request, a section presenting an event-study analysis, a section examining the effects of the reference point on behavior, and a section examining the role of political support of the Governor on compliance. We discuss the external validity of the estimates in section 6. Finally, section 7 summarizes the findings and concludes.

2 Polar vortex and natural gas fire events

Extreme cold weather events caused by disturbances to the polar vortex have recently received significant attention in the United States and Europe. Perhaps most notable was the 2021 polar vortex event that overwhelmed the electricity grid in Texas, killing 172 people and resulting in damages valued at levels ranging from $20 billion to $295 billion (NOAA, 2021; Perryman Group, 2021). Since 1980, winter disasters have resulted in 19 "billion-dollar climate disasters" in the United States, causing 1,223 deaths (NOAA, 2021). There is only weak evidence that climate change is contributing to the perceived increase in polar vortex events (Blackport and Screen, 2020); however, aging energy infrastructure in the United States and Europe may increase the costs of such events in the future.

Beginning on Tuesday, January 29, 2019, temperatures in the Midwest declined to nearly record-low levels as cold air in the stratosphere over the Arctic blew southward over North America (NOAA, 2019). Temperatures reached -23°F in Chicago, -13°F in Detroit, -11°F in Indianapolis, and as low as -45°F elsewhere in the United States (EIA, 2019b). On Wednesday, January 30, 2019, single-day estimated natural gas consumption in the United States hit an all-time high with 37.9 billion cubic feet consumed in a single day, and electricity demand in the Midwest approached all-time peaks (EIA, 2019a). In 2017, over 75 percent of Michigan homes used natural gas as the primary heating fuel (MPSC, 2019b).

Coinciding with this extreme demand-side stress, a supply-side emergency caused a near system-wide natural gas delivery failure in Michigan. On January 30, 2019 at 10:30 am, a fire broke out at the Ray Compressor Station in Macomb County, Consumers Energy's largest natural gas storage facility (MPSC, 2019b). Immediately after the fire broke out, the utility drew upon standby natural gas reserves to stabilize pipeline pressures (Consumers Energy Company, 2019). By 1:00 pm, Consumers Energy recognized the possibility that demand could exceed supply, which could cause total system failure, and contacted their highest demand industrial and commercial customers with requests to reduce consumption of natural gas. At 2:26 pm, Consumers issued a tweet (appendix figure 9a) requesting households to reduce thermostat settings and sent emails to residential and business customers

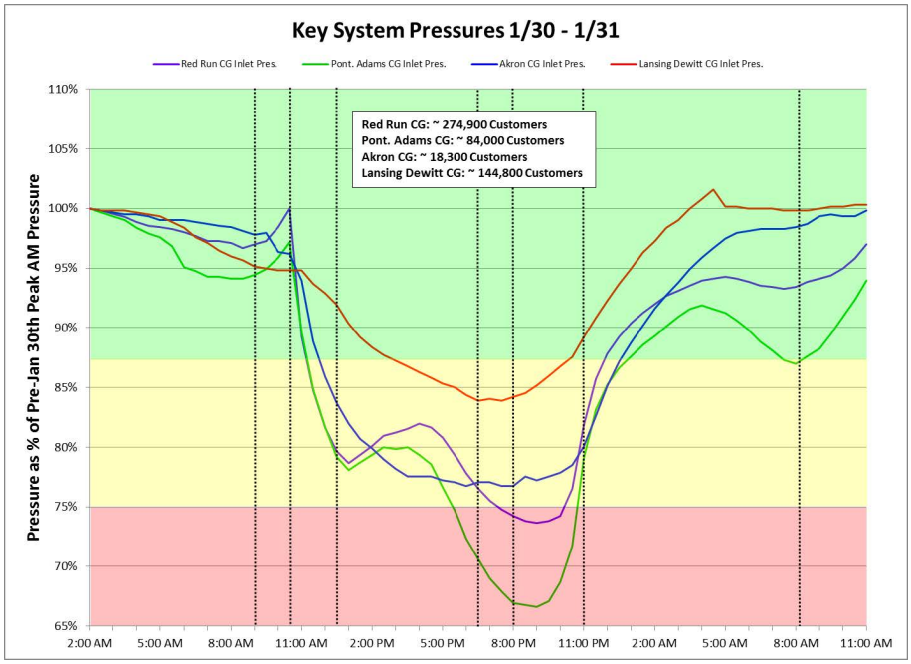


Figure 1: Each series corresponds to a Michigan natural gas pipeline instantaneous pressure, January 30-31, 2019. Image source: Michigan Public Service Commission Case No. U-20463 (Consumers Energy Company, 2019).

requesting reductions in natural gas use. Shortly thereafter, the CEO of Consumers Energy took to Facebook Live to urge households to reduce thermostat settings (appendix figure 9b). The utility ultimately sent over 500,000 external emails, made 21 social media posts, and responded to 130 media inquiries on January 30-31 (Consumers Energy Company, 2019). State-operated buildings reduced thermostat settings by 5°F and manufacturers reduced consumption of natural gas (DesOrmeau, 2019). In addition, the utility issued mandatory curtailment orders for industrial and commercial natural gas customers and requested that natural gas electricity generators reduce generation to preserve residential heating (Consumers Energy Company, 2019). On the supply side, Consumers Energy purchased 925 MMcf/day worth of same-day supply of natural gas for January 30th, of which only 61% was ultimately delivered due to supply constraints (Consumers Energy Company, 2019).[3]

[3]Same-day natural gas delivery is relatively rare compared to same-day electricity generation, for example. This event was the first time that Consumers Energy had attempted to secure same-day delivery (Consumers Energy Company, 2019). For extreme-weather events, utilities can increase pressure in natural gas pipelines ahead of time, storing gas within the system. Given that the flow of gas is not instantaneous, same-day supply is not typically used to balance supply and demand.

Despite efforts to reduce non-residential consumption and to procure natural gas on the supply side, the system was still unpacking (losing pressure) going into the evening. Figure 1 displays Michigan natural gas pipeline pressures on January 30th and 31st. Despite efforts to curb demand and increase supply, equilibrium pressures were dropping as the evening approached and temperatures continued to get colder. At 8:00 pm, Consumers Energy reached out to the Governor of Michigan, Gretchen Whitmer, to make a final public appeal to households to reduce thermostat settings (Consumers Energy Company, 2019).

At 10:01 pm, the Governor of Michigan tweeted a request for households to reduce thermostats to 65°F (appendix figure 9c), and at 10:30 pm activated FEMA's Wireless Emergency Alert system to send a text alert to all cell phones in Michigan asking households to reduce thermostat settings to 65°F (Gray, 2019). The text of the cell phone alert message read "Due to extreme temps Consumers asks everyone to lower their heat to 65 or less through Fri." Conversations with the Michigan State Police Emergency Management and Homeland Security department and Consumers Energy indicated that officials believed 65°F was achievable, comfortable, and likely to be lower than the usual thermostat setting, but the number was chosen arbitrarily.

Shortly after the Governor's text message at 10:40 pm, 30% of the Ray Compressor Station capacity came back online, which combined with demand reductions to begin to increase pressures (Consumers Energy Company, 2019). Using data provided by Consumers Energy, forecasted natural gas demand using realized weather conditions was 3.3 billion cubic feet on January 30th and 2.9 billion cubic feet on January 31st. After all reductions in consumption were accounted for, actual consumption was 3.0 billion cubic feet on January 30th and 2.6 billion cubic feet on January 31st, implying a 10.7% and 10.5% reduction in daily consumption from all sources (residential and non-residential). On January 31st at 4:30 pm, Consumers Energy tweeted an "all clear" time of midnight that night, after which households could resume heating normally (appendix figure 9d).

Did households listen and comply with the emergency requests issued by the utility and public officials? Furthermore, how did the phrasing of the request around a thermostat

setting of 65°F affect household compliance? Given the Governor's request, did political polarization affect which households were likely to comply? We introduce a theoretical framework to generate hypotheses and test them using high-frequency smart thermostat data.

3 Theoretical framework

Here we develop a theoretical model in which utility-maximizing households choose thermostat settings in the presence of an emergency request to reduce thermostat settings to a compliance target. Typically, households choose the thermostat setting that equates their marginal benefit from a degree Fahrenheit with the marginal cost of increasing the thermostat by an additional degree Fahrenheit.[4] We model the emergency request to reduce thermostat settings to 65°F as an additional moral payoff term in the utility function similar to the moral payoff term introduced by Levitt and List (2007) and considered in Ferraro and Price (2013). In our case, we model the emergency request as a moral tax on thermostat settings above the requested reference level, a moral subsidy for thermostat settings below the requested reference level, and having no marginal incentive for thermostat settings exactly equal to the requested reference level.

Our model generates predictions that the emergency request will cause households with baseline thermostat settings above 65°F to reduce thermostat settings and households with thermostat settings below 65°F to increase thermostat settings. Because of the nonlinearity in the marginal incentive at the 65°F target, households with baseline thermostat settings near the target will have an incentive to exactly comply with the request, while households far from the target will partially comply but have a larger treatment effect, all else equal. The model thus suggests two perverse framing effects of the 65°F target. First, households heating at temperatures below the target will increase the thermostat setting and consume more energy. Second, households with thermostat settings near 65°F will only have an

[4] A choice modeled in Brewer (2022).

incentive to reduce the thermostat setting by a small amount.

We consider a household i with characteristics θ_i choosing the thermostat setting T_i in any given time period. The regulator announces a requested thermostat setting R (which is equal to 65°F in our context). The utility function for household i is linearly separable in consumption benefits $B_i(\cdot)$, energy heating costs $C_i(\cdot)$, and the moral payoff $M_i(\cdot)$:[5]

$$U_i(T_i, R, s; \theta_i) = B_i(T_i; \theta_i) - C_i(T_i; \theta_i) - M_i(T_i, R, s; \theta_i). \tag{1}$$

We assume that locally, households weakly prefer higher temperatures $\partial B_i/\partial T_i \geq 0$ at a diminishing rate $\partial^2 B_i/\partial T_i^2 < 0$. In addition, heating costs weakly increase with higher thermostat settings $\partial C_i/\partial T_i \geq 0$ and are weakly convex $\partial^2 B_i/\partial T_i^2 \geq 0$. The moral payoff term exerts a moral marginal cost on consumption above the requested reference point, has no marginal influence when thermostat setting equals the requested reference point, and exerts a moral subsidy on consumption below the reference point. Similar to Ferraro and Price (2013), s represents the salience or strength of the request. Thus, the moral payoff term has the following marginal effects on the household's utility:

$$\partial M_i/\partial T_i = \begin{cases} \tau(T_i) > 0 & \text{if } T_i > R, \\ 0 & \text{if } T_i = R, \\ \sigma(T_i) < 0 & \text{if } T_i < R. \end{cases} \tag{2}$$

When there is no emergency request, the household maximizes utility by choosing the thermostat setting that equates the household's marginal benefit of indoor temperature with the marginal energy cost in the first-order condition:

$$\partial B_i/\partial T_i = \partial C_i/\partial T_i. \tag{3}$$

[5]Linear separability in heating costs is not necessary to derive the comparative statics in this section but simplifies the notation and exposition.

When there is an emergency request, the household's marginal cost of heating includes both the marginal moral cost and the marginal energy cost of heating. The household maximizes utility by equating the marginal benefit of indoor temperature with the new marginal cost of heating in the first order condition:

$$\partial B_i/\partial T_i = \partial C_i/\partial T_i + \partial M_i/\partial T_i, \tag{4}$$

or by choosing a corner solution of exact compliance with the request: $T_i = R$. Thus, a salient emergency request causes *all* households to move the thermostat closer to the reference level, whether they initially were heating above or below the reference point.

In this context, the strength or salience s varies based on the platform and identity of the messenger. The utility company broadcast the initial emergency request, which was also taken up by local traditional news media from 2:30 pm - 10:00 pm. Beginning at 10:01 pm, the Governor took up the emergency appeal and issued the cell phone alert at 10:30 pm. We hypothesize that the initial request was a very low-salience request, but that the Governor's requests substantially increased the salience. Prior work studying emergency appeals for energy conservation have found that appeals via channels such as local news fail to induce reductions in energy consumption or can perversely cause households to increase energy consumption in expectation of a potential outage (Holladay et al., 2015). Thus, we expect that most reductions in thermostat setting will occur after 10:00 pm, reflecting the change in s.

Figure 2 illustrates the model with constant marginal cost and linear demand curves for four hypothetical households. Households $i \in \{A, B, C, D\}$ have marginal willingness to pay for heating D_i. Prior to the emergency request, each household equates marginal willingness to pay with marginal cost of heating and choose thermostat setting T_i^0. After the request, the marginal cost of heating now includes the moral marginal cost. Household A partially complies, reducing the thermostat setting to $T'_A > 65$, while household B fully complies, choosing the corner solution $T'_B = 65$. The treatment effect for household B is thus

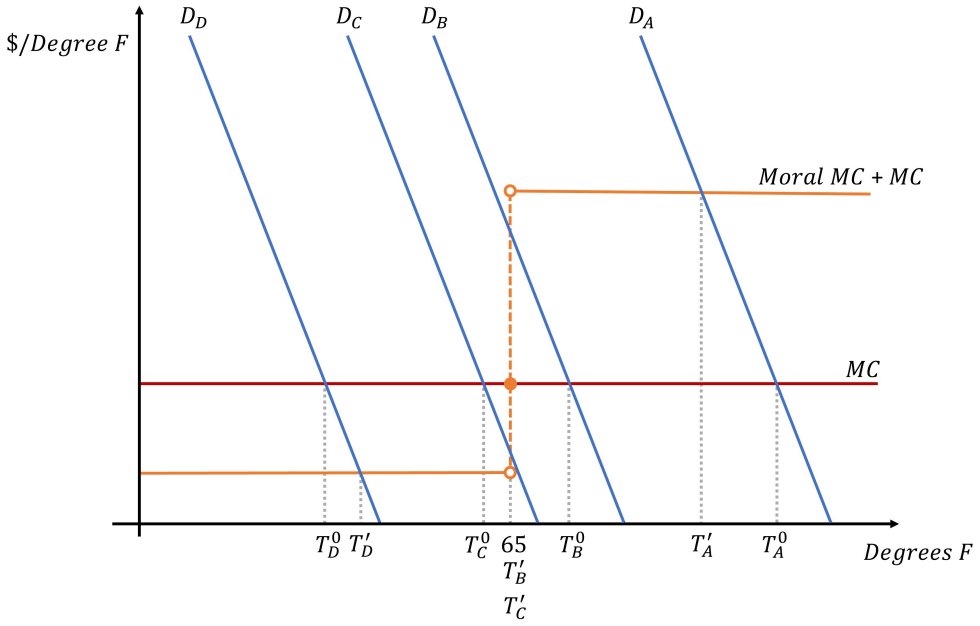


Figure 2: Theoretical effect of a request to reduce thermostats to 65°F for four household types with marginal willingness to pay for heating D_i for $i \in \{A, B, C, D\}$. Before the emergency request, households equate marginal willingness to pay with the marginal cost of heating and choose thermostat setting T_i^0. The request acts as a moral tax on temperatures above 65°F and a moral subsidy on temperatures below 65°F, causing households to choose thermostat setting T_i'. Household A partially complies, while household B fully complies. Households C and D perversely increase the thermostat setting toward the reference level.

limited by the reference point. The behavior of hypothetical households A and B generates our first two empirical hypotheses. The first hypothesis is that households with baseline thermostat settings T_i^0 near the reference point will be more likely to comply fully with the request relative to households far above the reference point. The second hypothesis is that households with baseline thermostat settings far from the reference point will have a larger average treatment effect as they are less likely to achieve the corner solution. We believe this effect is likely to attenuate for households with the highest baseline thermostat settings, as the compliance target may appear out of reach or unrealistic. This attenuation would be consistent with a moral cost function decreasing in thermostat setting, or $\tau'(T_i) < 0$.

Households C and D respond perversely to the emergency request and increase thermostat settings in response to the moral subsidy. Household C mirrors the behavior of household B and chooses the corner solution $T_C' = 65$, while household D mirrors the behavior of household A and increases the thermostat setting toward the reference level. For households with baseline thermostat settings below 65°F, we hypothesize that we will see increased thermostat settings with similar heterogeneity based on the distance of the baseline thermostat setting T_i^0 from the reference level. Thus, for households with baseline thermostat settings near but below 65°F, we expect to see more perverse compliance by setting the thermostat exactly to the reference point and a smaller perverse treatment effect relative to those with baseline thermostat settings far below the reference point.

Finally, this model suggests that an emergency request tied to a reference point is not the least-cost nudge required to achieve a given reduction in energy consumption. Equation 2 shows that households will face heterogeneous marginal costs of energy consumption, violating the equimarginal principle (also called the equal marginal principle or Gossen's second law), immediately implying this nudge is of higher cost relative to a nudge that exerts a moral marginal cost of consumption that applies no matter the household's baseline thermostat setting. Perhaps more intuitively, a nudge that perversely subsidizes additional consumption of energy for some households when energy is scarce increases the mismatch between the retail price of energy and the wholesale price of energy for those households.

While it is possible that the reference level increases the strength or salience of the request by providing households with a concrete action to perform, a request for a uniform reduction in the thermostat setting (e.g., a request to reduce the thermostat setting by 5°F) would still be concrete. Such a uniform request would rule out perverse behavior, but would still not satisfy the equimarginal principle. We revisit these implications for the design of an emergency request in the conclusion.

4 Smart thermostat data and research design

We use data on smart thermostat temperature settings provided by Ecobee as part of the 2022 release of the "Donate Your Data" program.[6] The raw data include 5-minute interval observations of thermostat settings and the amount of time the furnace fan was running. In addition, a small amount of information about the household is available, including the location up to city and state, the number of occupants, the size, age, and number of floors of the home, and when the smart thermostat was first connected. We augment these data with hourly outdoor temperature, humidity, wind speed, precipitation, snow depth, and cloud cover at the city level purchased from Visual Crossing. Consumers Energy only serves households in Michigan. We limit the sample of households to those in Michigan and the surrounding four states for controls: Ohio, Indiana, Illinois, and Wisconsin.[7] 99.89 percent of sample households heat with natural gas, compared to 75 percent of population households in Michigan. We include all observations between January 2nd and February 3rd, 2019. There are 3,036 households from Michigan and 9,221 control households in the final sample.

It is possible that the households in our sample responded to the emergency request

[6]This paper is among a few others studying the effects of smart thermostats or using smart thermostat data. Ge and Ho (2019) study how households change the thermostat in response to warm and cold weather and assess the degree of habit formation in thermostat settings. A working paper by Brandon et al. (2021) find that smart thermostats alone do not result in energy savings, partially due to users overriding smart thermostat algorithms. Another working paper by Blonz et al. (2021) studies an energy-efficiency program implemented by Ecobee that automatically reduces thermostat settings during peak pricing periods.

[7]We exclude households in the Upper Peninsula of Michigan because these households are on a separate natural gas network and it is unclear whether they were treated or were controls. The Upper Peninsula accounts for about 3% of the population of Michigan; dropped households represent 1% of the Michigan sample in the data.

Table 1: Summary statistics for households in Michigan and in the control states (Ohio, Indiana, Illinois, and Wisconsin). The sample includes observations from January 2nd-February 3rd.

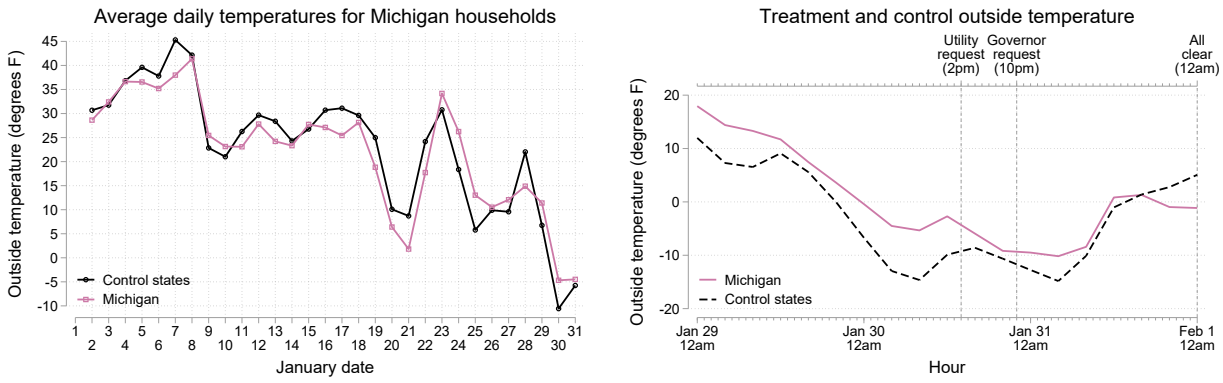
	(1) Michigan Mean/SD	(2) Controls Mean/SD	(3) Difference Diff./t-stat
Sq ft	2,387.18	2,545.96	158.78**
	(1,033.94)	(1,138.21)	(6.83)
Age of home (years)	32.73	33.96	1.23*
	(29.00)	(31.25)	(1.98)
Number of occupants	1.19	1.34	0.15**
	(1.70)	(1.75)	(4.27)
January 2 - 29 thermostat setting	66.87	67.45	0.58**
	(3.60)	(3.32)	(7.83)
January 30 thermostat setting before event	67.38	68.20	0.82**
	(3.93)	(3.79)	(9.88)
January 30 - 31 thermostat setting during event	67.09	68.67	1.58**
	(3.48)	(3.56)	(21.35)
January 2 - 29 outside temperature	23.86	25.08	1.22**
	(2.33)	(4.18)	(20.02)
January 30 - 31 outside temperature	-0.29	-0.50	-0.22**
	(0.24)	(0.55)	(-29.69)
Households	3,036	9,221	12,257
Observations	581,163	1,780,876	2,362,039

** p<0.01, * p<0.05

differently than the general population due to selection into smart thermostat ownership and the Donate-Your-Data program. On observable characteristics, the Ecobee Donate-Your-Data households are comparable to the average household in the nationally representative Residential Energy Consumption Survey sample, though the Ecobee households have slightly more members (Meier et al., 2019). Our primary selection concern is that Ecobee households who join the Donate-Your-Data program may be more likely to contribute to other public goods and therefore more likely to comply with the emergency request. Another concern we have is that the Ecobee smart thermostat may make compliance with the request easier than compliance using a conventional thermostat because Ecobee thermostats can be controlled remotely via an app. While these issues are not a problem for our research design because treatment and control households are the same (i.e., our research design is internally valid), it may be that our estimated treatment effects overstate the response of the average household. In section 6 we discuss the external validity of the estimates in more detail. We find that our estimated treatment effect is smaller than that estimated using aggregate natural gas consumption data provided by the utility, the opposite of what we would expect if smart thermostat users were more likely to comply with the request.[8] We also show in that section that early adopters of smart thermostats responded similarly to late adopters. These findings mitigate our concerns that our estimates may not generalize to those with conventional thermostats.

For computational tractability and to reduce noise, we aggregate the data into four-hourly time intervals, resulting in 581,163 household-time observations in Michigan and 1,780,876 household-time observations in the controls. We compute the four-hourly average thermostat setting and minutes per hour the furnace was running. Analysis at the hourly level or lower does not substantially change the estimates, which we display in appendix C. Table 1 displays summary statistics for the treatment and control groups. Due to the large sample size, most differences in means for treatment and control are statistically significant, but are not practically meaningful and do not pose a threat to our empirical strategy. The pri-

[8] Aggregate consumption data includes residential, commercial and industrial consumption.



(a) (b)

Figure 3: (a): Sample mean daily temperatures for Michigan and control households in January 2019. January 20-21 are used as a "placebo" event for the January 30-31 polar vortex and emergency request. (b): Sample mean four-hour average outdoor temperatures for Michigan and control households in the hours before and during the emergency.

mary differences we see are that sample homes in Michigan are slightly smaller and have fewer occupants on average. During January 2 - 29, Michigan and the control states experienced average temperatures around 24 and 25°F. The outside average temperature in both treatments and controls dropped to just under 0°F during the event. From January 2nd through 29th, Michigan household thermostat settings were 0.6°F lower than the control household thermostat settings. In the hours before the first appeal to lower thermostats, the gap in thermostat settings had increased to 0.78°F. After households were asked to reduce the thermostat, the gap increased to 1.58°F.

Our research design compares outcomes in Michigan to those in control states where there were no appeals to reduce natural gas consumption. We consider three outcomes: the thermostat setting, a binary variable equal to one if the thermostat setting is at or below 65°F, and the amount of time the furnace fan ran during the hour. Standard furnaces run at essentially one speed.[9] When the thermostat setting is reduced, the home cools to the new setting and the furnace does not run, saving energy. When the indoor temperature goes below the new thermostat setting, the furnace runs again at full speed for a short period to maintain

[9]Two-stage furnaces can run at full-speed and half-speed depending on the scenario to reduce energy use and ramping costs.

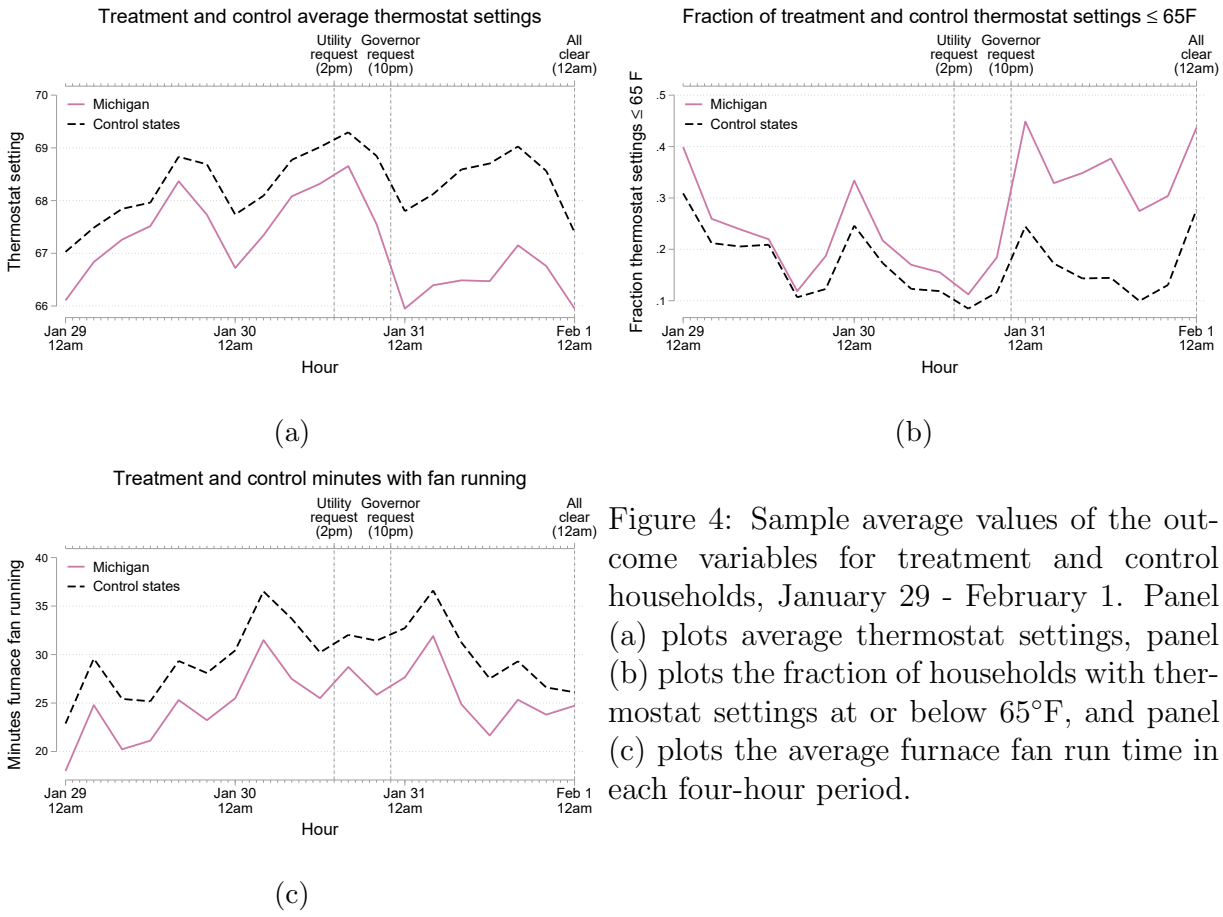


Figure 4: Sample average values of the outcome variables for treatment and control households, January 29 - February 1. Panel (a) plots average thermostat settings, panel (b) plots the fraction of households with thermostat settings at or below 65°F, and panel (c) plots the average furnace fan run time in each four-hour period.

the indoor temperature. Thus, furnace fan running time is our best proxy for natural gas consumption (Meier et al., 2019). Consumers Energy shared daily aggregate natural gas consumption and forecasts of expected consumption from their internal forecasting model, which we use to construct an estimate of total demand response from all sources.

Households in the control states (the "Great Lakes states": Ohio, Indiana, Illinois, and Wisconsin) have weather patterns and housing stocks similar to Michigan. Furthermore, these states also experienced extreme cold during the polar vortex event. Figure 3a plots daily average temperatures during January for Michigan and control states, showing the polar vortex event at the end of the month in addition to a similar cold wave on January 20th and 21st that we study in a placebo exercise in the appendix. Figure 3b plots the four-hour average outdoor temperature before and during the emergency, demonstrating that both treatment and control groups experienced similar conditions during the polar vortex.

Because we observe treatment and control households before and during the event, the difference-in-differences framework is a natural candidate to estimate the effect of the emergency request on thermostat settings. The key assumption needed in a difference-in-differences design is a parallel trends assumption in the evolution of the potential untreated outcome. To demonstrate the validity of the difference-in-differences assumption, we demonstrate visually that our outcome variables exhibit parallel trends prior to the event (we also provide event-study estimates with 10 days of pre-trend estimates in section 5.2). Figure 4a plots four-hour sample average thermostat settings, figure 4b plots the fraction of households with thermostat settings at or below 65°F, and figure 4c plots the number of minutes per hour the fan was running in Michigan and the control states from January 29th through February 1st. The first vertical dashed line indicates when the utility company first broadcast a request to residential customers to reduce natural gas consumption by reducing thermostats, the second vertical dashed line indicates when the Governor broadcast the emergency appeal, and the final vertical dashed line indicates the all-clear time.

Prior to the event, the treatment and control thermostat settings, fraction of households with thermostat settings at or below 65°F, and the number of minutes the fan was running in Michigan exhibit roughly parallel trends even without conditioning on covariates. When the event begins, a take-up lag can be observed where households have either not received the message or are not home and able to respond. A few hours after the event begins, the average thermostat setting in Michigan breaks trend and significantly decreases. Despite the Governor's request that households reduce thermostats to 65°F or lower, the average observed smart thermostat setting in Michigan is above 65°F during the entire event. It appears that compliance with the request does not begin until after the Governor's appeal. The differences in furnace fan running time after the request are more difficult to detect visually than thermostat setting and fraction of compliant households.

5 Empirical analysis and results

We compare differences in outcomes between households in Michigan and surrounding states before and after the emergency requests. We consider three outcomes. The household's thermostat setting is a continuous measure of the household's compliance with the emergency request and encompasses the thermal discomfort that the household incurred to contribute to the public good. The second outcome is a binary variable equal to one when the thermostat setting is at or below 65°F. This binary variable captures whether households complied with the request to the letter. The final outcome variable, average number of minutes the fan ran during the hour, is the best proxy available for the amount of energy conserved.

The analysis is divided into four subsections. We begin with a standard difference-in-differences framework to estimate the average treatment effect of the program. We then move to an event-study framework that allows for dynamic effects by sample time period as word of the emergency reached more households. Next, we test whether support for the Governor of Michigan affected compliance rates. Finally, we study how the phrasing of the emergency request around 65°F influenced household behavior.

5.1 Average treatment effect estimates

The first set of regressions we consider are two-way fixed effects specifications on all January 2019 observations. We consider outcomes $Y_{i,t}$ and code a binary variable $D_{i,t} = 1$ for all Michigan observations beginning January 30th at 2:00 pm and zero beforehand. Our preferred specification takes the following form:

$$Y_{i,t} = \alpha_i + \lambda_t + \beta D_{i,t} + \gamma X_{i,t} + \delta_{s,h,d} + \varepsilon_{i,t}, \tag{5}$$

where α_i are household fixed effects, λ_t are time-of-sample indicator variables, $X_{i,t}$ are controls for weather variables (including outside temperature, humidity, wind speed, precipitation, snow depth, and cloud cover), $\delta_{s,h,d}$ are state by day-of-week by time-of-day indicator variables, and $\varepsilon_{i,t}$ is mean-zero heterogeneity. The state by day-of-week by time-of-day indi-

Table 2: Estimates of the regressions from equation 5. The sample includes observations from January 2nd-January 31st.

Two-way fixed effects regressions

VARIABLES	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
Michigan x Post	-1.072**	0.101**	-1.470*
	(0.133)	(0.007)	(0.514)
Constant	67.429**	0.224**	24.820**
	(0.042)	(0.002)	(0.788)
Observations	2,126,336	2,126,336	2,135,114
R-squared	0.710	0.504	0.752
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

cator variables $\delta_{s,h,d}$ allow us to control flexibly for differences in time-varying heterogeneity across states. Equation 5 is a two-way fixed-effects specification. The ordinary-least-squares estimate $\hat{\beta}$ is a difference-in-differences estimate that identifies the causal average treatment effect on the treated under the standard parallel trends, no spillovers, and strict exogeneity assumptions.

Table 2 presents the coefficient estimates of the difference-in-differences regressions for each of the three outcome variables: thermostat setting, a binary variable for setting the thermostat at or below 65°F, and the average number of minutes per hour the furnace fan ran. We cluster the standard errors at the state level.[10] When the thermostat setting is the outcome variable (column 1), the coefficient on $D_{i,t}$ is an estimate of the average treatment effect and is the mean difference in thermostat settings for Michigan and control states before

[10]We cluster at the state level because that is the level of treatment variation (Abadie et al., 2017), but given that there are only five state clusters, it is possible that the cluster-robust standard error estimators will not asymptotically converge (Cameron et al., 2008) and may either overstate or understate the precision of the estimates. As a robustness check, we implement the Donald and Lang (2007) estimator of the average treatment effect in appendix F, which provides valid inference for five clusters and find that the average treatment effects are still statistically significant, so we are not concerned about the precision of our estimates.

and after the treatment. We estimate a reduction of 1.1°F after the emergency request for Michigan households relative to neighbor state households. This reduction is about 0.25 standard deviations in the thermostat setting from January 2 - 29.

Column 2 presents estimates using an indicator variable for having the thermostat at or below 65°F as the outcome variable, which we interpret as full compliance with the request. Given that at any time, some fraction of Michigan households would already have thermostat settings at 65°F or below, the difference-in-differences estimate accounts for this by differencing out the within-household and within-time average incidental compliance. The coefficient on $D_{i,t}$ is an estimate of the additional fraction of households induced to set the thermostat at or below 65°F. We estimate a 10.1 percentage point increase in the fraction of households with thermostat settings at or below 65°F for Michigan households relative to neighbor state households. Using the "constant" term reported in the two-way fixed effects estimates, the expected number of households in Michigan that already would have had thermostat settings less than or equal to 65°F was 22 percent, which we consider the incidental compliance.

Finally, column 3 presents the coefficient estimates of the difference-in-differences regressions using furnace fan run time as the dependent variable, which is the closest proxy to energy consumption in the smart thermostat data. We estimate a 1.5 minute per hour average reduction in furnace fan run time for Michigan households relative to neighbor state households. Relative to the predicted mean furnace fan run time for Michigan households during the emergency period, this is a 6 percent decrease. Given the lack of natural gas consumption data at the household level, this is the best estimate of the amount of natural gas savings caused by the emergency request.[11] Using aggregate daily consumption data provided by Consumers Energy, the total reduction in natural gas consumption from all sources was about 10 percent, which is in line with our estimates.

[11]Natural gas consumption at the household level is measured at the monthly level by the utility.

5.1.1 Robustness checks, alternative specifications, and placebo analysis

In this section and the appendices, we consider alternative behavioral responses (appendix B) and discuss the sensitivity of the main estimates to estimation at the hourly level (appendix C), a series of robustness checks and alternative specifications (appendix D), a placebo analysis of an earlier cold wave with no emergency response (appendix E), and alternative Donald and Lang (2007) inference (appendix F). We find little evidence that people responded to the request via alternative behavioral channels such as turning off the heat or using smart thermostat settings, and we find that the results are generally consistent across the sensitivity tests. As such, the results of these checks are located in the appendices.

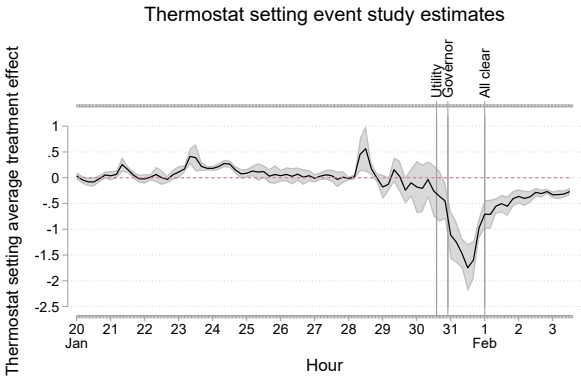
First, we consider the possibility that households responded to the emergency request in ways other than the thermostat setting. For instance, households may have turned off the heat, changed the thermostat mode to "hold" to override previously programmed thermostat setting changes, changed the thermostat to an automated program designed by Ecobee, or may have spent more or less time at home because of the emergency request. The smart thermostat data contains information on whether the furnace was turned off, whether the thermostat was on "hold" (which selects a constant temperature setting), or whether the thermostat was utilizing Ecobee's automation settings which adjust automatically to the household's schedule.[12] In addition, the thermostat contains a motion sensor that registers when there is motion in front of the thermostat, which allows us to test whether households were at home more or less due to the emergency request. We use these variables as outcomes in the main specification as described in equation 5 and display the estimates in appendix section B.

We find that the emergency request induced a 2 percentage point increase in "hold" thermostat settings and a 3 percentage point increase in the use of thermostat automation, which we interpret as small in magnitude. We find no evidence that households turned off the

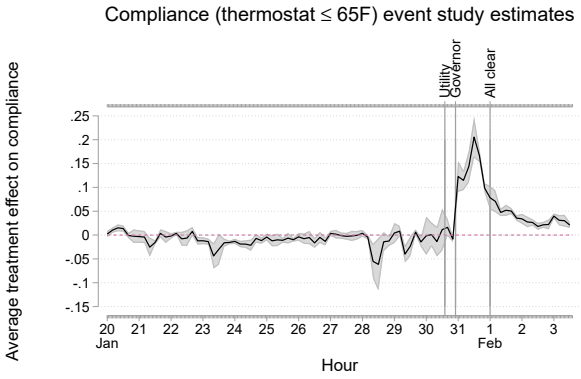
[12]Ecobee's automation includes "smart recovery mode," which will adjust the temperature in anticipation of a user's normal schedule and usual time it takes to heat or cool the home. For instance, if a user arrives home from work at 6 pm and typically increases the thermostat setting upon arrival, the smart recovery program may begin heating the home at 5:45 pm. This is the only automation mode described in the data.

heat or that households were activating the motion sensor more or less during the event. In addition, we include all mode variables and motion sensor variables in a robustness check for the thermostat setting, compliance, and fan regressions, finding that adding these controls does not substantially impact the average treatment effect estimates in table 2. Overall, these results suggest that thermostat setting was the main behavioral channel through which households responded to the request.

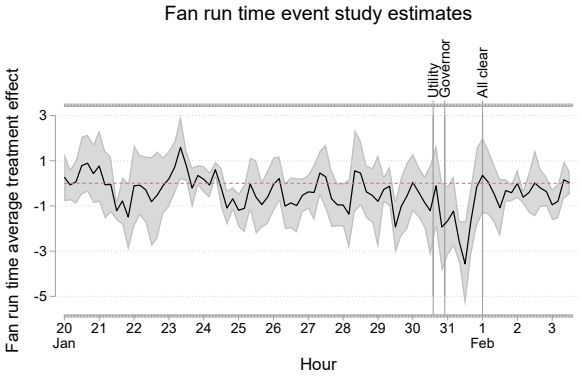
Next, we repeat the analysis using hourly data to test the sensitivity of the results to the choice of time interval in appendix C. We find that the point estimates of our main treatment effects are almost identical, but the estimates on furnace fan run time are less precise and are not statistically different from zero. This is because aggregating to four-hour intervals reduces noise not captured by the fixed effects and controls, improving the precision of the estimates. The robustness checks in section D test the sensitivity of the average treatment effects to alternative difference-in-differences specifications, omitting households who join the sample late or leave early (i.e., using a balanced panel), and allowing for spillovers to counties bordering Michigan. We find that the estimated effects do not change substantially. In section E, the placebo test analyzes a cold wave in Michigan that occurred ten days earlier on January 20-21, 2019, where temperatures dropped by a similar magnitude. We find that Michigan's heating behavior remains parallel to the control households during this placebo event, and that estimation using regression equation 5 with the placebo treatment yields estimates of zero, suggesting that our findings are not an artifact of differential responses by Michigan households to cold waves. In our regular specifications, this cold wave is included in the data and thus serves as a control, lending credibility to the research design. Finally in section F, we calculate the Donald and Lang (2007) estimate of the treatment effect (which has valid inference for five clusters), and find comparable and statistically significant estimates using this approach.



(a)



(b)



(c)

Figure 5: Event-study coefficients estimated using regression equation 6 with (a) thermostat setting, (b) compliance, and (c) minutes of furnace fan run time as the dependent variables. 95 percent confidence intervals constructed from standard errors cluster-robust to heteroskedasticity.

5.2 Event-study estimates

Given the repeated requests for reductions in thermostat settings over time, we next account for a dynamic response in an event-study framework. We estimate a two-way fixed effects regression using the following specification:

$$Y_{i,t} = \alpha_i + \lambda_t + \sum_{k=-57}^{-1} \beta_k^{lead} \mathbf{1}[k = t - g] + \sum_{k=0}^{29} \beta_k^{lag} \mathbf{1}[k = t - g] + \gamma X_{i,t} + \delta_{s,h,d} + \varepsilon_{i,t}, \quad (6)$$

where g is the time period the utility made its first emergency request to households. Thus, we estimate 57 lead coefficients and 30 lag coefficients to include 10 and a half days of pre-trends and four days of dynamic treatment effects (including two days after the treatment ends).[13] We hypothesize that prior to the end of working hours, household responses will be muted and that the largest responses will occur after the Governor's use of the emergency text message alert at 10:30 pm. Further, we suspect that the treatment effect persisted after the "all clear" time due to barriers to receiving the all-clear message or adjusting the thermostat (e.g., if the home is vacant or all occupants are sleeping).

Figure 5 plots the dynamic treatment effects estimated using the event-study regressions specified in equation 6 using thermostat setting, compliance with the request, and fan run time as outcome variables. Leading up to the emergency, the difference in thermostat settings between Michigan and the control states is typically small and positive when the confidence interval does not overlap zero. To the extent that the pre-trends are not parallel, we expect that our event-study estimates may underestimate the treatment effect on thermostat setting during some time periods, though this effect is likely to be small given the size of the estimated treatment effect relative to the pre-treatment noise. We see similar results for the compliance estimates and note that the compliance estimates may also be conservative. The pre-treatment trends are more noisy for fan run time, but do not display systematic trends,

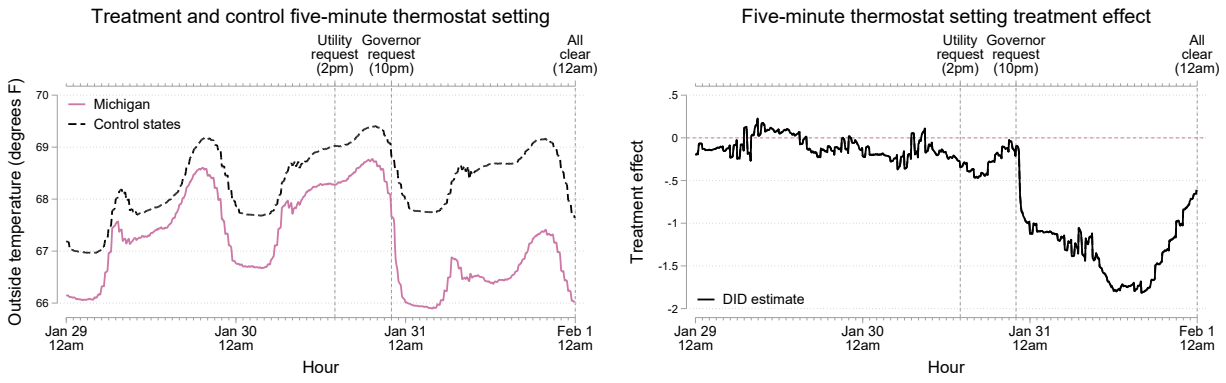
[13] Choosing greater or fewer leads and lags does not substantially change the coefficient estimates, but increases computational cost. We chose the window to allow us to test for the presence of pre-trends, observe when thermostat settings returned to a normal level after the event, and to keep computational times reasonable.

and the confidence interval contains zero for most pre-period estimates.

The first finding of note is that the Governor's alert was essential to increasing compliance. Averaging over the event-study coefficients for the eight hours prior to the Governor's alert, the utility's emergency request only resulted in an average additional compliance rate of 0.4 percent, resulting in an average thermostat reduction of just 0.4°F. Following the Governor's alert, the average additional compliance rate was 14.2 percent (peaking at over 20 percent), resulting in an average thermostat reduction of 1.4°F (with a peak of 1.7°F). Given the utility's actions of sending emails to customers, posting on social media, and reaching out to traditional news media, we do not think the lack of responsiveness was solely due to a lack of reach. While it is possible that the initial lukewarm response was caused by households not being at home to change the thermostat setting, we see this as unlikely because the Ecobee smart thermostat setting can be changed remotely via app. Instead, it is likely that households did not take the request seriously until it became clear that there was a true emergency. The additional authority of the Governor and the repeated request to reduce thermostat settings likely increased the salience of the request, inducing additional compliance.

After the emergency request and before the "all clear," thermostat settings begin to trend upward and compliance begins to fall. Sustained compliance is likely increasingly costly, so participation rates decline toward the end of the event. This trend may also be due to households choosing low thermostat settings when sleeping and leaving the home for work in the morning. Upon returning from work, households may increase the thermostat setting to a slightly higher level. This behavior is similar to the "backsliding" dynamic reported by Allcott and Rogers (2014) in which households conserve electricity after receiving a home energy report, but the effect lessens over time. To sustain high levels of compliance in an emergency, repeated requests are likely necessary.

Another interesting finding is that the effect persists after the "all clear." This persistence suggests that some households which had programmed their thermostats to reduce the temperature setting in keeping with the emergency request had not yet re-programmed



(a) (b)

Figure 6: (a): Five-minute sample mean thermostat settings for Michigan and control households from January 30th 12 pm - January 31st 11:59 pm. (b): Five-minute difference-in-differences estimate.

them after the all clear due to the fixed cost of interacting with the thermostat. This result is consistent with previous work that finds that changes in thermostat settings in response to a cold or hot period tend to persist after the cold or hot period ends (Ge and Ho, 2019).

In addition, one can see that the reductions in furnace fan running time are only transitory. Because furnaces essentially run at one speed, reducing the thermostat at night or when out of the home will reduce energy use while the home cools, but upon increasing the thermostat, the furnace will need to run again and incur a ramp-up cost to increase the temperature. We can see in the fan run time event study estimates that on January 31st, there were savings during the early morning and day, but when households returned home in the evening that furnaces had to run at essentially full intensity to warm the home again. After the all clear, the estimates return to mean zero more quickly than the thermostat setting and compliance rate estimates.

In appendix section C, we replicate the event-study analysis at the hourly level. We find similar results across all three variables, and one can see that the treatment effect begins the hour of the Governor's request, providing additional evidence that the Governor's request was key to increasing the strength and salience of the nudge. The main difference is that the pre-treatment coefficients are more noisy, which leads us to favor the four-hour analysis.

We supplement the event study with a graphical analysis of the five-minute thermostat-setting data. In figure 6a, we plot five-minute thermostat setting data for Michigan and the control states from January 29th through February 1st. In figure 6b, we plot a difference-in-differences estimate of the treatment effect, which we construct as the average difference between Michigan and the control states in five-minute thermostat settings less same day-of-week and hour-of-day thermostat settings from before the event. In these figures, one can only see a clear decline in Michigan thermostat settings after the Governor's request. In appendix section E.2, we replicate the five-minute analysis during the placebo cold wave. The difference-in-differences estimates are zero throughout most of the placebo period other than a slight increase in thermostat settings for Michigan after the placebo all clear time, lending credibility to the difference-in-differences estimates in figure 6b.

5.3 Heterogeneity analysis

Next, we analyze support for the Michigan Governor and the effect of the reference point on household behavior in a triple-differences framework. Our approach analyzes both forms of heterogeneity in the same estimating equation with additional controls to account for the possibility that county-level support for the Governor is correlated with baseline thermostat setting or other demographic factors. We supplement the smart thermostat data with data on gubernatorial election county vote shares for each state's most recent election obtained from the Voting and Elections Collection maintained by the CQ Press (2019). We include household-level controls available in the Ecobee data as well as demographic controls at the county level obtained from the American Community Survey (US Census Bureau, 2019). Below, we discuss how we define the baseline thermostat setting and support for the Governor separately before presenting a single regression equation where we estimate the effect of baseline thermostat setting and support for the Governor on the treatment effect. We estimate the effects in the same regression in case homes in Republican-voting counties have differing baseline thermostat settings than homes in Democrat-voting counties (or vice versa).

Households whose thermostats would have been at 65°F or lower essentially received information that they were already keeping the thermostat low enough and may have felt that they did not need to reduce the thermostat further.[14] Furthermore, we hypothesize that the distance from the reference point may also affect household behavior as outlined in the theoretical framework in section 3. To test our hypotheses, we estimate the effect of the emergency request allowing for different responses by expected baseline thermostat settings. We construct a non-parametric estimate of baseline expected thermostat setting $\hat{T}_{i,t}$ for each household by calculating the household's sample average thermostat setting for each day-of-week and time-of-day combination from the pre-treatment period. Denote $\mathcal{T} = \{[0, 59), [59, 61), [61, 63), ..., [73, 75), [75, 100]\}$ as the collection of 2-degree intervals from 59°F to 75°F with binned endpoints for higher and lower temperatures, and $b \in \mathcal{T}$ the interval with upper bound b. We interact indicator variables for belonging in each interval with the treatment variable to create a third difference and estimate heterogeneous effects by baseline temperature category.

In the same regression, we analyze heterogeneity by political support for the Governor. The data on gubernatorial election returns is at the county level. In the 2018 election, the distribution of the Michigan Governor's county vote share ranged from 31 percent to 73 percent. We expect the effect of political support to be non-linear so we create 5-percentile indicator variables between 30 and 75 percent. Denote $\mathcal{P} = \{[30, 40), [40, 45), [45, 50), ..., [70, 75]\}$ as the collection of a 10-percentile interval between 30 and 40 and 5-percentile intervals between 40 and 75, and $a \in \mathcal{P}$ the interval with upper bound a.[15] Similarly to the baseline thermostat setting, we interact indicator variables for belonging in each interval with the treatment variable.

[14] This analysis generally treats larger reductions as welfare-improving, but we note that thermostat settings that are too low increase the risk of frozen pipes.

[15] We combined the 30-35 and 35-40 percentile intervals because only 0.38 percent of households lived in counties with a Democratic Party vote share between 30 and 35 percent, which lead to extremely imprecise estimates.

Thus, our estimating equation is

$$\begin{aligned} Y_{i,t} = \; & \alpha_i + \lambda_t + \sum_{b \in \mathcal{T}} \beta_b D_{i,t} \times \mathbf{1}[\hat{T}_{i,t} \in b] + \phi_b \mathbf{1}[\hat{T}_{i,t} \in b] + \sum_{a \in \mathcal{P}} \beta_a D_{i,t} \times \mathbf{1}[P_{county} \in a] \\ & + \gamma_1 D_{i,t} \times Z_i + \gamma_2 X_{i,t} + \delta_{s,h,d} + \varepsilon_{i,t}, \end{aligned} \tag{7}$$

where Z_i is a vector of controls for household-level characteristics available in the smart thermostat data as well as county-level demographics to address correlation between county vote share and demographics.[16] Given our hypotheses, we expect compliance to fall with an increased baseline expected thermostat setting. Thus, we expect β_b to be lower for higher levels of b. Our theory predicts that the treatment effect will increase with a higher baseline expected thermostat setting. Moreover, it is possible that very cold baseline households may increase thermostat setting when introduced to the reference level of 65, thus we expect β_b to be zero or positive for $b \leq 65$.[17] For the thermostat setting and compliance outcome variables, we hypothesize that the coefficients on the interaction with vote share β_a will be increasing as Democratic vote share increases and that the opposite will be true for the fan running outcome variable regression, indicating that the appeal was more effective for households in counties that supported the Governor's election.

Table 3 displays the estimates of equation 7 for each outcome variable. The standard errors are cluster-bootstrapped to incorporate the uncertainty due to sampling error from estimating the baseline thermostat setting.[18] We discuss the vote share and baseline thermostat setting results separately in the following sections.

[16] From the smart thermostat data, we include square feet of the home, number of occupants, whether the home is detached or an apartment, and the age of the home. From the ACS, we include county level median income, median age, population, fraction male, fraction white, fraction with a high school education or more, and fraction of non-US citizen residents.

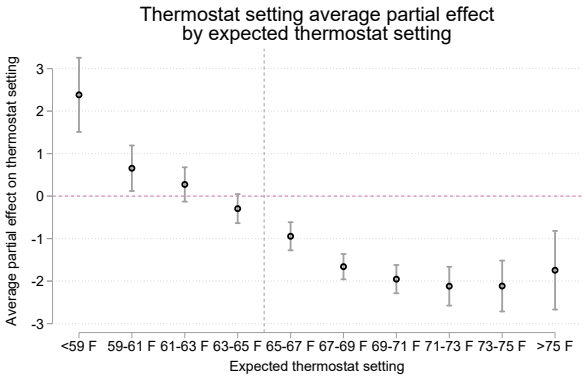
[17] This backfire is called the "boomerang effect" in the social-comparison literature where if a household receives information that they are consuming less than the average they may decide they were being too conservative and increase consumption (Allcott, 2011).

[18] The bootstrap procedure first draws observations with replacement from within household, day of week, and time period strata to estimate 100 different baseline temperatures for each household, day of week, and time of day combination. It then samples from this empirical distribution and draws 100 bootstrap samples clustered at the city level. Ultimately, these standard errors differ very little from clustered standard errors that ignore the uncertainty from the first-stage estimation.

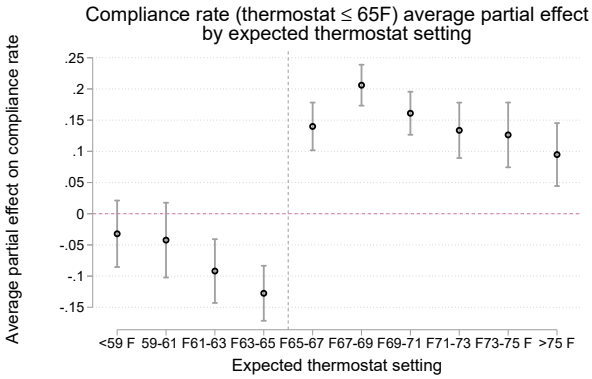
Table 3: Results from the estimation of equation 7. Standard errors cluster-bootstrapped to incorporate the sampling error from estimation of the baseline thermostat setting. The sample includes observations from January 2nd-January 31st.

	(1)	(2)	(3)
	Thermostat setting	Thermostat $\leq$ 65F	Fan run time
59 F or lower expected X Treatment	2.68**	0.095**	1.32
	(0.45)	(0.024)	(0.83)
59-61 F expected X Treatment	0.95**	0.085**	0.57
	(0.21)	(0.023)	(0.64)
61-63 F expected X Treatment	0.57**	0.036	0.32
	(0.13)	(0.020)	(0.47)
65-67 F expected X Treatment	-0.65**	0.27**	-0.70*
	(0.082)	(0.015)	(0.31)
67-69 F expected X Treatment	-1.37**	0.33**	-1.15**
	(0.096)	(0.018)	(0.31)
69-71 F expected X Treatment	-1.66**	0.29**	-1.45**
	(0.099)	(0.017)	(0.33)
71-73 F expected X Treatment	-1.82**	0.26**	-1.92**
	(0.19)	(0.021)	(0.55)
73-75 F expected X Treatment	-1.82**	0.25**	-2.23*
	(0.25)	(0.022)	(0.90)
Higher than 75 F expected X Treatment	-1.45**	0.22**	-2.31
	(0.43)	(0.022)	(1.40)
40-45% Democrat X Treatment	-0.44	0.077**	0.19
	(0.29)	(0.026)	(0.80)
45-50% Democrat X Treatment	-0.14	0.036	1.23
	(0.25)	(0.025)	(0.91)
50-55% Democrat X Treatment	-0.34	0.048*	-0.38
	(0.21)	(0.023)	(1.08)
55-60% Democrat X Treatment	-0.28	0.051	2.10
	(0.36)	(0.038)	(1.61)
60-65% Democrat X Treatment	-0.57	0.095*	1.59
	(0.40)	(0.046)	(2.12)
65-70% Democrat X Treatment	-0.51	0.078	-0.43
	(0.38)	(0.051)	(1.81)
70-75% Democrat X Treatment	-0.87*	0.091	2.18
	(0.41)	(0.047)	(2.09)
Observations	1,959,762	1,959,762	1,960,072
R-squared	0.81	0.69	0.76
FE	YES	YES	YES
Hour	YES	YES	YES
Controls	YES	YES	YES
Expected thermostat level	YES	YES	YES

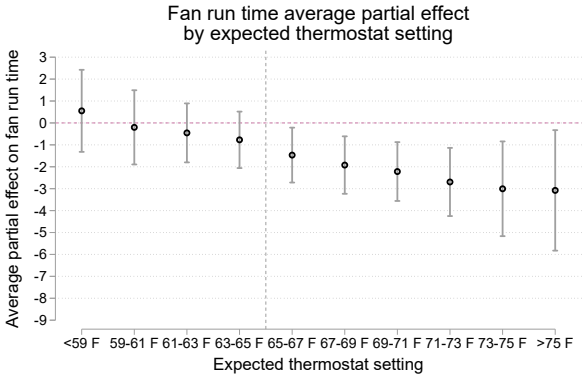
Standard errors cluster-bootstrapped at the city level.
** p<0.01, * p<0.05



(a)



(b)



(c)

Figure 7: Effect of the emergency request by expected thermostat setting estimated using equation 7 with (a) thermostat setting, (b) compliance, and (c) minutes of furnace fan run time as the dependent variables. 95 percent confidence intervals constructed from bootstrapped standard errors that account for first-stage estimation of expected thermostat setting and are cluster-robust to heteroskedasticity.

5.3.1 Reference point effect

The coefficients on the interaction between expected thermostat setting and treatment in table 3 are the difference in the average treatment effect by expected thermostat setting relative to households in time periods with an expected thermostat setting of 63-65°F (the omitted category). We find that households in time periods with baseline thermostat settings at or below 65°F were less responsive to the appeal or even increased their thermostat setting. As the baseline thermostat setting increases above 65°F, the magnitude of the response increases and then decreases at higher temperatures. Compliance with the emergency request falls as the baseline thermostat setting increases above the requested level and then decreases for baseline temperatures 69°F and above. Households with baseline thermostat settings just below 65°F were more likely to increase the thermostat setting to above 65°F, coming out of compliance. The estimates for the fan run time variable show that for the coldest baseline

thermostat setting, fan running times increased relative to the base category. For baseline thermostat settings above 65°F, fan run time decreased relative to the base category, with the largest effect (though not statistically significant) for the highest baseline thermostat settings.

One concern we had was whether these estimates were an artifact of statistical mean reversion rather than a meaningful pattern.[19] To test this alternative hypothesis, we estimate equation 7 using the placebo cold wave event. Appendix section E.3 displays the results of the placebo analysis. The estimates from the placebo analysis are the same sign as the estimates from the polar vortex, but the magnitude of the estimated coefficients in the placebo analysis are substantially smaller than during the polar vortex. Furthermore, the placebo estimates are relatively flat as the baseline category gets further away from 65°F. Thus, we conclude that mean reversion may play a small role in the main heterogeneity estimates but the effect is not large enough to alter our conclusions in this section. As an additional robustness check, we replicate the analysis on hourly data in appendix C and find similar results to our primary analysis.

Figure 7 displays the average partial effects of treatment on the outcome variables by baseline thermostat setting.[20] When the expected thermostat setting is less than 65°F, the appeal does not decrease the thermostat setting and is likely to increase the thermostat setting in the coldest homes. On average, when the expected thermostat setting is below 65°F, the appeal corresponds with households increasing the thermostat above 65°F. These perverse effects are consistent with the appeal anchoring low thermostat settings to the norm of 65°F. Alternatively, the appeal may have increased the sense of danger, causing households to either store heat in their home in case of an outage or to stay home when they typically would have left for work and reduced the thermostat setting.

Households with high baseline thermostat settings were also less likely to fully comply

[19]That is, do our estimates merely reflect that households with high or low temperatures in the past are mechanically more likely to have average temperatures when measured later?

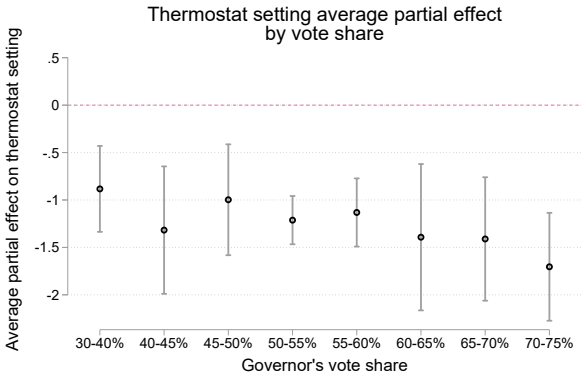
[20]Formally, the average partial effect of treatment by baseline thermostat setting $\hat{T}_{i,t} \in b$ is $E[Y_{i,t}|W_{i,t}, \hat{T}_{i,t} \in b, D_{i,t} = 1] - E[Y_{i,t}|W_{i,t}, \hat{T}_{i,t} \in b, D_{i,t} = 0]$, where $W_{i,t}$ is a vector of all other control variables in equation 7.

with the reference level of 65°F, which may be caused by several mechanisms with different implications. One potential mechanism is that the request may have appeared out of reach, which suggests that a reference level can induce larger contributions of effort for households near the reference level, but it discourages effort for households far away from that reference level. For households with expected thermostat settings above 75°F, the average partial effect on compliance rate is the lowest of those above the reference level; however, the average partial effect on fan run time was the second largest, and in general higher baseline thermostat settings resulted in larger energy savings. This suggests that those households that did comply generated substantial energy savings, although the heterogeneity of this effect leads to a wide confidence interval. Another potential explanation may be that households that prefer warm temperatures have stronger preferences for deviating from their preferred thermostat setting. We see this as unlikely, given the large average reductions in thermostat settings upon the request but cannot rule it out. Without variation in the reference level, we are cautious not to draw more firm conclusions.

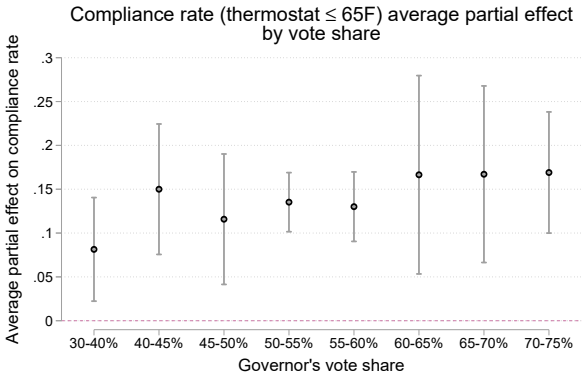
5.3.2 Vote share effect

The coefficients on the interaction between county vote share and treatment in table 3 are the difference in the average treatment effect for households in counties with a given level of support for the Governor relative to households in counties where the Governor's vote share was 30-40% (the omitted category).[21] The estimates show that the average thermostat reduction and compliance rate is higher in all counties relative to those with the lowest Governor's vote share. In addition, the treatment effects are generally increasing in magnitude as the Governor's vote share increases. The results indicate that households in the counties that supported the Governor the most reduced thermostats by up to 0.9°F more on average and had a compliance rate 9 percentage points higher relative to the least supportive county (although the difference in compliance rate was not statistically different from zero). Despite this, the estimates of the effect on fan run time are the opposite of the

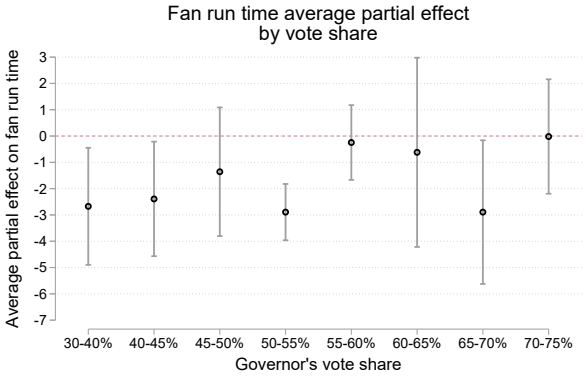
[21] Appendix section 5.3 replicates this analysis using the placebo cold wave and does not find the same patterns in the estimated coefficients.



(a)



(b)



(c)

Figure 8: Effect of the emergency request by Governor's vote share estimated using equation 7 with (a) thermostat setting, (b) compliance, and (c) minutes of furnace fan run time as the dependent variables. 95 percent confidence intervals constructed from standard errors cluster-robust to heteroskedasticity.

expected sign and are imprecise with all confidence intervals containing zero. Given that fan run time is a function of underlying energy efficiency of the furnace and home, we believe that the fan estimates reflect unobserved differences in energy efficiency that are correlated with political affiliation.

Figure 8 displays average partial effects of treatment on the outcome variables by Governor's vote share.[22] The average partial effect on thermostat setting and compliance is increasing in the Governor's vote share. On average, the appeal induced about a 7 percent compliance rate in the counties most opposed to the Governor and a 16 percent compliance rate in the counties most in support of the Governor (difference = 9.1%, p-value = 0.053). In the counties most opposed to the Governor, the appeal induced a 0.9°F decrease in thermostat settings while the most favorable counties saw a 1.7°F decrease (difference = 0.9°F, p-value < 0.05). While this effect is large (the same size as the average treatment effect), it does not outweigh the increased compliance rates generated by the Governor's amplification of the appeal on social media and via the emergency text alert system, without which there may have been no response from households at all.

Thus, we conclude that support for the Governor is correlated with a stronger response to the public appeal, although the implications for energy use are unclear. We caution against interpreting these estimates causally, but our findings are consistent with distrust arising out of affective political polarization. In the increasingly polarized political environment of the United States (Iyengar et al., 2019), a public appeal may be met with defiance rather than compliance. An alternative explanation is that political ideology may be correlated with willingness to contribute to public goods or thermostat setting behavior more broadly. Given our inability to distinguish the effect of polarization from political ideology, we cannot reject either explanation.

[22]The average partial effect of treatment by baseline county-level vote share $P_{county} \in a$ is $E[Y_{i,t}|W_{i,t}, P_{county} \in a, D_{i,t} = 1] - E[Y_{i,t}|W_{i,t}, P_{county} \in a, D_{i,t} = 0]$, where $W_{i,t}$ is a vector of all other control variables in equation 7.

6 External validity

Here, we discuss how likely our results are to generalize to non-smart thermostat households and to future energy crises. Our research design takes advantage of the exceptional nature of the emergency event and availability of the smart thermostat data during this period to construct well-identified estimates of the effect of the emergency request on behavior, but it may be unclear to what extent these unique data and event are representative of other energy users in other events. To orient our discussion, we make use of the "SANS" conditions for generalizability suggested by List (2020). In the SANS framework, the external validity of a study can be assessed by discussing selection, attrition, naturalness, and scaling. In our context, there was very little attrition from the sample (we assess the sensitivity of the results to keeping and including those who enter or leave the sample early in appendix D), which leaves selection, naturalness, and scaling for discussion. In our discussion, we find little evidence that selection plays a role in our results. Given that the emergency request is a natural experiment and that similar emergency requests are made during other energy emergencies, we believe the intervention is natural and likely to be representative of interventions in other energy contexts. Finally, we discuss the ability of this intervention to scale vertically to the grid (ISO) level, and horizontally across states and other energy emergencies. We believe our results are likely to generalize to other energy emergencies based on the availability of a wireless emergency alert system, the political climate, and the frequency of repeated requests which may result in habituation to the alerts.

6.1 Selection

Our primary concern about generalizability is the selection into owning an Ecobee thermostat and sharing data used for this study. As other studies have noted (Burkhardt et al., 2019; Blonz et al., 2021), most papers in the residential energy literature rely on data from self-selected households—this paper is no different. Given that our treatment is unrelated to selection, selection into the sample does not affect the internal validity of the research design,

Table 4: Estimates of equation 5 by smart thermostat adoption date. The sample includes observations from January 2nd-January 31st.

Thermostat setting estimates by smart thermostat adoption date

	(1)	(2)	(3)	(4)
VARIABLES	Pre-2016 adopters	2016 adopters	2017 adopters	2018 adopters
Michigan x Post	-1.152**	-1.223**	-1.027**	-1.081**
	(0.111)	(0.107)	(0.146)	(0.151)
Constant	66.865**	67.434**	67.433**	67.544**
	(0.106)	(0.059)	(0.051)	(0.062)
Observations	163,033	372,872	714,402	846,655
R-squared	0.696	0.701	0.717	0.710
Weather controls		YES	YES	YES
Household FE			YES	YES
Time FE				YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

but may be relevant to the external validity of the estimates. Previous work has found that, on observable characteristics, the Ecobee Donate-Your-Data households are comparable to the average household in the nationally representative Residential Energy Consumption Survey sample, though the Ecobee households have slightly more members (Meier et al., 2019). Another study argues that the Ecobee smart thermostat features do not differ substantially from other smart thermostats (Blonz et al., 2021). Despite being comparable on observable characteristics, there remains a concern that households that choose to share data may be more willing to contribute to other public goods such as compliance during an emergency. Furthermore, the Ecobee is marketed as being eco-friendly, which may also be correlated with pro-social attitudes. Finally, smart thermostat features differ from traditional thermostats. The largest and most relevant difference is that smart thermostat users can adjust the thermostat remotely via app, which reduces the cost of compliance. Brandon et al. (2021) find little evidence that smart thermostats on their own cause households to consume energy differently relative to households with conventional thermostats, but the concern remains that these households may find it easier to comply with the request. If these hypotheses

about selection are correct, our estimates from smart-thermostat household behavior may overstate compliance relative to non-smart-thermostat households.

We assess the degree to which selection might affect our estimates in two ways. First, we compare our average treatment effect to estimates of the reduction in residential natural gas consumption calculated using aggregate data provided by Consumers Energy, and second we consider differential responses to the emergency request by early adopters of the Ecobee thermostat. During the event, forecasted natural gas demand using realized weather conditions was 3.3 billion cubic feet on January 30th and 2.9 billion cubic feet on January 31st. After all reductions in consumption were accounted for, actual consumption was 3.0 billion cubic feet on January 30th and 2.6 billion cubic feet on January 31st, implying a 10.7% and 10.5% reduction in daily consumption from all sources (residential and non-residential). In a mostly-residential area of Detroit, Consumers Energy claimed to see a 10% reduction in energy consumption. Using furnace fan run time as a proxy variable for energy consumption, we estimate a 6% reduction. This estimate is similar but smaller in magnitude, which is the opposite of what we would expect if selection was substantially affecting the estimates.

The second test of selection examines whether early adopters of the smart thermostats responded differently to the request than late adopters. In the smart thermostat data, we observe the date the smart thermostat account became active. We hypothesize that early adopters of smart thermostats are more highly selected relative to the general population, whereas late adopters are more representative. If early adopters have differential responses to the emergency request, this evidence would suggest that smart thermostat users are not representative of the general population.

Table 4 displays the results of estimating the main specification from equation 5 on samples segmented by smart thermostat adoption year using thermostat setting as the outcome variable.[23] The estimates are similar across adoption years, and none of the differences are statistically different from zero. Thus, to the extent that early adopters are more highly

[23]We pool households who adopted in 2016 or earlier, and we do not include households that adopted the smart thermostat in January 2019, as any differences may be attributed to lack of experience with the thermostat.

selected, the treatment effect does not vary by that selection.

Overall, our diagnostics suggest a limited role for selection. Hypothetically, all of the selection effects point to a potentially larger responsiveness of households in our data to the requests, but we do not find any evidence that suggests our estimates will not generalize out of sample.

6.2 Naturalness

Given that this is a natural experiment, the treatment subjects are subjected to a natural source of variation in the emergency request. Moreover, emergency appeals with requested compliance levels are a standard response in the toolkit of public utilities responding to emergency shortages. For example, California's Flex Alert system regularly includes a thermostat setting target of 78°F.[24] Since this polar vortex in 2019, thermostat-targeted emergency appeals have been used in Texas and California in response to extreme heat events in 2021 and 2022.[25] Emergency appeals of this form are used widely, though future research should analyze the effect of variation in the reference level and requests for fixed reductions from the baseline (e.g., "please reduce thermostat settings by 5°F").

6.3 Scaling

The Michigan emergency appeal was moderately successful at the state level, suggesting that similar appeals can be successful at a large scale. Here, we discuss which policy design elements are necessary for similar emergency appeals to scale vertically to the grid (ISO) level, horizontally from state to state, and to future emergencies. Our analysis focuses on the availability and use of the wireless emergency alert system, the political climate, and the frequency of requests.

Our event-study analyses in section 5.2 suggest that the Governor's messaging and use

[24] See https://flexalert.org/.

[25] Texas ERCOT request to increase thermostats to 78°F: https://www.ercot.com/news/release?id=8b772e9e-51d0-4c3c-e653-1e5079f28e89. California Governor request to increase thermostats to 78°F: https://twitter.com/CAgovernor/status/1567316274849660928.

of the wireless emergency alert system were critical to achieving compliance. Previous studies of requests for voluntary reductions have demonstrated low compliance rates when the request is made only by the utility or by local news media (Holladay et al., 2015). Thus, a wide-reaching message from an authority figure is crucial for scaling this effect. In other contexts, these appeals will be limited by the reach of the emergency communications systems. Areas with low media and technology penetration or with limited cell phone service may see limited success of emergency appeals. Similarly, the size of the treatment effect peaked mid-day, after people had time to respond, suggesting that the timing of the messaging is pivotal in future emergencies. ISOs spanning multiple states should have emergency communications relationships with state emergency agencies to gain access to key communications infrastructure.

Our estimates suggest that political polarization may undermine the authority of a leader's request for emergency compliance. A request from a popular official or institution may receive more favorable responses than an unpopular official or institution. Given that the Governor of Michigan had only just assumed office after winning the election with 53% of the vote and a 10% margin of victory, we believe that the effect of political affiliation could be much stronger in other contexts.

Finally, we caution that similar emergency appeals may not scale when made repeatedly. Ito et al. (2018) find that repeated conservation nudges result in habituation or desensitization, reducing their effectiveness over time. The Michigan polar vortex appeal was a unique emergency and to our knowledge was the only such appeal in recent years. Thus, the stimulus was novel, likely increasing the salience of the request. In states such as Texas and California where requests are relatively commonplace, habituation may decrease the request's effectiveness.

7 Conclusion

This paper studies an acute natural gas shortage during the 2019 polar vortex in Michigan. During near-record- low temperatures, a fire at a compressor plant resulted in natural gas demand that nearly outpaced supply. In response, the utility requested that households voluntarily conserve natural gas, and the Governor of Michigan subsequently issued an emergency text alert that requested households voluntarily reduce thermostat settings to 65°F.

We use smart thermostat data to analyze consumer responses to the emergency request, finding robust evidence of voluntary compliance with the request. Using a difference-in-differences strategy with four control states, we obtain estimates of the average treatment effect. On average, households lowered their thermostats by 1.1°F, roughly a 25% reduction of the typical variation in the average thermostat setting. 10 percent of households complied with the request fully by reducing their thermostats to 65°F or lower, while 22 percent of household thermostat settings were already at or below the threshold. Finally, we find evidence that the emergency request reduced furnace fan run times (our best proxy for natural gas consumption) by 1.5 minutes per hour; a 6% decrease. This reduction is smaller than the 10% reduction in all consumption of natural gas we calculate using aggregate consumption data provided by the utility. Our estimate is comparable with reductions in energy consumption observed in field experiments that use moral suasion to induce conservation (see e.g., Brandon et al. (2018)), but falls short of field experiments that use price incentives to induce conservation (see e.g., Ito et al. (2018)).

Our analysis highlights the importance of wide-reaching emergency messaging for governments and utilities. An event study analysis reveals that prior to the Governor's announcement, the utility's emergency request only induced 0.4 percent of households to reduce thermostat settings to 65°F or less. After the Governor's amplification of the emergency request, the fraction of households in compliance with the request increased to 20 percent at its height. The emergency directives communicated via social media and news media suffered from low visibility and were likely lost in the large amount of other content on

these platforms. As the emergency progressed, compliance with the request waned as households likely found it increasingly costly to maintain low thermostat settings. In addition, we find evidence of persistence in low thermostat settings in the day after the emergency. These habits appeared to be driven by programmed thermostat settings left in place by households.

The particular phrasing of the emergency request around the reference point of 65°F played a large role in determining household behavior. We identify three perverse effects of the reference point. First, households that typically heat at 65°F or lower did not reduce thermostats in response to the emergency request. Second, those that typically have the lowest thermostat settings even increased the thermostat setting after receiving the request. Third, those with the highest thermostat settings were less likely to comply with the request. For households with thermostat settings typically above 65°F, the average treatment effect at first increases and then decreases with distance from the reference point. These findings are consistent with our theoretical model, which suggests that a nudge with a compliance target will not achieve the least-cost reduction in energy consumption because of the difference in marginal incentives on either side of the reference point created by the target.

Setting a more aggressive target trades off a larger effect of compliance with the cost of compliance, which increases defiance. This suggests that a particular reference point may induce a larger response. While this natural experiment did not provide the necessary variation to determine the highest-impact reference points in this context, these could be determined via purposeful experimentation. Furthermore, smart thermostat technology offers the possibility for targeted requests and automated compliance. Alternatively, requests for a uniform reduction in thermostat setting may induce broader compliance than a uniform compliance goal because they avoid the problems driven by users in the tails of the energy consumption distribution.

Political affiliation also correlated with compliance. Households in counties that supported the Governor the least in the 2018 election responded the least to the request, and households in counties that supported the Governor the most had high levels of compliance. Thus, it appears that political polarization can lead to defiance from outsider groups; how-

ever, this does not outweigh the benefits of the Governor's amplification of the emergency appeal via social media and the text alert system.

Ultimately, the 2019 Michigan polar vortex crisis was resolved by a combination of residential and non-residential demand reductions and supply side efforts. After the crisis, the Governor of Michigan issued an executive order transferring energy emergency response management from the Michigan Energy Agency to the Michigan Public Service Commission (MPSC) and commissioned an assessment of Michigan energy resources from the MPSC (MPSC, 2019a). The report includes an overview of energy supply systems for natural gas, electricity, and propane, as well as a section on energy emergency management. The section on emergency management states in general terms that utilities can pursue a variety of curtailment strategies to reduce demand, including voluntary requests and rate increases, but the guidelines are vague.

This paper shows that emergency demand response programs can help provide stability during times of crisis, but the efficacy of the program depends heavily on its design and implementation. While the emergency request in Michigan was largely successful and the worst-possible outcome was averted, the low overall level of compliance and perverse reference-point effects highlight the need for well-designed emergency measures to reduce energy demand. To be successful, a voluntary emergency demand-response program needs a communication platform that enables it to reach households, can induce compliance from households that receive the request, and that has a demonstrated effectiveness so that utilities and balancing authorities know the size of the demand reductions to expect. Rather than relying on voluntary requests with unknown efficacy, utilities should develop, test, and optimize programs that can be called upon when needed.

To avoid compliance issues, utilities can invest in voluntary programs that eliminate compliance barriers by purchasing centralized control of energy consumption long before emergency events occur. Interruptible-load demand response programs are not new, but applications involving smart thermostats may provide a new opportunity to enhance emergency management. For households who do not wish to surrender control or who have conventional

thermostats, incentive-based emergency curtailment program design remains critical.

We see several results as likely to generalize beyond energy consumption. First, compliance with requests is likely to be higher when the messenger is a trusted public figure. This effect may be reduced by political polarization or distrust of institutions. Second, a low-cost method of widespread emergency notification such as the cell phone alert system is key for communicating timely requests during a crisis. Emergency communication via social media is likely to suffer from low reach and must compete with other content for visibility. The incentives for compliance also matter. Requests for uniform compliance goals with targets are likely to be too ambitious for some and too conservative for others. Instead, a simple request for a marginal contribution to the public good or a menu of marginal contributions that can be dialed up or down avoids this problem without the need to explicitly tailor requests. Finally, this event highlights the need for testing emergency planning and incorporating design elements that explicitly consider economic incentives and behavioral responses.

References

Abadie, A., S. Athey, G. Imbens, and J. Wooldridge (2017). When should you adjust standard errors for clustering? NBER working paper 24003, National Bureau of Economic Research.

Allcott, H. (2011). Social norms and energy conservation. *Journal of Public Economics 95*(9), 1082–1095. Special Issue: The Role of Firms in Tax Systems.

Allcott, H., L. Boxell, J. Conway, M. Gentzkow, M. Thaler, and D. Yang (2020). Polarization and public health: Partisan differences in social distancing during the coronavirus pandemic. NBER working paper 26946, National Bureau of Economic Research.

Allcott, H. and T. Rogers (2014, October). The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation. *American Economic Review 104*(10), 3003–37.

Allcott, H. and D. Taubinsky (2015, August). Evaluating behaviorally motivated policy: Experimental evidence from the lightbulb market. *American Economic Review 105*(8), 2501–38.

Anderson, S. T. and J. M. Sallee (2011, June). Using loopholes to reveal the marginal cost of regulation: The case of fuel-economy standards. *American Economic Review 101*(4), 1375–1409.

Barrios, J. and Y. Hochberg (2020). Risk perception through the lens of politics in the time of the covid-19 pandemic. NBER working paper 27008, National Bureau of Economic Research.

Beatty, T. K. M., J. P. Shimshack, and R. J. Volpe (2019). Disaster preparedness and disaster response: Evidence from sales of emergency supplies before and after hurricanes. *Journal of the Association of Environmental and Resource Economists 6*(4), 633–668.

Blackport, R. and J. Screen (2020). Weakened evidence for mid-latitude impacts of arctic warming. *Nature Climate Change 10*, 1065–1066.

Blonz, J., K. Palmer, C. Wichman, and D. Wietelman (2021). Smart thermostats, automation, and time-varying prices. Technical Report 21-20, Resources for the Future.

Brandon, A., C. M. Clapp, J. A. List, R. Metcalfe, and M. Price (2021). Smart tech, dumb humans: The perils of scaling household technologies. Working paper.

Brandon, A., J. List, R. Metcalfe, and F. Rundhammer (2018). Testing for crowd out in social nudges: Evidence from a natural field experiment in the market for electricity. *Proceedings of the National Academy of Sciences 116*(12), 5293–5298.

Brent, D. A., J. H. Cook, and S. Olsen (2015). Social comparisons, household water use, and participation in utility conservation programs: Evidence from three randomized trials. *Journal of the Association of Environmental and Resource Economists 2*(4), 597–627.

Brewer, D. (2022). Equilibrium sorting and moral hazard in residential energy contracts. *Journal of Urban Economics 129*, 103424.

Breza, E., F. C. Stanford, M. Alsan, B. Alsan, A. Banerjee, A. G. Chandrasekhar, S. Eichmeyer, T. Glushko, P. Goldsmith-Pinkham, K. Holland, E. Hoppe, M. Karnani, S. Liegl, T. Loisel, L. Ogbu-Nwobodo, B. Olken, C. Torres, P. L. Vautrey, E. T. Warner, S. Wootton, and E. Duflo (2021). Effects of a large-scale social media advertising campaign on holiday travel and covid-19 infections: a cluster randomized controlled trial. *Nature Medicine 27*, 1622 – 1628.

Brown, Z., N. Johnstone, I. Haščič, L. Vong, and F. Barascud (2013). Testing the effect of defaults on the thermostat settings of oecd employees. *Energy Economics 39*, 128–134.

Burkhardt, J., K. Gillingham, and P. Kopalle (2019). Experimental evidence on the effect of information and pricing on residential electricity consumption. Technical Report 25576, NBER.

Camerer, C., L. Babcock, G. Loewenstein, and R. Thaler (1997, 05). Labor Supply of New York City Cabdrivers: One Day at a Time*. *The Quarterly Journal of Economics 112*(2), 407–441.

Cameron, A. C., J. B. Gelbach, and D. L. Miller (2008, 08). Bootstrap-Based Improvements for Inference with Clustered Errors. *The Review of Economics and Statistics 90*(3), 414–427.

Carlsson, F., C. Gravert, O. Johansson-Stenman, and V. Kurz (2021). The use of green nudges as an environmental policy instrument. *Review of Environmental Economics and Policy 15*(2), 216–237.

Cialdini, R. (2006). *Influence: The Psychology of Persuasion, Revised Edition.* Harper Business.

Consumers Energy Company (2019). Ray compressor station fire, Jan. 30, 2019. Report submitted to Michigan Public Service Commission Case No. U-20463.

Costa, F. and F. Gerard (2021). Hysteresis and the welfare effect of corrective policies: Theory and evidence from an energy-saving program. *Journal of Political Economy 129*(6), 1705–1743.

CQ Press (2014-2019). Voting and elections collection. Governor election returns, county detail by year, http://library.cqpress.com/elections/download-data.php, accessed 05-2020.

DesOrmeau, T. (2019). Michiganders answered call, cut gas usage 10 percent after emergency plea. Technical report, MLive.

Donald, S. G. and K. Lang (2007, 05). Inference with Difference-in-Differences and Other Panel Data. *The Review of Economics and Statistics 89*(2), 221–233.

Edwards, J. T. and J. A. List (2014). Toward an understanding of why suggestions work in charitable fundraising: Theory and evidence from a natural field experiment. *Journal of Public Economics 114*, 1–13.

EIA (2019a). Extreme cold in the Midwest led to high power demand and record natural gas demand. Technical report, United States Energy Information Administration.

EIA (2019b). Record cold temperatures in the upper midwest increase u.s. natural gas heating demand. Technical report, United States Energy Information Administration.

Engström, P., K. Nordblom, H. Ohlsson, and A. Persson (2015, November). Tax compliance and loss aversion. *American Economic Journal: Economic Policy 7*(4), 132–64.

Farber, H. S. (2008, June). Reference-dependent preferences and labor supply: The case of new york city taxi drivers. *American Economic Review 98*(3), 1069–82.

Ferraro, P. J., J. J. Miranda, and M. K. Price (2011, May). The persistence of treatment effects with norm-based policy instruments: Evidence from a randomized environmental policy experiment. *American Economic Review 101*(3), 318–22.

Ferraro, P. J. and M. K. Price (2013, 03). Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. *The Review of Economics and Statistics 95*(1), 64–73.

Foster, A., E. Gutierrez, and N. Kumar (2009, May). Voluntary compliance, pollution levels, and infant mortality in mexico. *American Economic Review 99*(2), 191–97.

Foster, A. D. and E. Gutierrez (2013, May). The informational role of voluntary certification: Evidence from the mexican clean industry program. *American Economic Review 103*(3), 303–08.

Ge, Q. and B. Ho (2019). Energy use and temperature habituation: Evidence from high frequency thermostat usage data. *Economic Inquiry 57*(2), 1196–1214.

Gray, K. (2019). How the Consumers Energy polar vortex emergency unfolded. Technical report, The Detroit Free Press.

Hallsworth, M., J. A. List, R. D. Metcalfe, and I. Vlaev (2017). The behavioralist as tax collector: Using natural field experiments to enhance tax compliance. *Journal of Public Economics 148*, 14–31.

Harding, M. and A. Hsiaw (2014). Goal setting and energy conservation. *Journal of Economic Behavior & Organization 107*, 209–227.

Holladay, J. S., M. K. Price, and M. Wanamaker (2015). The perverse impact of calling for energy conservation. *Journal of Economic Behavior & Organization 110*, 1–18.

Ito, K., T. Ida, and M. Tanaka (2018, February). Moral suasion and economic incentives: Field experimental evidence from energy demand. *American Economic Journal: Economic Policy 10*(1), 240–67.

Iyengar, S., Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood (2019). The origins and consequences of affective polarization in the united states. *Annual Review of Political Science 22*(1), 129–146.

Kim, J. and S. S. Oh (2015). Confidence, knowledge, and compliance with emergency evacuation. *Journal of Risk Research 18*(1), 111–126.

Levitt, S. D. and J. A. List (2007, June). What do laboratory experiments measuring social preferences reveal about the real world? *Journal of Economic Perspectives 21*(2), 153–174.

List, J. A. (2020, July). Non est disputandum de generalizability? a glimpse into the external validity trial. Working Paper 27535, National Bureau of Economic Research.

Luyben, P. D. (1982). Prompting thermostat setting behavior: Public response to a presidential appeal for conservation. *Environment and Behavior 14*(1), 113–128.

Meier, A., U. Tsuyoshi, M. Pritoni, L. Rainer, A. Daken, and D. Baldewicz (2019). What can connected thermostats tell us about american heating and cooling habits? In *ECEEE Summer Study Proceedings*. European Council for an Energy Efficient Economy.

Montero, J. (1999). Voluntary compliance with market-based environmental policy: Evidence from the U.S. acid rain program. *Journal of Political Economy 107*(5), 998–1033.

MPSC (2019a). Michigan statewide energy assessment. Michigan Public Service Commission.

MPSC (2019b). Michigan's statewide energy assessment fact sheet. Michigan Public Service Commission.

NOAA (2019). The science behind the polar vortex: You might want to put on a sweater. Technical report, United States Department of Commerce National Oceanic and Atmospheric Administration.

NOAA (2021). U.s. billion-dollar weather and climate disasters. https://www.ncdc.noaa.gov/billions/ accessed 09-06-2021).

Perryman Group (2021). Preliminary estimates of economic costs of the february 2021 Texas winter storm. Technical report, Perryman Group. https://www.perrymangroup.com/media/uploads/brief/perryman-preliminary-estimates-of-economic-costs-of-the-february-2021-texas-winter-storm-02-25-21.pdf, accessed 05-24-2021.

Reiss, P. and M. White (2008). What changes energy consumption? prices and public pressures. *RAND Journal of Economics 39*(3), 636–663.

Seibold, A. (2021, April). Reference points for retirement behavior: Evidence from german pension discontinuities. *American Economic Review 111*(4), 1126–65.

Sigman, H. and H. F. Chang (2011, May). The effect of allowing pollution offsets with imperfect enforcement. *American Economic Review 101*(3), 268–72.

Thakral, N. and L. T. Tô (2021, August). Daily labor supply and adaptive reference points. *American Economic Review 111*(8), 2417–43.

Thaler, R. and C. Sunstein (2008). *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.

Tversky, A. and D. Kahneman (1981). The framing of decisions and the psychology of choice. *Science 211*(4481), 453 – 458.

US Census Bureau (2019). American community survey 1-year sample.

Whitehead, J. C., B. Edwards, M. V. Willigen, J. R. Maiolo, K. Wilson, and K. T. Smith (2000). Heading for higher ground: factors affecting real and hypothetical hurricane evacuation behavior. *Global Environmental Change Part B: Environmental Hazards 2*(4), 133–142.

Wichman, C. J., L. O. Taylor, and R. H. von Haefen (2016). Conservation policies: Who responds to price and who responds to prescription? *Journal of Environmental Economics and Management 79*, 114–134.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*(2), 1281–1301.

Wooldridge, J. M. (2010). *Econometric Analysis of Cross Section and Panel Data* (2 ed.). The MIT Press.

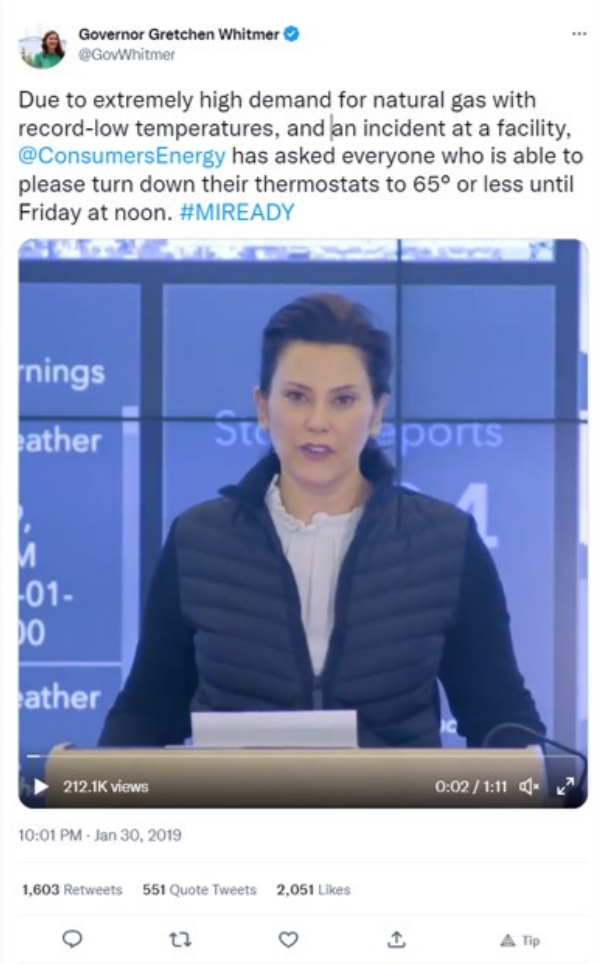
Zou, E. Y. (2021, July). Unwatched pollution: The effect of intermittent monitoring on air quality. *American Economic Review 111*(7), 2101–26.

Appendices

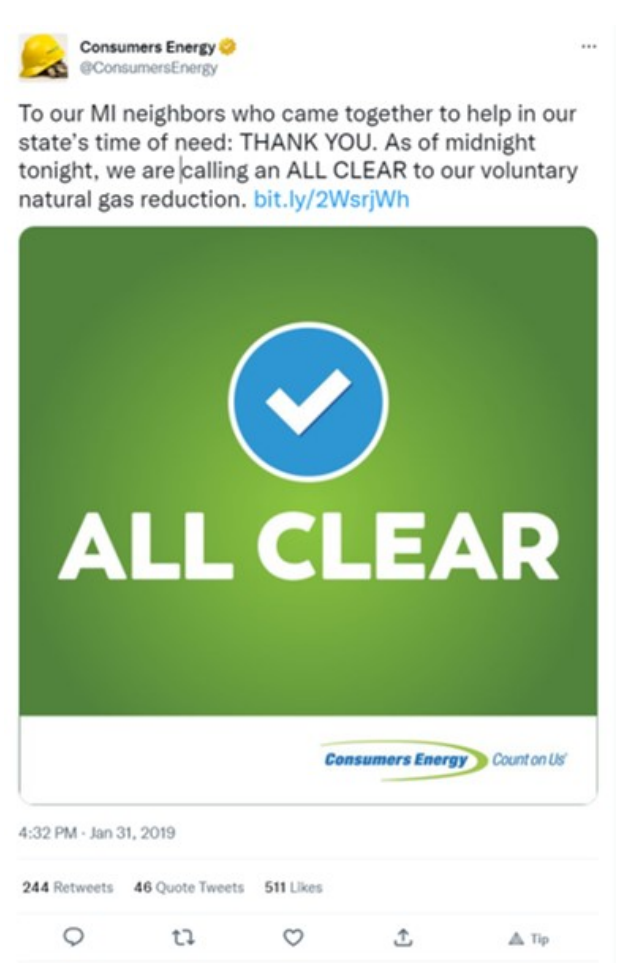
A Social media posts

(a) https://twitter.com/ConsumersEnergy/status/1090692811081551885

(b) https://www.facebook.com/85543026043/videos/357785638397997/



(c) https://twitter.com/GovWhitmer/status/1090807363811065857



(d) https://twitter.com/ConsumersEnergy/status/1091086892592959495

Figure 9: Social media appeals.

B Tests of alternative behavioral responses

Here, we consider potential alternative behavioral responses to the emergency request outside of the thermostat setting. Households may have altered the thermostat mode or turned the furnace off in order to comply with the emergency request. In addition, the emergency request may have either induced households in Michigan to stay at home during the emergency or to seek shelter elsewhere, potentially differently than households in the control states. The smart thermostat data contain variables on the thermostat's mode and whether the thermostat's motion sensor detects motion. Our thermostat mode variables contain information on whether the furnace is turned off, set to "hold" (which keeps a constant thermostat setting), or set to a smart automation setting. The Ecobee thermostat's smart automation feature "smart recovery" can pre-heat or cool a home based on a household's typical behavior (for example, if a person typically arrives home from work at 6:00 pm and increases the thermostat setting, the smart recovery feature may automatically begin heating the home at 5:45 pm in anticipation of arrival).

We test whether the emergency event induced households to set the thermostat to "hold," use the automation feature, turn the furnace off, and spend more or less time at home by using these variables as outcomes in our main specification (equation 5). Each of these variables is a binary indicator equal to one if the thermostat was in the indicated mode or detected motion during the period, making these linear probability models. Table 5 displays these estimates. We estimate a 2 percentage point increase in thermostat settings on "hold" and a 3 percentage point increase in the use of automation, which we interpret as small changes. We find no statistically significant change in whether the furnace was turned off or whether motion was detected by the thermostat.

Although we do not see large responses via these channels, we test the robustness of our thermostat setting estimates to controlling for thermostat mode and the motion sensor indicator in the thermostat setting, compliance, and fan run time regressions. Table 6 contains these estimates. While the furnace fan estimate is slightly larger than the estimates in the main text (table 2), the estimates are overall similar.

Table 5: Estimates of regressions from equation 5 using alternate potential outcome variables. The sample includes observations from January 2nd-January 31st.

	Potential alternative behavior responses			
	(1)	(2)	(3)	(4)
VARIABLES	Thermostat hold	Thermostat automation	Furnace off	Motion detected
Michigan x Post	0.02*	0.03*	-0.00	-0.00
	(0.01)	(0.01)	(0.00)	(0.00)
Constant	0.30**	0.20**	0.01**	0.58**
	(0.01)	(0.00)	(0.00)	(0.01)
Observations	2,144,796	2,144,796	2,144,796	2,144,796
R-squared	0.51	0.30	0.69	0.55
Weather controls	YES	YES	YES	YES
Household FE	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
DOW × HOD × state	YES	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

Table 6: Estimates of the regressions from equation 5 controlling for thermostat mode and motion sensor. The sample includes observations from January 2nd-January 31st.

	Controlling for thermostat mode and automation settings		
	(1)	(2)	(3)
VARIABLES	Thermostat setting	Thermostat ≤ 65F	Fan run time
Michigan x Post	-1.07**	0.10**	-1.63*
	(0.12)	(0.01)	(0.50)
Constant	66.89**	0.32**	16.37**
	(0.20)	(0.01)	(1.04)
Observations	2,126,336	2,126,336	2,135,114
R-squared	0.74	0.53	0.76
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week × hour of day × state	YES	YES	YES
Thermostat modes	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

Table 7: Estimates of the regressions from equation 5 using hourly data from January 2nd-January 31st.

Two-way fixed effects regressions

VARIABLES	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
Michigan x Post	-1.116**	0.126**	-1.420
	(0.134)	(0.008)	(0.521)
Constant	67.439**	0.248**	25.121**
	(0.039)	(0.002)	(0.755)
Observations	8,474,273	8,474,273	8,509,460
R-squared	0.666	0.476	0.631
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

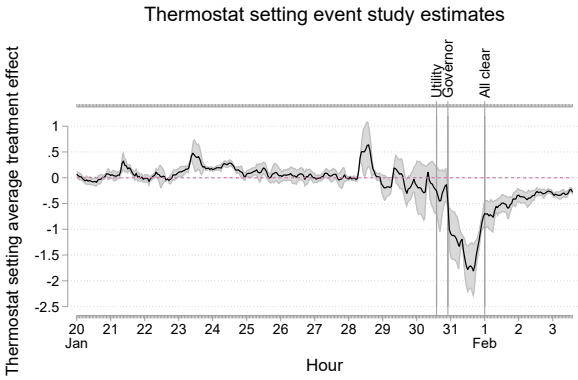
Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

C Hourly analysis

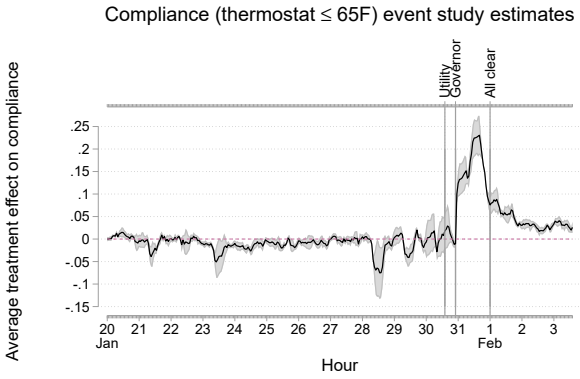
In this section, we replicate the analyses from the main text using hourly data rather than data aggregated to four-hour intervals. Table 7 contains average treatment effect estimates, figure 10 displays the event-study estimates, and table 8 displays the heterogeneity estimates. We find that the choice of time frequency makes very little difference for the main estimates, but the four-hour intervals reduce noise in the pre-period of the event-study analysis, resulting in cleaner pre-trends. Given the reduction in noise, we favor the four-hour analysis displayed in the main text.

D Robustness checks

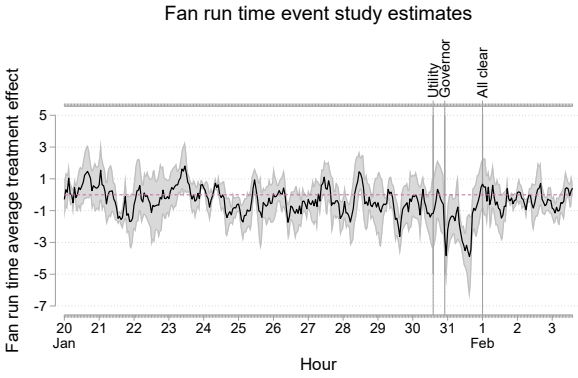
In this section, we test the sensitivity of the average treatment effect estimates to specification, sample selection, and potential spillovers. We find that the estimates are not affected by these potential confounders.



(a)



(b)



(c)

Figure 10: Event-study coefficients estimated on hourly data using regression equation 6 with (a) thermostat setting, (b) compliance, and (c) minutes of furnace fan run time as the dependent variables. 95 percent confidence intervals constructed from standard errors cluster-robust to heteroskedasticity.

Table 8: Results from the estimation of equation 7 using hourly data from January 2nd-January 31st. Standard errors cluster-bootstrapped to incorporate the sampling error from estimation of the baseline thermostat setting.

	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
59 F or lower expected X Treatment	2.70**	0.064**	0.75
	(0.40)	(0.023)	(0.76)
59-61 F expected X Treatment	1.01**	0.045*	0.41
	(0.17)	(0.019)	(0.58)
61-63 F expected X Treatment	0.59**	0.019	-0.0024
	(0.099)	(0.017)	(0.48)
65-67 F expected X Treatment	-0.68**	0.25**	-0.61
	(0.063)	(0.012)	(0.34)
67-69 F expected X Treatment	-1.43**	0.34**	-1.32**
	(0.078)	(0.016)	(0.38)
69-71 F expected X Treatment	-1.72**	0.29**	-1.79**
	(0.085)	(0.012)	(0.38)
71-73 F expected X Treatment	-1.87**	0.25**	-2.14**
	(0.15)	(0.014)	(0.52)
73-75 F expected X Treatment	-1.87**	0.25**	-2.63**
	(0.25)	(0.019)	(0.89)
Higher than 75 F expected X Treatment	-1.40**	0.21**	-2.86
	(0.40)	(0.019)	(1.53)
40-45% Democrat X Treatment	-0.43	0.069*	0.28
	(0.27)	(0.029)	(0.71)
45-50% Democrat X Treatment	-0.11	0.034	1.32
	(0.27)	(0.029)	(0.99)
50-55% Democrat X Treatment	-0.33	0.054*	-0.22
	(0.23)	(0.027)	(0.97)
55-60% Democrat X Treatment	-0.25	0.049	2.43
	(0.30)	(0.039)	(1.56)
60-65% Democrat X Treatment	-0.51	0.085	2.05
	(0.40)	(0.055)	(1.86)
65-70% Democrat X Treatment	-0.53	0.086	-0.22
	(0.43)	(0.059)	(1.76)
70-75% Democrat X Treatment	-0.82	0.088	2.65
	(0.42)	(0.051)	(1.84)
Observations	7,811,227	7,811,227	7,812,552
R-squared	0.80	0.69	0.64
FE	YES	YES	YES
Hour	YES	YES	YES
Controls	YES	YES	YES
Expected thermostat level	YES	YES	YES

Standard errors cluster-bootstrapped at the city level.
** p<0.01, * p<0.05

Table 9: Alternate difference-in-differences specifications with temperature as the outcome variable. The sample includes observations from January 2nd-January 31st.

Thermostat setting DID

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Michigan	-0.588	-0.591			
	(0.356)	(0.349)			
Post	1.203**	1.054*	0.858**	0.675**	
	(0.180)	(0.287)	(0.149)	(0.110)	
Michigan x Post	-0.992**	-1.171**	-1.196**	-1.100**	-1.052**
	(0.180)	(0.223)	(0.161)	(0.135)	(0.129)
Constant	67.460**	68.055**	68.266**	67.572**	67.472**
	(0.356)	(0.597)	(0.091)	(0.062)	(0.047)
Observations	2,126,410	2,126,338	2,126,336	2,126,336	2,126,336
R-squared	0.008	0.012	0.687	0.705	0.708
Weather controls		YES	YES	YES	YES
Household FE			YES	YES	YES
Time FE					YES
Day of week				YES	
Hour of day				YES	

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

Table 10: Alternate difference-in-differences specifications with compliance as the outcome variable. The sample includes observations from January 2nd-January 31st.

Thermostat $\leq$ 65F LPM DID

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Michigan	0.049	0.049			
	(0.025)	(0.025)			
Post	-0.081**	-0.074*	-0.061**	-0.043**	
	(0.010)	(0.023)	(0.009)	(0.007)	
Michigan x Post	0.093**	0.108**	0.111**	0.102**	0.098**
	(0.010)	(0.014)	(0.007)	(0.008)	(0.007)
Constant	0.221**	0.163*	0.142**	0.211**	0.220**
	(0.025)	(0.045)	(0.014)	(0.006)	(0.004)
Observations	2,126,410	2,126,338	2,126,336	2,126,336	2,126,336
R-squared	0.004	0.007	0.471	0.497	0.502
Weather controls		YES	YES	YES	YES
Household FE			YES	YES	YES
Time FE					YES
Day of week				YES	
Hour of day				YES	

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

Table 11: Alternate difference-in-differences specifications with fan run time as the outcome variable. The sample includes observations from January 2nd-January 31st.

Fan minutes running DID

	(1)	(2)	(3)	(4)	(5)
VARIABLES	Model 1	Model 2	Model 3	Model 4	Model 5
Michigan	-2.503*	-2.997*			
	(0.559)	(0.894)			
Post	10.254**	1.409*	0.661**	0.898**	
	(0.735)	(0.419)	(0.128)	(0.102)	
Michigan x Post	-2.206*	-0.933	-1.338*	-1.450*	-1.483*
	(0.735)	(0.591)	(0.451)	(0.498)	(0.483)
Constant	20.593**	26.604**	25.473**	26.080**	25.225**
	(0.559)	(1.479)	(0.568)	(0.445)	(0.669)
Observations	2,135,188	2,135,116	2,135,114	2,135,114	2,135,114
R-squared	0.020	0.071	0.737	0.749	0.751
Weather controls		YES	YES	YES	YES
Household FE			YES	YES	YES
Time FE					YES
Day of week				YES	
Hour of day				YES	

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

First, we examine the effect of specification choice on the difference-in-difference estimates. In the main text, we display results using a two-way fixed effects approach. Tables 9 - 11 display alternative specification coefficient estimates for each outcome variable. Column 1 displays estimates using a standard difference in differences specification with indicator variables for being in Michigan, being in the post-treatment period, and the interaction of these two. Column 2 adds time-varying controls for temperature and humidity, column three replaces the treatment group indicator with household fixed effects, and column four replaces the post-treatment indicator with day-of-week and time of day indicator variables. The results do not differ substantially across all specifications for each outcome variable.

Table 12: Estimates of the regressions from equation 5 on a balanced panel of households. The sample includes observations from January 2nd-January 31st.

Robustness check: Estimated on balanced panel

	(1)	(2)	(3)
VARIABLES	Thermostat setting	Thermostat $\leq$ 65F	Fan run time
Michigan x Post	-1.09**	0.10**	-1.53*
	(0.13)	(0.01)	(0.52)
Constant	67.40**	0.23**	25.25**
	(0.04)	(0.00)	(0.80)
Observations	1,850,760	1,850,760	1,850,760
R-squared	0.71	0.50	0.75
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

Next, we consider the possibility of sample selection. In the sample, 6.1% of households enter late or leave early. This is best thought of as a sample selection problem as we do not observe the households before or after these points. In addition, 5.2% of observations are missing data on thermostat setting or weather data. Because entry, exit, and missingness are unrelated to the treatment, we consider the missing observations to be "missing completely at random" and are therefore unrelated to the error term (Wooldridge, 2007). Nonetheless, we

Table 13: Estimates of the regressions from equation 8, allowing the treatment to potentially spill over into border counties. The sample includes observations from January 2nd-January 31st.

	Spillovers		
	(1)	(2)	(3)
VARIABLES	Thermostat setting	Thermostat $\leq$ 65F	Fan run time
Michigan x Post	-1.08**	0.10**	-1.47*
	(0.13)	(0.01)	(0.52)
Border county x Post	-0.12	0.00	-0.07
	(0.22)	(0.01)	(0.60)
Constant	67.43**	0.22**	24.82**
	(0.04)	(0.00)	(0.79)
Observations	2,126,336	2,126,336	2,135,114
R-squared	0.71	0.50	0.75
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

interpolate missing values and drop households that enter the sample late or leave early and estimate the average treatment effect using a balanced panel of households using the two-way fixed effects specification of equation 5. Table 12 displays the results of this estimation. The estimates are slightly smaller than those in the main text but are not substantially different.

Our next robustness check allows for the possibility of spillovers to counties bordering Michigan. Because the text alerts go to cellphones based upon the closest cell phone tower, it is possible that households living near the border in Indiana and Ohio were also treated. Illinois and Wisconsin do not border Michigan's lower peninsula. 3.4 percent of Ohio and 11.6 percent of Indiana sample households live in counties bordering Michigan. We estimate the following regression, which allows for a spillover treatment effect for households living in border counties:

$$Y_{i,t} = \alpha_i + \lambda_t + \beta D_{i,t} + \sigma S_{i,t} + \gamma X_{i,t} + \delta_{s,h,d} + \varepsilon_{i,t}, \tag{8}$$

where $S_{i,t}$ is a treatment variable equal to one for households in counties that border Michigan's lower peninsula during the post-treatment period. The estimated coefficient σ on $S_{i,t}$ should be equal to zero if there are no spillovers into the bordering counties. Table 13 displays the estimated coefficients using all three outcome variables. In each regression, the estimated spillover coefficient is small and the confidence interval contains zero. The coefficient on the main treatment variable is not substantially different from the estimates in the main text. We have experimented with modeling potential spillovers as far as two counties away from the border and we do not see much difference in the estimated coefficients on the main treatment variable, so we do not display the results here. In addition to these empirical results, searches of the archives of Toledo, Ohio's main newspaper, *The Blade*, using the keywords "polar vortex" and "natural gas" does not turn up any news coverage of the Michigan event. These results suggest that any potential spillover effects are not affecting the estimates.

E Placebo analyses

Ten days before the polar vortex, Michigan experienced a similar cold wave that did not coincide with a supply-side emergency causing an emergency request for voluntary thermostat reductions. Figure 3a displays mean daily temperatures for sample households in Michigan in January 2019. Temperatures on January 20-21 dropped from 25°F to below 10°F, making these days a good placebo event for the January 30-31 emergency. Because there was no emergency request to reduce thermostat settings on the 20th and 21st, we would expect to see no difference in heating behavior for Michigan and control states.

E.1 Average treatment effects

Figure 11 displays average thermostat settings, fraction of households with thermostat settings at or below 65°F, and fan running times for Michigan and control states from January 19-22. Unlike in the main text, we do not see a change in heating behavior between Michigan

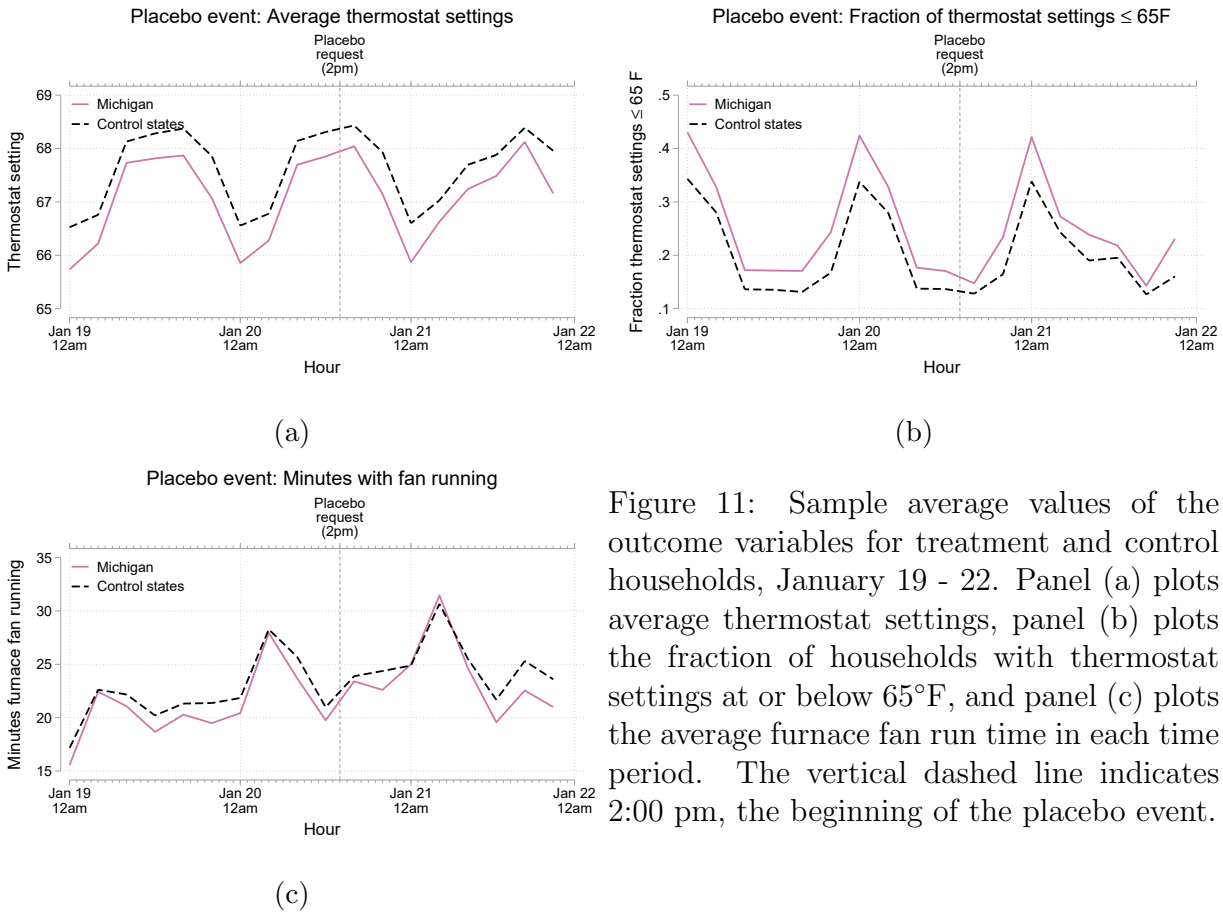


Figure 11: Sample average values of the outcome variables for treatment and control households, January 19 - 22. Panel (a) plots average thermostat settings, panel (b) plots the fraction of households with thermostat settings at or below 65°F, and panel (c) plots the average furnace fan run time in each time period. The vertical dashed line indicates 2:00 pm, the beginning of the placebo event.

and the controls states after the placebo treatment time.

We then estimate the two-way fixed-effects specification of regression equation 5 for the placebo event. To do so, we include all observations observed between January 1st and January 21st, 2019 and treat 2:00 pm on January 20th as the placebo treatment time. Table 14 displays the results of this estimation using thermostat setting, an indicator variable for having the thermostat at or below 65°F, and fan running time as outcome variables. As expected, the estimates are close to zero. The only statistically significant estimate suggests that Michigan households increased thermostat settings by 0.06°F, which we interpret as a precisely estimated zero. This placebo procedure demonstrates that households in Michigan respond similarly to cold spells as households in the control states absent a request to reduce energy consumption.

Table 14: Estimates of the regressions from equation 5 on the sample of households observed between January 1st and January 21st, treating January 20th at 2:00 pm as the placebo treatment time. We expect the estimates from this placebo estimation to be close to zero.

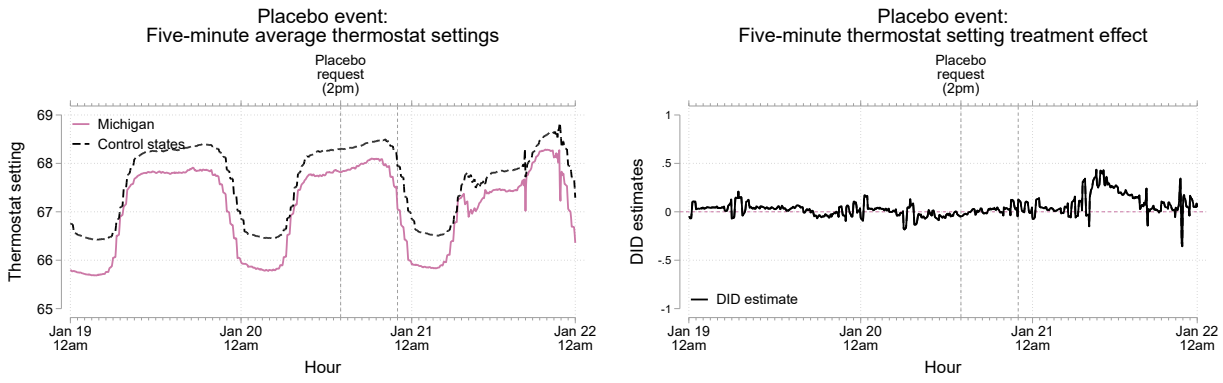
Placebo estimates from Jan 20-21 cold wave

VARIABLES	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
Michigan x Post	0.06*	-0.00	-0.08
	(0.02)	(0.00)	(0.39)
Constant	67.24**	0.23**	23.64**
	(0.06)	(0.01)	(0.57)
Observations	1,413,411	1,413,411	1,419,364
R-squared	0.72	0.52	0.77
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

Robust standard errors clustered at the state level.
** p<0.01, * p<0.05

E.2 Five-minute thermostat settings

In figure 12a, we plot five-minute thermostat setting data for Michigan and the control states between 12:00 pm on January 20th and 11:59 pm on January 21st. In figure 12b, we plot a difference-in-differences estimate of the placebo treatment effect, which we construct as the difference between five-minute thermostat setting for Michigan and control states during the event minus the average difference in same time-of-day and day-of-week five-minute thermostat settings before the event. In these figures, we see a discrete increase in thermostat setting in the five-minute periods beginning at 7:00 pm and 12:00 am for both treatment and control households. These discrete changes correspond to commonly programmed times for the thermostat to automatically change. Other than in a few five-minute periods on January 21st, the measured treatment effect is zero in the placebo period. These treatment effects are driven by seemingly spurious five-minute jumps in the average thermostat setting for Michigan households. Overall, these placebo plots demonstrate the validity of the difference-in-differences approach and verify that discrete increases at 7:00



(a) (b)

Figure 12: (a): Five-minute sample mean thermostat settings for Michigan and control households from 12:00 pm January 20th - 11:59 pm January 21st. (b): Five-minute difference-in-differences estimate.

pm and 12:00 am are common and not artifacts of the emergency event.

E.3 Heterogeneity analysis

One concern that we had was that the estimates from the heterogeneity analysis in the main text reflect statistical reversion to the mean rather than a meaningful pattern of results for the sub-groups (particularly for the baseline thermostat-setting results). To test this, we estimate the heterogeneity regression analysis from equation 7 during the placebo period. Table 15 displays the estimated coefficients and bootstrapped standard errors that account for the first-stage estimation of the baseline thermostat setting. The magnitude of the estimated coefficients in the placebo analysis are almost all two to five times smaller than during the polar vortex and do not display as clear trends as you move away from 65°F. Thus, we conclude that mean reversion may play a small role in the main heterogeneity estimates but the effect is not large enough to alter our conclusions in section 5.3.

Table 15: Results from the estimation of equation 7 using the placebo cold wave. The sample includes observations from January 2nd-January 21st.

	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
59 F or lower expected X Treatment	1.59**	0.091**	0.23
	(0.39)	(0.026)	(0.79)
59-61 F expected X Treatment	0.50*	0.061**	0.089
	(0.20)	(0.022)	(0.62)
61-63 F expected X Treatment	0.41**	0.023	0.014
	(0.12)	(0.018)	(0.48)
65-67 F expected X Treatment	-0.29**	0.16**	-0.23
	(0.073)	(0.015)	(0.36)
67-69 F expected X Treatment	-0.53**	0.19**	-0.088
	(0.077)	(0.016)	(0.36)
69-71 F expected X Treatment	-0.66**	0.19**	-0.34
	(0.088)	(0.015)	(0.38)
71-73 F expected X Treatment	-0.63**	0.17**	-0.27
	(0.13)	(0.016)	(0.60)
73-75 F expected X Treatment	-0.99**	0.17**	-0.38
	(0.23)	(0.021)	(0.92)
Higher than 75 F expected X Treatment	-0.81*	0.17**	0.40
	(0.41)	(0.021)	(1.49)
40-45% Democrat X Treatment	0.12	0.0026	0.73
	(0.27)	(0.026)	(0.86)
45-50% Democrat X Treatment	-0.16	-0.018	0.71
	(0.25)	(0.026)	(1.02)
50-55% Democrat X Treatment	-0.20	0.012	0.55
	(0.23)	(0.023)	(1.06)
55-60% Democrat X Treatment	-0.057	0.0090	1.47
	(0.40)	(0.035)	(1.61)
60-65% Democrat X Treatment	0.028	0.0056	2.41
	(0.37)	(0.050)	(1.67)
65-70% Democrat X Treatment	-0.35	-0.011	1.35
	(0.41)	(0.051)	(1.77)
70-75% Democrat X Treatment	-0.39	0.030	1.57
	(0.43)	(0.040)	(1.94)
Observations	1,302,219	1,302,219	1,302,511
R-squared	0.85	0.75	0.78
FE	YES	YES	YES
Hour	YES	YES	YES
Controls	YES	YES	YES
Expected thermostat level	YES	YES	YES

Standard errors cluster-bootstrapped at the city level.
** p<0.01, * p<0.05

F Donald and Lang inference

Here, we use an aggregation approach inspired by Donald and Lang (2007) to estimate the average treatment effects and provide valid inference for five clusters. In our approach, we average our outcome variables to the state level s and estimate the following ordinary-least-squares regression:

$$Y_{s,t} = \alpha + \lambda_t + \beta D_{s,t} + \gamma X_{s,t} + \delta_{s,h,d} + \varepsilon_{s,t}, \tag{9}$$

where $Y_{s,t} = \bar{Y}_{i,t}$ and $X_{s,t} = \bar{X}_{i,t}$ are the state sample averages. Under standard exogeneity assumptions and large state-cluster group sizes (implying normality of $\varepsilon_{s,t}$ via the central limit theorem) estimation of β in equation 9 is consistent and standard inference is valid (Wooldridge, 2010). When there are equal cluster sizes and a balanced panel, the aggregated Donald and Lang estimates are exactly equal to the estimates from the individual-level regression, but in our case the unbalanced panel and unequal cluster sizes will result in a slight difference in estimates.

Table 16 displays the Donald and Lang estimates of the average treatment effect for each outcome variable. The estimated average treatment effects are slightly smaller than those in the main text, but are not substantially different. Importantly, each estimated effect is statistically significant even under the conservative Donald and Lang inference test, which alleviates concerns that clustering at the state level may result in over-rejection of the null hypothesis.

Table 16: Donald and Lang estimation of average treatment effects with valid inference for very few clusters. The sample includes observations from January 2nd-January 31st.

VARIABLES	(1) Thermostat setting	(2) Thermostat $\leq$ 65F	(3) Fan run time
Michigan x Post	-1.008**	0.104**	-0.981**
	(0.052)	(0.005)	(0.344)
Constant	67.220**	0.233**	23.824**
	(0.079)	(0.008)	(0.516)
Observations	900	900	900
R-squared	0.990	0.989	0.985
Weather controls	YES	YES	YES
Household FE	YES	YES	YES
Time FE	YES	YES	YES
Day of week $\times$ hour of day $\times$ state	YES	YES	YES

Donald and Lang standard errors reported.
** p<0.01, * p<0.05