E2e Working Paper 022

The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons

Hunt Allcott and Judd B. Kessler

October 2016

Revised version published in

American Economic Journal: Applied Economics,

2019, (11)1, 236-276.

The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons

Hunt Allcott and Judd B. Kessler*

October 23, 2016

Abstract

"Nudge"-style interventions are typically evaluated on the basis of their effects on behavior, not social welfare. We use a field experiment to measure the welfare effects of one especially policy-relevant intervention, home energy conservation reports. We measure consumer welfare by sending introductory reports and using an incentive-compatible multiple price list to determine willingness-to-pay to continue the program. We combine this with estimates of implementation costs and externality reductions to carry out a comprehensive welfare evaluation. We find that this nudge increases social welfare, although traditional program evaluation approaches overstate welfare gains by a factor of six. To exploit significant individual-level heterogeneity in welfare gains, we develop a simple machine learning algorithm to optimally target the nudge; this would increase welfare gains by more than 150 percent. Our results highlight that nudges, even those that are highly effective at changing behavior, need to be evaluated based on their welfare implications.

JEL Codes: C44, C53, D12, L94, Q41, Q48.

Keywords: Behavioral interventions, energy efficiency, machine learning, program evaluation, randomized field experiments, smart defaults, social comparisons, welfare analysis.

*Allcott: New York University, NBER, E2e, JPAL, and ideas42. hunt.allcott@nyu.edu. Kessler: Wharton, NBER, and JPAL. judd.kessler@wharton.upenn.edu. We are grateful to Paula Pedro for outstanding research management, and we thank Opower and Central Hudson Gas and Electric for productive collaboration and helpful feedback. We also thank Nava Ashraf, Stefano DellaVigna, Avi Feller, Michael Greenstone, Ben Handel, Guido Imbens, Kelsey Jack, David Laibson, John List, Todd Rogers, Dmitry Taubinsky, and seminar participants at Berkeley, Chicago Booth, the Consumer Financial Protection Bureau, Cornell, Microsoft Research, New York University, Stanford, Stanford Institute for Theoretical Economics, Wesleyan, the World Bank, and Yale for comments. We are grateful to the Sloan Foundation and the Jameel Poverty Action Lab for grant funding. This RCT was registered in the American Economic Association Registry for randomized control trials under trial number 713. Code to replicate the analysis is available from Hunt Allcott's website.

Policymakers and academics are increasingly interested in "nudges," such as information provision, reminders, social comparisons, default options, and commitment contracts, which can affect behavior without changing prices or choice sets. Nudges are being used to encourage a variety of privately-beneficial and socially-beneficial behaviors, such as healthy eating, exercise, organ donation, charitable giving, retirement savings, and environmental conservation. The US, British, and Australian governments, among others, have set up "nudge units" to infuse these ideas into the policy process.[1] A growing list of academic papers evaluate nudge-style interventions in various domains.[2]

With only a few exceptions discussed below, nudges are typically evaluated based on the magnitude of behavior change or on cost effectiveness. When a nudge significantly increases a positive behavior at low cost, policymakers often advocate that it be broadly adopted. A full social welfare evaluation could produce different policy prescriptions, however, because people being nudged often experience two types of benefits and/or costs that typical evaluations do not consider. First, nudge recipients often incur costs in order to change behavior. For example, people who quit smoking save money on cigarettes but give up any enjoyment from smoking, and healthy eating might mean paying more for vegetables and giving up tasty desserts.[3] Second, the nudge itself may directly impose positive or negative utility. For example, seeing cigarette warning labels with graphic images of smoking-related diseases can be unpleasant, and body weight report cards could make children feel guilty or uncomfortable. Building on Caplin (2003) and Loewenstein and O'Donoghue (2006), Glaeser (2006) argues that many nudges are essentially emotional taxes that reduce utility but do not raise revenues.

This paper presents a social welfare evaluation of Home Energy Reports (HERs), one-page letters that compare a household's energy use to that of its neighbors and provide energy conservation tips. While HERs are just one case study, they are one of the most prominent and frequently-studied nudges. As of mid-2015, Opower, the leading HER provider, was working with 95 utility companies in nine countries, sending HERs regularly to 15 million households. There has been significant academic interest in HERs, including seminal studies by Schultz *et al.* (2007) and Nolan *et al.* (2008) and many follow-on evaluations of social comparisons and other "behavior-based" energy conservation interventions.[4] There are also a plethora of industry studies and regulatory evaluations of

[1]In September 2015, the US "nudge unit," the Social and Behavioral Sciences Team, released results from 15 experiments, and President Obama signed an executive order that directs federal agencies to use behavioral insights when they "may yield substantial improvements in social welfare and program outcomes" (EOP 2015). See Whitehead *et al.* (2014) for an overview of the influence of nudge units worldwide.

[2]One indicator of academic interest is that the book Nudge (Thaler and Sunstein 2008) has been cited more than 7000 times.

[3]Of course, if the policymaker has correctly designated a "good" behavior to nudge people toward, this typically means that the behavior change generates net benefits for the individual. However, the magnitude of these net benefits would ideally be calculated and weighed against a nudge's other costs and benefits.

[4]Academic papers on energy use social comparison reports include Kantola, Syme, and Campbell (1984), Allcott (2011, 2015), Ayres, Raseman, and Shih (2013), Costa and Kahn (2013), Dolan and Metcalfe (2013), Allcott and Rogers (2014), and Sudarshan (2014). Delmas, Fischlein, and Asensio (2013) review 156 published field trials studying

such programs.[5]

These existing evaluations of behavior-based energy conservation programs often make policy recommendations by comparing program implementation costs to the value of energy saved. This approach is so well-established that energy industry regulators have a name for it: the "program administrator cost test." As with most evaluations of other nudges, this ignores benefits and costs (other than energy cost savings) experienced by nudge recipients. For example, what financial costs did consumers incur to generate the observed energy savings (e.g., to install improved insulation)? What is the cost of time devoted to turning off lights or adjusting thermostats? What is the value of comfort from better-insulated homes or the discomfort from setting thermostats to energy-saving temperatures? Are there meaningful psychological benefits or costs of using social comparisons to inspire or guilt people into conserving energy?

Home Energy Reports have two features that we leverage to conduct a social welfare analysis that considers the full range of the nudge's benefits and costs. First, they are a private good that can be sold. Second, the standard policy is to deliver them regularly (e.g., every two months) over several years. These two features mean that it is both possible and policy-relevant to measure willingness-to-pay (WTP) for future HERs in a sample of experienced past recipients. In simple terms, our approach is to send people one year of HERs, each of which has a similar structure but includes new conservation tips and updated energy use feedback, and then ask them how much they are willing to pay to receive HERs for a second year. Because these people have experience with HERs from the first year, we respect their WTP as an accurate measure of their welfare from receiving more of them. We then use standard economic tools to evaluate the welfare effects of the second year of HERs, including effects on consumer welfare along with implementation costs and reductions in uninternalized externalities.

More specifically, we study a program providing HERs to about 10,000 residential natural gas consumers at a utility in upstate New York over the 2014-2015 and 2015-2016 winter heating seasons. At the end of winter 2014-2015, we surveyed all HER recipients by mail and phone with multiple price lists (MPLs) that trade off next winter's HERs with checks for different amounts of money. We designed the MPL to allow negative WTP as well as positive WTP, as some households opt out of HER programs even though the reports are free. The MPLs were incentive-compatible — depending on their responses, each household received a check from the utility and/or more HERs in winter 2015-2016. The initial HER treatment group was randomly assigned from a larger population as part of a randomized control trial, so we can easily estimate the effects of HERs on energy use, which we then translate to a value of uninternalized externalities using parameters such as the social cost of carbon.

We find that the average household is willing to pay just under $3 for a second year of Home

social comparisons and other informational interventions to induce energy conservation.

[5]These include Violette, Provencher, and Klos (2009), Ashby *et al.* (2012), Integral Analytics (2012), KEMA (2012), Opinion Dynamics (2012), and Perry and Woehleke (2013), among many others.

Energy Reports. While most people like HERs, 34 percent have weakly negative WTP — that is, they prefer not to be nudged even if the nudge is free. In support of our revealed preference approach, the data suggest that WTP is a reliable measure of how much people like HERs: for example, WTP is highly correlated with qualitative evaluations of the HERs and beliefs about savings made possible by future HERs. We estimate that WTP equals about 54 percent of retail energy cost savings, meaning that the remaining 46 percent represents net financial, time, comfort, and psychological costs required to generate the energy savings. This high ratio of energy savings to costs suggests that, leaving aside the implementation cost, HERs provide privately-useful conservation information and/or psychological benefits. However, this 46 percent "non-energy cost" is not included in previous HER evaluations, nor in most evaluations of similar nudges in other domains.

Our main estimates suggest that the second year of this HER program increases social welfare by \$0.49 per household. However, the standard approach of ignoring non-energy costs overstates this welfare gain by a factor of six. We find the same qualitative results in a more speculative calculation where we generalize the 46 percent non-energy cost rule of thumb to the full course of a typical HER program: under this assumption, the typical program likely increases welfare, but ignoring non-energy costs overstates welfare gains by a factor of 2.2.

The nudge's welfare effects are driven down by the fact that about 60 percent of nudge recipients are not willing to pay the social marginal cost of the nudge, including many who have negative WTP. On the other hand, more than 30 percent of recipients are willing to pay more than twice the social marginal cost. A natural response to heterogeneous valuations would be to price the nudge at expected net social cost and let people opt in if they want to. In this context, however, inertia is extremely powerful.[6] We show that even under generous assumptions, an opt-in program is unlikely to enroll enough people to generate welfare gains that are much larger than the current opt-out policy. Instead, we train a simple machine learning algorithm to set a "smart default" — that is, to target the program at consumers that would generate the largest welfare gains if nudged. The smart default approach can increase welfare gains by more than 150 percent, holding constant the number of nudge recipients.

These results have important but nuanced implications for energy policy. Many utilities send HERs to help comply with regulations called Energy Efficiency Portfolio Standards (EEPS), which require utilities to induce a specific quantity of energy savings each year. While this paper finds that net benefits of HERs are less than previously reported, benefit-cost analyses of alternative energy efficiency programs such as home retrofits also may suffer from systematic biases.[7] Thus,

[6] HERs involve much lower stakes than other contexts where defaults have been shown to have significant power, such as health insurance and retirement savings plans as studied by Madrian and Shea (2001), Kling *et al.* (2012), Handel (2013), Ericson (2014), and others

[7] One potential source of systematic bias is that actual energy savings may be different than simulation-based assumptions; see Nadel and Keating (1991) and more recent studies such as Allcott and Greenstone (2015) and Fowlie, Greenstone, and Wolfram (2015). A second source of bias is that according to Kushler *et al.* (2012), only 30

substituting to alternative programs that have not been subjected to a complete social welfare analysis may not be better than continuing an HER program. At a minimum, our results suggest that there is much work to be done to correctly measure the welfare effects of energy efficiency programs.

We are not the first or only researchers to consider the welfare effects of nudges. A handful of previous empirical and theoretical analyses of behaviorally-motivated policies have recognized the difference between effects on behavior and effects on welfare, including Carroll, Choi, Laibson, Madrian, and Metrick (2009) and Bernheim, Fradkin, and Popov (2015) on optimal retirement savings plan defaults; Ito, Ida, and Tanaka (2015) on peak electricity use; Handel (2013) on insurance plan choice; Ambuehl, Bernheim, and Lusardi (2014) on financial education; Bhattacharya, Garber, and Goldhaber-Fiebert (2015) on exercise commitment contracts; and Reyniers and Bhalla (2013) and Cain, Dana, and Newman (2014) on charitable giving. There is an active literature debating the welfare gains from cigarette graphic warning labels, including Weimer, Vining, and Thomas (2009), FDA (2011), Chaloupka *et al.* (2014), Ashley, Nardinelli, and Lavaty (2015), Chaloupka, Gruber, and Warner (2015), Cutler, Jessup, Kenkel, and Starr (2015), Jin, Kenkel, Liu, and Wang (2015), and others. Even within these papers that are grounded in a welfare framework, however, most do not actually implement an empirical social welfare analysis of a nudge because measuring consumer welfare can be so challenging.

Although not a study of a nudge intervention, DellaVigna, List, and Malmendier (2012) is similar in spirit: they point out that charitable donation appeals could increase utility by activating warm glow of donors or instead decrease utility by imposing social pressure. They combine an "avoidance design" — measuring whether people avoid opportunities to donate — with a structural model, concluding that door-to-door fundraising drives can reduce welfare even as they raise money for charity. Herberich, List, and Price (2012) use the same design to show that both altruism and social pressure motivate people to buy energy efficient lightbulbs from door-to-door salespeople, and Andreoni, Rao, and Trachtman (2011) and Trachtman *et al.* (2015) use a different avoidance design to study motivations for charitable giving, although none of these latter three papers includes a social welfare analysis. Avoidance designs achieve the same conceptual goal as our MPL: both allow the analyst to observe people opting in or out of a nudge (or opportunity to donate) at some cost. Our MPL design is especially useful, however, because it immediately gives a WTP, whereas avoidance behaviors require additional assumptions or structural estimates to be translated into dollars.

Section I formally defines a "nudge" and derives a welfare effect formula. Sections II and III present the experimental design and data. Sections IV and V present the empirical results and social welfare calculation. Section VI evaluates targeting and opt-in policies, and Section VII concludes.

---

percent of energy efficiency programs measure non-energy benefits and costs such as the financial, time, and utility costs discussed above. Depending on the program, these factors could bias welfare estimates in either direction.

I Theoretical Framework

This section lays out a simple theoretical framework that formalizes what we mean by a "nudge" and derives an equation for welfare effects.

I.A Consumers and Producers

We model a population of heterogeneous consumers who derive utility from consuming numeraire good x and a continuous choice e, which in our application is energy use. With slight modifications to the below, e could also represent healthful eating, exercise, using preventive health care, charitable giving, or other actions. e generates consumption utility $f(e;\alpha)$, where α is a taste parameter. To capture imperfect information or behavioral bias, we allow a factor γ that affects choice but not experienced utility. For example, γ could represent noise in a signal of an unknown production function for health or household energy services, or it could represent a mistake in evaluating the private net benefits of e, perhaps due to inattention or present bias. Consumers have perceived consumption utility $\hat{f}(e;\alpha,\gamma)$, which may or may not equal $f(e;\alpha)$.

e is produced at constant marginal cost c_e and sold at constant price p_e, giving markup $\pi_e = p_e - c_c$ per unit. For another application, one might extend the model to endogenize price. In our application, e is sold by a regulated utility that is allowed a constant markup over marginal cost, so π_e is exogenous and positive. e imposes constant externality ϕ_e per unit. Consumers have income y and pay lump-sum tax T to the government.

We include a "moral utility" term $M = m - \mu e$. Following Levitt and List (2007), moral utility arises when actions impose externalities, are subject to social norms, or are scrutinized by others. This concept is especially appropriate for our setting, where energy production causes environmental externalities and Home Energy Reports scrutinize energy use and present social norms. The moral price μ can be thought of as a "psychological tax" or "moral tax" on e, as in Glaeser (2006, 2014) and Loewenstein and O'Donoghue (2006), or as fear of future consequences of e, as in Caplin (2003). More positive μ can also represent a moral subsidy for reducing e. To model a moral subsidy, imagine that consumers receive utility μ for every unit of e *not* consumed, up to m_s, where $m_s > e$. Moral utility is then $M = \mu(m_s - e)$, which equals $m - \mu e$ when we set $m = \mu m_s$. This framework can also allow moral utility to depend on consumption relative to a social norm s: if $M = m_s - \mu(e - s)$, this equals $m - \mu e$ when we set $m = m_s + \mu s$. m also captures any "windfall" utility change, if recipients like or dislike the nudge regardless of e.

Let the vector $\theta = \{y, \alpha, \gamma, m, \mu\}$ summarize all factors that vary across consumers. We assume that utility is quasilinear in x, so $\hat{f}' > 0$, $\hat{f}'' < 0$, $\hat{f}'(0) = \infty$, and the consumer maximizes

$$\max_{x,e} \hat{U}(\theta) = x + \hat{f}(e;\alpha,\gamma) + m - \mu e, \tag{1}$$

subject to budget constraint

$$y - T \geq x + ep_e. \tag{2}$$

Consumers' equilibrium choice of e, denoted $\tilde{e}(\theta)$, is determined by the following first-order condition:

$$\hat{f}'(\tilde{e};\alpha,\gamma) - \mu = p_e. \tag{3}$$

This equation shows that increasing the moral price μ can have the same effect on behavior as increasing the price p_e. However, we discuss below how a price increase vs. a moral price increase are very different from a welfare perspective.

Two market failures can cause equilibrium $\tilde{e}(\theta)$ to differ from the social optimum. First, γ (imperfect information or other factors) affects choice but not experienced utility — in other words, consumers choose based on $\hat{f}(e;\alpha,\gamma)$ instead of $f(e;\alpha)$. Second, price p_e may differ from social marginal cost $c_e + \phi_e$ because of the externality ϕ_e and markup π_e. In the first best, $p_e = c_e + \phi_e$ and the consumer would maximize experienced utility, which is

$$U(\theta) = x + f(e;\alpha) + m - \mu e. \tag{4}$$

I.B Nudges

The policymaker can implement a nudge at cost C_n per consumer and maintains a balanced budget using lump-sum tax $T = C_n$. We formalize the nudge as a binary instrument $n \in \{1, 0\}$ that changes consumers' γ, m, and μ. Specifically, each consumer has possibly different potential outcomes θ_n for $n = 0$ vs. $n = 1$, in which γ, m, and μ could differ. We define $\Theta = \{\theta_0, \theta_1\}$ and let $F(\Theta)$ denote its distribution. In words, a nudge provides information, reduces bias, and/or persuades people by activating moral utility. This is intended to be consistent with the practical examples of Thaler and Sunstein (2008), and it is closely analogous to the formal definition in Farhi and Gabaix (2015).

I.C Private and Social Welfare Effects of Nudges

We define "pre-tax consumer welfare" as $V(\theta_n) = U(\theta_n) + T$, and we use Δ to represent effects of a nudge, e.g. $\Delta V \equiv V(\theta_1) - V(\theta_0)$. The effect of the nudge on pre-tax consumer welfare is

$$\Delta V = -\Delta\tilde{e} \cdot p_e + \Delta f + \Delta M. \tag{5}$$

Social welfare is consumer welfare plus profits minus the externality:

$$W(n) = \int U(\theta_n) + (\pi_e - \phi_e)\tilde{e}(\theta_n)\ dF(\Theta). \tag{6}$$

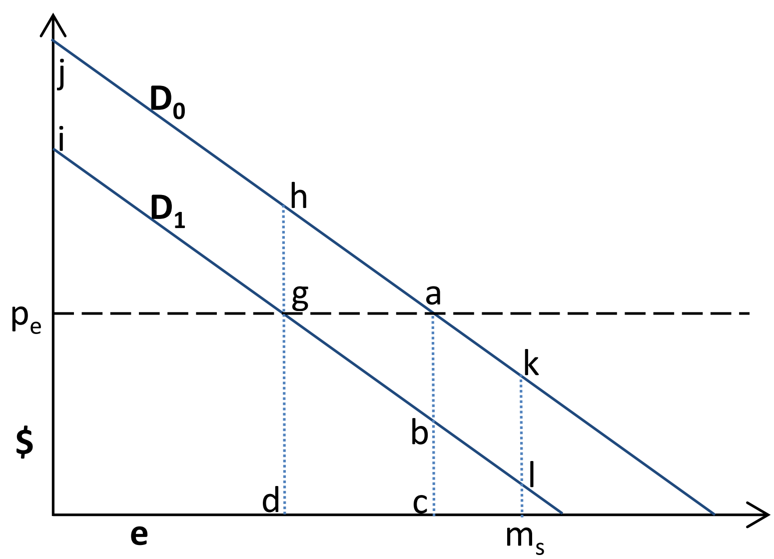
The effect of the nudge on social welfare is

$$\Delta W = \int \Delta V - C_n + (\pi_e - \phi_e)\Delta\tilde{e} \; dF(\Theta). \tag{7}$$

The first term in Equation (7) reflects the net benefit to consumers, ignoring the fact that they must pay for the nudge through the lump-sum tax. The second term C_n then accounts for the cost of the nudge. The final term reflects the change in the pricing distortion.

Nudges with the same effect on behavior $\Delta\tilde{e}$ and the same cost C_n, and thus the same cost effectiveness, can have very different effects on consumer welfare, and thus very different social welfare effects. Figure 1 helps present several distinct mechanisms through which demand could shift from D_0 to D_1, giving the same $\Delta\tilde{e} < 0$ as the equilibrium shifts from point a to point g. First, imagine that there is no moral utility, and the nudge only sets $\hat{f} = f$, i.e. it only provides information or eliminates bias. In this example, D_0 represents perceived $\hat{f}$, while D_1 represents true f. The nudge saves consumers money $-\Delta\tilde{e} \cdot p_e$ (rectangle acdg), which is only partially offset by reduction in consumption utility $f(e;\alpha)$ (trapezoid bcdg). To a first order approximation, the nudge generates $\Delta V \approx -\frac{1}{2}\frac{(\Delta\tilde{e})^2}{de/dp_e} > 0$; that is, it eliminates deadweight loss triangle abg.

Figure 1: **Illustrating the Effects of a Nudge on Consumer Welfare**



Now imagine that $\hat{f} = f$ without the nudge, and the nudge only raises the moral price from $\mu_0 = 0$ to μ_1, generating the same $\Delta\tilde{e}$. In this example, D_0 reflects consumption utility $f(e;\alpha)$, both with and without the nudge. As in the first example, this saves consumers money $-\Delta\tilde{e} \cdot p_e$, but this is outweighed by consumption utility loss shown by trapezoid acdh. In addition, moral utility M decreases by $\mu_1 \tilde{e}(\theta_1)$, which is area ghji. In sum, the moral tax reduces consumer welfare by the same amount as a standard tax: $\Delta V = \frac{1}{2}\frac{(\Delta\tilde{e})^2}{de/dp_e} - \mu_1 \tilde{e}(\theta_1) < 0$, i.e. trapezoid agij. Unlike a standard tax, however, the moral tax does not generate revenues — it simply reduces utility. The welfare effect is negative even if the first-best $\tilde{e}$ is achieved.

Alternatively, the nudge could be a moral subsidy on every unit of e not consumed up to m_s. In this case, consumer welfare would change by $\Delta V = \frac{1}{2}\frac{(\Delta\tilde{e})^2}{de/dp_e} + \mu_1(m_s - \tilde{e}(\theta_1)) > 0$, i.e. trapezoid aklg. More broadly, the nudge can have unbounded positive or negative effects on ΔV unless further restrictions are placed on m.

This discussion highlights how traditional evaluation metrics can be misleading guides for policy decisions: large behavior change $\Delta\tilde{e}$ and low implementation cost C_n are neither necessary nor sufficient for a nudge to increase welfare.

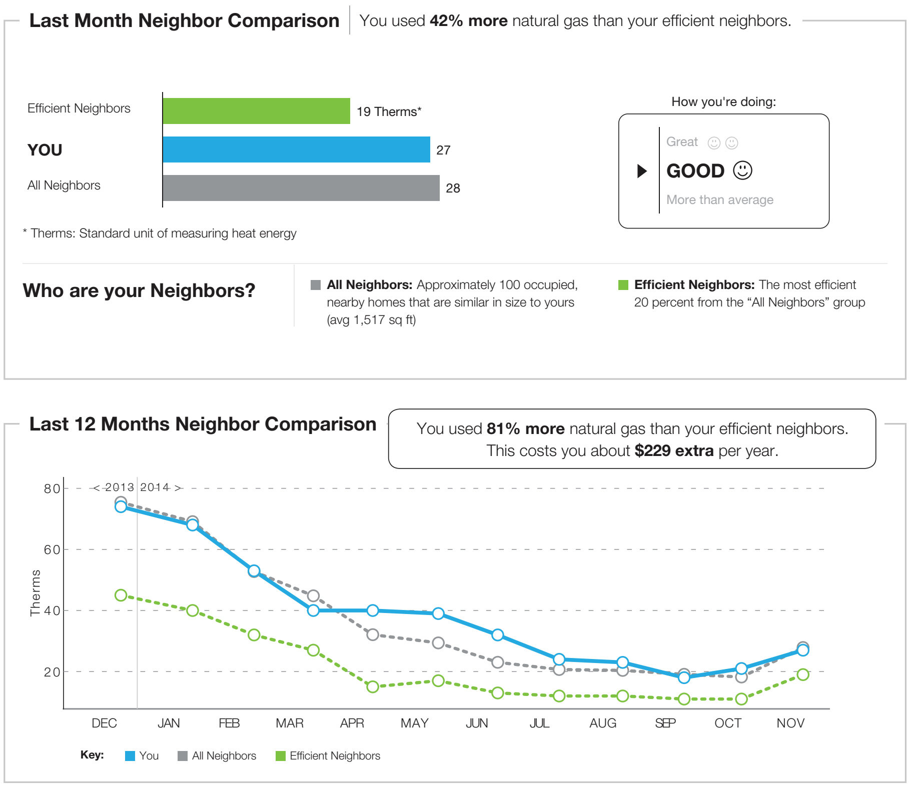
I.D Estimation

In the remainder of the paper, we estimate Equation (7) for a specific nudge: Home Energy Reports. We estimate the change in energy use $\Delta\tilde{e}$ by implementing HERs as a randomized control trial. We use outside estimates of energy use externalities ϕ_e, and we estimate markup π_e and nudge cost C_n from pricing and cost data. To estimate the change in consumer welfare ΔV, we elicit willingness-to-pay for the nudge. In doing this, we assume that our experimental design correctly elicits WTP and that consumers are "sophisticated" in the sense that their WTP for the nudge equals its true effect on their welfare. Sections II-IV present evidence on the plausibility of this assumption, and we formalize it before performing the welfare analysis in Section V.

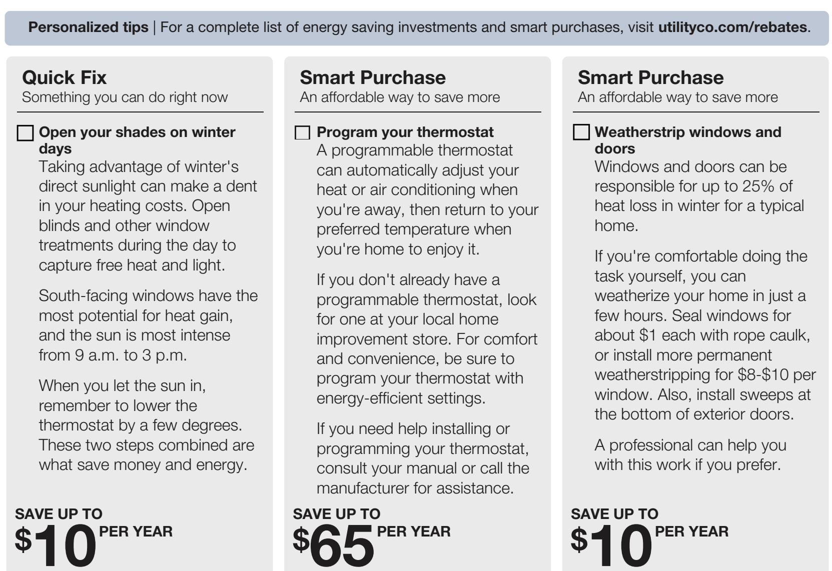
II Experimental Design

The Opower Home Energy Report is a one-page letter (front and back) with two key features illustrated in Figure 2. The Social Comparison Module in Panel (a) compares a household's energy use to that of its 100 geographically nearest neighbors in similar house sizes whose energy use meters were read on approximately the same date. In the neighbor comparison graphs, "All Neighbors" refers to the mean of the neighbor distribution, while "Efficient Neighbors" refers to the 20th percentile. To the right of the three-bar neighbor comparison graph is a box presenting "injunctive norms" intended to signal virtuous behavior (Schultz *et al.* 2007): consumers earn one smiley face for using less than their mean neighbor and two smiley faces for using less than their Efficient Neighbors. The Action Steps Module in Panel (b) gives energy conservation tips; these suggestions are tailored to each household based on past usage patterns. The HERs are thus designed to both provide information and activate "moral utility."

Figure 2: **The Opower Home Energy Report**



(a) **Social Comparison Module**



Personalized tips | For a complete list of energy saving investments and smart purchases, visit **utilityco.com/rebates**.

Quick Fix
Something you can do right now

Open your shades on winter days
Taking advantage of winter's direct sunlight can make a dent in your heating costs. Open blinds and other window treatments during the day to capture free heat and light.

South-facing windows have the most potential for heat gain, and the sun is most intense from 9 a.m. to 3 p.m.

When you let the sun in, remember to lower the thermostat by a few degrees. These two steps combined are what save money and energy.

SAVE UP TO $10 PER YEAR

Smart Purchase
An affordable way to save more

Program your thermostat
A programmable thermostat can automatically adjust your heat or air conditioning when you're away, then return to your preferred temperature when you're home to enjoy it.

If you don't already have a programmable thermostat, look for one at your local home improvement store. For comfort and convenience, be sure to program your thermostat with energy-efficient settings.

If you need help installing or programming your thermostat, consult your manual or call the manufacturer for assistance.

SAVE UP TO $65 PER YEAR

Smart Purchase
An affordable way to save more

Weatherstrip windows and doors
Windows and doors can be responsible for up to 25% of heat loss in winter for a typical home.

If you're comfortable doing the task yourself, you can weatherize your home in just a few hours. Seal windows for about $1 each with rope caulk, or install more permanent weatherstripping for $8-$10 per window. Also, install sweeps at the bottom of exterior doors.

A professional can help you with this work if you prefer.

SAVE UP TO $10 PER YEAR

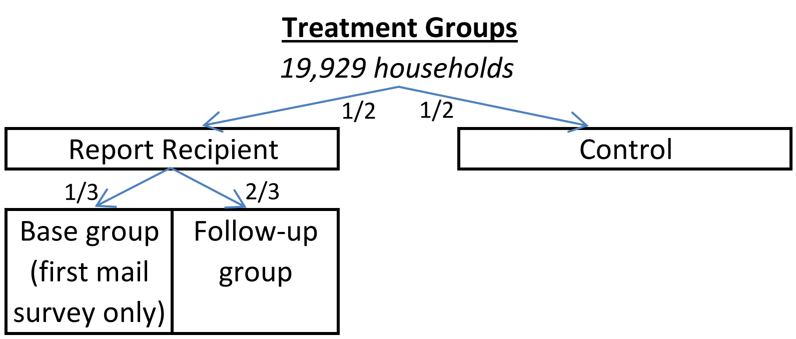
(b) **Action Steps Module**

Notes: The Home Energy Report is a one-page (front and back) letter including the Social Comparison Module in Panel (a) and the Action Steps Module in Panel (b).

As of mid-2015, Opower had implemented HER programs at 95 utilities in nine countries. We focus on one program at Central Hudson Gas and Electric, which serves 300,000 electric customers and 78,000 natural gas customers in eight New York counties. Like 23 other states, New York has an Energy Efficiency Portfolio Standard, which requires that utilities cause consumers to reduce energy demand by a specified amount each year (ACEEE 2015). As part of compliance with the standard, Central Hudson had already planned a multi-year Home Energy Report program for residential natural gas customers. Central Hudson and Opower agreed to modify the program to incorporate this study.

Figure 3 summarizes the experimental design. Starting with an eligible population of 19,929 households, Opower randomly assigned half to treatment and half to control. The treatment group received up to four HERs during the "heating season" from late October 2014 through late April 2015. Central Hudson employees read each household's natural gas meter every two months, and an HER was generated and mailed shortly after each meter read in order to provide timely and relevant information. Some households received fewer than four HERs for standard technical reasons such as not having enough neighbors to generate valid comparisons. Like almost all other HER programs, this is an "opt out" program, so households continue to receive HERs unless they contact the utility to opt out. Sixteen households had opted out by September 2015, and thus did not receive reports in the program's second year. Households also stop receiving HERs if they move addresses.

Figure 3: **Experimental Design**



Process

1. **Four Reports** (October 2014-April 2015)
2. **First mail survey** (in final Report)
3. **Follow-up mail survey** (own envelope, May 2015)
4. **Phone survey** (June-August 2015)
5. **Next four Reports and/or check** (October 2015-April 2016)

Opower included our one-page survey and postage-paid Business Reply Mail return envelope in the same envelope as the final HER of the 2014-2015 heating season. Figure 4 reproduces the survey. The first seven questions were a multiple price list (MPL) that asked recipients to trade off four more HERs with checks for different amounts of money. The responses can be used to bound willingness-to-pay. For example, consumers who prefer "four more Home Energy Reports plus a $9 check" instead of "a $10 check" value the four HERs at $1 or more. Consumers who prefer "a $10 check" instead of "four more Home Energy Reports plus a $5 check" value the four HERs at $5 or less. A consumer who answered as in these two examples therefore has WTP between $1 and $5.

The survey letters included three variations intended to remind consumers of different features of the HERs. Figure 4 was the "Standard" version. In the "Comparison" version, the sentence "Remember that **Home Energy Reports compare your energy use to your neighbors' use**" was added after "we want to know what you think about them" in the introductory paragraph. In the Environmental version, "Remember that **Home Energy Reports help you to reduce your environmental impact**" was added in that same place.

In typical HER programs, including this one, a few consumers dislike HERs enough to take the time to opt out. If time has any positive value, this implies a strictly negative WTP for HERs for these consumers. To correctly measure the distribution of WTP in such an opt-out program, it is thus necessary to allow consumers to reveal negative WTP. We designed the MPL to do this, by asking consumers to choose between "four more HERs plus a $10 check" and checks of less than $10. For example, consumers who choose "a check for $9" instead of "four more HERs plus a $10 check" are giving up $1 to *not* receive four more HERs, meaning that their WTP must be no greater than $-1. Answers to the seven-question MPL place a respondent's WTP into eight ranges, which are symmetric about zero: $(-\infty, -9], [-9, -5], [-5, -1], [-1, 0], [0, 1], [1, 5], [5, 9]$, and $[9, \infty)$.

The survey's final question was, "Think back to when you received your first Home Energy Report. Did you find that you used more or less energy than you thought?" This measures the extent to which HERs caused consumers to update beliefs about relative usage.

A randomly-selected two-thirds of HER recipients were sent a follow-up mail survey on May 26th, 2015. We call this group the "follow-up group," while the other one-third is the "base group." This was not part of an HER and was sent through a separate vendor, so the outbound envelope had a different originating address than the HERs. The survey and Business Reply Mail return envelope were identical to the first mail survey.

Figure 4: **Mail Survey**

People. Power. Possibilities.
Central Hudson
A FORTIS COMPANY

Account Number: xxxx-xxxx-xx-x

Tell us what you think — and earn a check for up to $10!

Central Hudson has been sending you Home Energy Reports since last fall, and we want to know what you think about them. Would you take a moment to complete the survey below? For each question, please fill in one box with your answer.

What happens next?

1. When you're finished, mail the survey back to us in the enclosed prepaid envelope.
2. We will use a lottery to draw one of the first seven questions, and we'll mail you what you chose in that question — either a check or a check plus four more Home Energy Reports.

Thank you!

Your participation will help us make these reports even more useful for you. If you have any questions, please email us at **HERSurvey@cenhud.com** or call **(845) 486-5221**.

1.	Which would you prefer?	4 + $10 — 4 more Home Energy Reports PLUS a **$10** check	☐	OR	☐	$1 — A **$1** check
2.	Which would you prefer?	4 + $10 — 4 more Home Energy Reports PLUS a **$10** check	☐	OR	☐	$5 — A **$5** check
3.	Which would you prefer?	4 + $10 — 4 more Home Energy Reports PLUS a **$10** check	☐	OR	☐	$9 — A **$9** check
4.	Which would you prefer?	4 + $10 — 4 more Home Energy Reports PLUS a **$10** check	☐	OR	☐	$10 — A **$10** check
5.	Which would you prefer?	4 + $9 — 4 more Home Energy Reports PLUS a **$9** check	☐	OR	☐	$10 — A **$10** check
6.	Which would you prefer?	4 + $5 — 4 more Home Energy Reports PLUS a **$5** check	☐	OR	☐	$10 — A **$10** check
7.	Which would you prefer?	4 + $1 — 4 more Home Energy Reports PLUS a **$1** check	☐	OR	☐	$10 — A **$10** check

8.	Think back to when you received your first Home Energy Report. Did you find that you used more or less energy than you thought?	☐ Much less	☐ Somewhat less	☐ About what I thought	☐ Somewhat more	☐ Much more

CHGE_0009_WELCOME_LETTER_SURVEYA

In June, July, and early August 2015, an independent survey research firm surveyed the entire HER treatment group by phone. Each phone number was called up to eight times until the household completed the survey or declined to participate. The beginning of the phone survey parallels the mail survey, except that we used a three-question version of the same MPL that dynamically eliminated questions whose answers were implied by earlier answers.[8] We then asked a belief update question parallel to the mail survey and a series of additional questions to elicit beliefs about energy cost savings and qualitative evaluations of the HERs. Appendix A presents the full phone survey questionnaire. A condensed version is:

1. [Multiple price list]
2. Did your first Report say you were using more or less than you thought?
3. Do you think that receiving four more Reports this fall and winter would help you reduce your natural gas use by even a small amount?
    (a) *If Yes:* How much money do you think you would save on your natural gas bills if you receive four more Reports?
4. How much money do you think the average household has saved since last fall?
5. How would you like the Reports if they didn't have the neighbor comparison graph?
6. Do the Reports make you feel inspired, pressured, neither, or both?
7. Do the Reports make you feel proud, guilty, neither, or both?
8. Do you agree/disagree with: "The Reports gave useful information that helped me conserve energy."
9. Do you have any other comments about the Reports that you'd like to share?

If the phone survey respondent reported that he or she had already returned the mail survey, the phone survey skipped directly to question 3. Questions 6 and 7 were designed to measure whether the HERs tend to generate positive or negative affect, to provide suggestive evidence on whether HERs affect "moral utility" or act as a psychological tax or subsidy. The words "inspired," "proud," and "guilty," were drawn from the Positive and Negative Affect Schedule (Watson, Clark, and Tellegan 1988), a standard measure in psychology. We added the word "pressured" because we hypothesized that it might be relevant in this context.

[8] We began by asking question 4 from the mail survey. If the respondent preferred HERs+a $10 check, we asked question 6. If the respondent preferred HERs+a $5 check on question 6, we asked question 7, whereas if the respondent preferred a $10 check on question 6, we asked question 5. If the respondent preferred a $10 check on question 4, we asked question 2. If the respondent preferred HERs+a $10 check on question 2, we asked question 3, whereas if the respondent preferred a $5 check on question 2, we asked question 1.

Both the mail and phone MPLs clearly stated at the outset that they were incentive-compatible. The mail survey stated, “We will use a lottery to draw one of the first seven questions, and we’ll mail you what you chose in that question.” The phone survey script stated, “These are real questions: Central Hudson will use a lottery to pick one question and will actually mail you what you chose.” Once all survey responses were collected, we randomly selected one of the seven MPL questions for each respondent, and the respondent received what he or she had chosen in that question: either a check from Central Hudson or a check plus four more HERs in the program’s second winter.[9] As a result of their survey responses, 146 households were dropped from the program’s second year, including all households who responded on all seven MPL questions that they preferred not to receive HERs. The consequences of non-response were not communicated to households in the survey or otherwise. In practice, households that did not respond to the survey did not receive a check and did receive HERs over the 2015-2016 winter heating season.

Our design elicits WTP for the program’s second year, and thus allows a welfare evaluation only of the second year. Why study only the second year of a program? First, the revealed preference approach that is central to the paper — that is, taking WTP seriously as a measure of consumer welfare — is much more plausible when consumers have experience with the nudge they are evaluating. Second, while most utilities that currently send HERs to households do so for multiple years, our analysis helps address an active debate about how long to continue treating the same households with HERs.

Relatedly, one might wonder whether the first few HERs provide the bulk of informational or motivational benefits of an HER program. Perhaps WTP would be much higher for the first HER or first few HERs? It is not clear that this intuition is correct. Allcott and Rogers (2014) show that continued HERs cause incremental conservation even after receiving eight to 24 reports over two years, implying that there is additional value well after the first year. The fact that additional HERs continue to affect energy consumption likely arises both because additional HERs are a motivational reminder and because they provide new information. For example, about half of households see their ranking relative to their mean neighbor or Efficient Neighbors change across reports over a given year of the Central Hudson program we study.[10] Furthermore, the energy conservation tips change with every report. It is thus unlikely that the first few HERs provide the bulk of the benefits, and it is not obvious the extent to which consumers would value the first few HERs vs. later HERs differently. This discussion highlights why it is both interesting and relevant to evaluate the program’s second year.

[9]Because Central Hudson needed to continue sending HERs to most households to satisfy regulatory requirements under the Energy Efficiency Portfolio Standard, we placed 98.6 percent probability on the first question, on which 94 percent of respondents chose HERs. The remaining six questions were each selected with 0.2 percent probability.

[10]These changes occur largely because of standard month-to-month variation in household energy use, not due to conservation actions induced by the HERs. The average treatment effect of HERs is very small relative to the standard within-household and between-household variation.

III Data

There are five data sources: the utility's natural gas bill data, neighbor comparisons, customer demographic data, mail surveys, and phone surveys.

Central Hudson reads customers' natural gas meters on very regular bi-monthly cycles: 94 percent of billing period durations are between 55 and 70 days. Central Hudson measures natural gas use in hundred cubic feet (ccf). As we discuss further in Section V, Central Hudson uses a decreasing block tariff, and the average marginal retail price is \$0.99/ccf during the program's first winter and \$0.80/ccf during the program's second winter. We observe natural gas use for each household in treatment or control for all meter read dates between July 1, 2013, and September 23, 2016. Households that move do not appear in the energy bill data (and do not receive HERs) after they move.

The key feature of the Social Comparison Module in Panel (a) of Figure 2 is a bar graph comparing the household's use on its previous bill to the mean and 20th percentile of the distribution of neighbors' use. We observe that mean and 20th percentile for all HERs, including HERs that control group households would have received.

Table 1 presents demographic variable summary statistics. All variables other than baseline use and hybrid auto share are from a demographic data vendor and are matched to the utility account holder. These variables are from a combination of public records, survey responses, online and offline purchases, and statistical predictions, and most are likely measured with error. Some households in the population could not be matched to demographic data, in which case we use mean imputation. We made every effort to acquire the best data possible, because measurement error and missing data make our inverse probability weights and prediction algorithm less effective.

These data may overestimate household income, but the population is relatively wealthy: according to Census data, the mean household is in a census block group with median household income of \$64,000. Green consumer is a binary measure of environmentalism based on income, age, and purchases of organic food, energy efficient appliances, and environmentally responsible brands. Wildlife donor is an indicator for whether the consumer has contributed to animal or wildlife causes. These two variables could proxy for environmentalism and thus interest in energy conservation. Home improvement is an indicator for home improvement transactions or product registrations, which could proxy for interest in making energy-saving improvements in response to HERs.

Our household covariates, denoted $\mathbf{X}$, are these same variables, except that we take natural logs of income, net worth, house value, age, and house age.[11] Appendix Tables A1 and A2 confirm that these covariates are not more correlated with HER recipient group or survey group assignment than would be expected by chance.

[11] Some households have negative net worth, so before taking the natural log, we add a constant to all observations such that the minimum value is \$1.

Table 1: **Demographic Variable Summary Statistics**

Variable	Non-missing Observations	Mean	Standard Deviation	Min	Max
Baseline use (ccf/day)	19,921	2.12	1.67	0	19.1
Income ($000s)	15,557	94.4	81.9	10	450
Net worth ($000s)	15,557	195	288	-30	1500
House value ($000s)	16,741	271	173	18	2527
Education (years)	19,475	13.6	2.44	10	18
Male	16,811	0.51	0.50	0	1
Age	17,282	50.7	16.1	19	99
Retired	16,728	0.04	0.20	0	1
Married	15,406	0.59	0.49	0	1
Rent	17,561	0.30	0.46	0	1
Single family home	17,734	0.68	0.46	0	1
House age	14,885	59.7	40.2	0	115
Democrat	18,080	0.16	0.55	-1	1
Hybrid auto share	19,728	1.03	2.78	0	18.2
Green consumer	18,883	0.15	0.35	0	1
Wildlife donor	16,728	0.06	0.24	0	1
Profit score	19,784	0.00	1.00	-1.65	2.09
Buyer score	14,967	0.00	1.00	-2.03	1.47
Mail responder	17,734	0.47	0.46	0	1
Home improvement	16,728	0.13	0.33	0	1

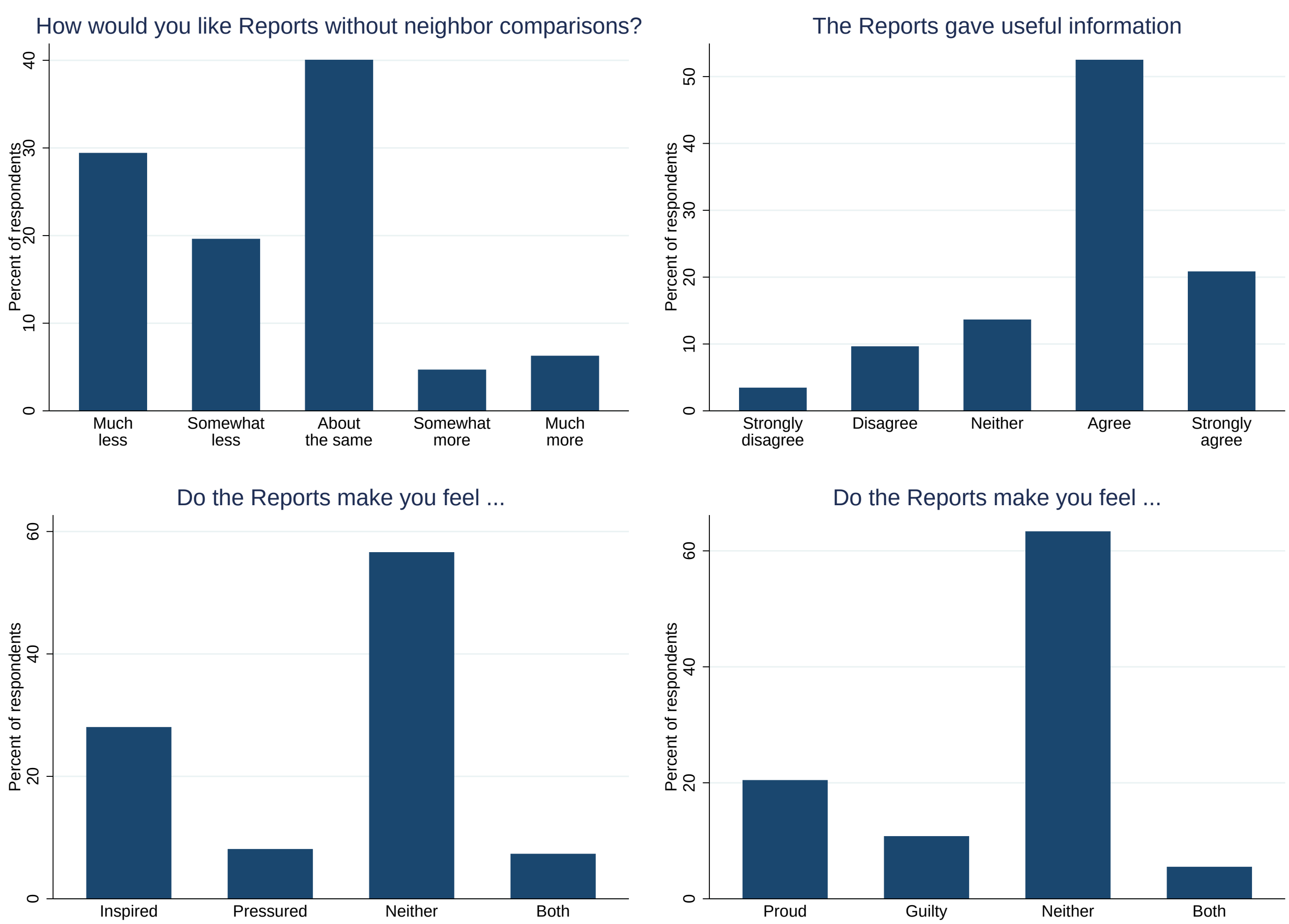
Notes: This table summarizes the demographic variables. Baseline use is mean natural gas use (in hundred cubic feet per day) between July 2013 and June 2014. Hybrid auto share is the share (from 0-100) of vehicles registered in the census tract in 2013 that were hybrids. All other variables are from a demographic data provider. Education is top-coded at 18 years for people with any graduate degree. Democrat takes value 1 for Democrats and -1 for Republicans. Green consumer is a binary measure of environmentalism based on income, age, and purchases of organic food, energy efficient appliances, and environmentally responsible brands. Wildlife donor is an indicator for whether the consumer has contributed to animal or wildlife causes. Profit score and buyer score measure the consumer's likelihood of paying debts and making purchases; we normalize both to mean 0, standard deviation 1. Mail responder is an indicator for whether anyone in the household has purchased by direct mail. Home improvement is an indicator for home improvement transactions or product registrations.

Table 2 summarizes response rates. Households that were sent the follow-up mail survey were more than twice as likely to respond as base group households, which only received the survey in their final Home Energy Report. 899 households (9.5 percent of households that were surveyed) responded to the mail survey, and 1690 households (17.9 percent) completed the phone survey. 2312 households (24.5 percent) responded to one or both surveys. This overall response rate is lower than official government surveys such as the Current Population Survey, but higher than most other non-government surveys, and indeed higher than we expected. We discuss our strategy for extrapolating to the population of non-respondents in Section IV.B.

Table 2: **Survey Response Rates**

	Response rate (%)
Mail survey	9.5
Base mail survey group	4.5
Follow-up mail survey group	12.0
Phone survey	17.9
Both mail and phone surveys	2.9
Mail and/or phone surveys	24.5

Figure 5: **Qualitative Evaluations of Home Energy Reports**



Notes: This figure presents qualitative evaluations of Home Energy Reports from the phone survey.

Figure 5 summarizes responses to the qualitative evaluations of the HERs from the phone survey. Forty-nine percent would like HERs less if the neighbor comparisons were removed, against only 11 percent who would like them more. Seventy-three percent of respondents agree or strongly agree that HERs provide useful information. For most respondents, the HERs did not generate positive or negative affect: 57 percent said that the HERs made them feel neither "inspired" nor "pressured," and 63 percent said that HERs made them feel neither "proud" nor "guilty." When the HERs did induce some positive or negative affect, it was much more likely to be positive (inspired or proud) instead of negative (pressured or guilty). These qualitative results suggest that most people "like" HERs, i.e. that they would want HERs if they were free.

III.A Constructing Willingness-to-Pay

Complete and internally-consistent responses to the multiple price list allow us to place each respondent's willingness-to-pay into one of eight ranges. For simplicity, we assign one unique WTP for each range. For the six interior ranges, we assign the mean of the endpoints. For example, we assign a WTP of \$-3 for all responses on $[-5, -1]$ and a WTP of \$0.50 for all responses on $[0, 1]$. For the unbounded ranges, i.e. WTP less than \$-9 or greater than \$9, we assume that the conditional distribution of WTP is triangular, with initial density equal to the average density on the adjacent range.[12] This gives \$14.45 and \$-12.31, respectively, as the conditional mean WTPs on $[9, \infty)$ and $(-\infty - 9]$. We also present results under alternative assumptions.

For the 2.9 percent of households that responded to both the phone and mail surveys, we use the phone survey WTP in order to be consistent with the phone survey's additional qualitative questions. For the 87 households that returned more than one mail survey with valid WTP, we use the first survey we received. These two sets of households with multiple survey responses provide an opportunity to show the stability of our WTP elicitations within a household, both within the same MPL format (i.e., mail vs. mail) and across different formats (i.e., mail vs. phone). We therefore give them special attention in Section III.B.

III.B Do the Surveys Correctly Measure Willingness-to-Pay?

While standard in academic economics and lab settings, multiple price list surveys are relatively unusual in field settings. One concern in designing this study was that respondents would not understand the MPL, rendering WTP estimates noisy or meaningless. We devoted substantial effort to designing easily-understandable surveys and piloting the mail and phone instruments.

[12] For example, the density on $[5, 9]$ is 2.49 percent of respondents per dollar, and the mass above \$9 is 20.30 percent of respondents. We assume that this 20.30 percent of respondents is distributed triangular on $[9, \infty)$, with maximum density of 2.49 percent per dollar at \$9 decreasing to zero density above some upper bound. This gives an upper bound of \$25.34. The mean of WTP on $[9, \infty)$ is thus \$14.45. The mean WTP on $(-\infty - 9]$ is determined by an analogous calculation, given that the density on $[-9, -5]$ is 1.27 percent per dollar and the mass below \$-9 is 6.32 percent. We will also present welfare estimates under alternative assumptions.

Table 3 shows that the vast majority of returned mail surveys were filled out in a way that allows us to construct a valid WTP. 14.7 percent of mail surveys were incomplete, usually because the respondent answered only one of the seven questions. 11.1 percent of phone respondents heard the introduction to the MPL but terminated the interview before completing all three questions. Only 2.1 percent of mail MPL responses were complete and internally inconsistent. Three mail MPL responses (0.3 percent) were both incomplete and internally inconsistent. Because the phone MPL was shortened by not asking questions whose answers were implied by previous responses, there was no opportunity to be internally inconsistent on the phone survey. These figures suggest that consumers generally understood the MPL and gave meaningful answers.[13]

Table 3: **Multiple Price List Response Statistics**

	Mail	Phone
Percent incomplete	14.7	11.1
Percent complete and internally inconsistent	2.1	N/A
Percent complete and internally consistent	83.2	88.9

In addition, WTP relates to other survey responses in expected ways. WTP is very strongly correlated with the qualitative assessments of the HERs from questions 3-9 of the phone survey. As would be expected, WTP is strongly positively correlated with reporting that future HERs would save them more money (question 3), feeling inspired and proud (questions 6 and 7), agreeing that HERs give useful information (question 8), and with positive additional comments about the HERs (question 9).[14] Also as expected, WTP is strongly negatively correlated with preferring that HERs not have neighbor comparisons (question 5) and with feeling pressured (question 6). The only result that we did not expect was that feeling guilty is positively associated with WTP, but the relationship is not significant after conditioning on predicted savings, which suggests that consumers do not like guilt *per se* — they like guilt only because it helps them reduce expenditures. See Appendix Table A4 for formal results.

As we detail below, 34 percent of respondents reported negative WTP. In Appendix Table A5, we confirm that negative WTP is strongly associated with the same set of qualitative assessments in

[13]We listened to about 25 early phone survey interviews. Because the MPL questions are unusual, respondents would sometimes pause to process the first question but would then provide a considered answer to that and the next two MPL questions.

[14]456 phone survey respondents offered comments in response to our open-ended question 9. Of these, 170 were positive, such as "They're terrific. I like the way they're laid out and easy to understand," and "I think you did it right. It has all the information owners need. I think it's an excellent idea," and "Detailed and a great thing. Helps me monitor my usage." 213 were neutral on the HERs, often including complaints about high energy prices. 73 were negative, such as "I do not understand it; it does not make sense," and "It's a waste of paper. If they did not send those reports maybe they could lower the delivery charges," and "The money would be better spent reducing the cost of energy rather than sending the reports."

expected ways. Furthermore, all six households that opted out and also responded to the survey had negative WTP. These strong correlations build confidence that both the MPL and the qualitative questions elicited meaningful responses.

87 households returned more than one mail survey with valid WTP. These could have been filled out by different people in the same household, or by one person who wanted to ensure that his or her response was received. Thus, one might expect responses to be correlated, but not perfectly correlated. WTP is indeed very highly correlated across the two responses within these households, implying that people understood the mail MPL well enough that responses were consistent within a person or household. See Appendix Table A6 for formal results.

277 households responded to both the phone and mail surveys, of which 224 have valid WTP from both surveys and 259 responded to the belief update question on both surveys. Because the phone survey called for skipping these questions if the respondent reported already returning the mail survey, it seems likely that duplicate mail and phone responses came from different people in the same household. Here again, one might thus expect responses to be correlated, but not perfectly correlated. Appendix Table A6 confirms this: WTP, an indicator for negative WTP, and belief updates are all strongly correlated within household across the mail and phone surveys. WTP and answers to the belief update question within household are almost equally strongly correlated across the mail and phone surveys, which suggests that the MPL questions to elicit WTP were no more confusing or cognitively demanding than the belief update question, where responses were on the familiar Likert scale. Across the 224 households with valid WTP from both surveys, the mean WTP and the share of negative WTPs are almost exactly identical between the mail and phone surveys. This implies that the survey formats did not generate differential biases in mean WTP.

In general, these results suggest that respondents understood the MPLs and that the survey instruments correctly elicited WTP. Here we address some remaining concerns about how well our MPL measure elicited WTP.

First, time discounting could affect WTP. For example, if respondents have annual discount rates of six percent and thought that checks would arrive six months before the HERs' benefits, their WTP would be about three percent lower than if they thought that checks would arrive at the same time as the benefits. Such a small difference would not be enough to meaningfully affect the welfare calculations below. Conceptually, we want all components of welfare to be discounted to the time at which the implementation costs are incurred for the second year of HERs. In practice, the checks were sent in December 2015, although we intentionally did not say this on the survey because we did not want to make time discounting salient.

Second, WTP might be lower if paying out of pocket instead of trading off against an unexpected windfall from a check. If this results from a behavioral bias, it is not obvious what WTP to respect for welfare analysis.

Third, WTP might be higher with per-month subscription pricing instead of a one-time check. Because the monetary amounts are small and respondents pay for HERs from a future windfall instead of from their existing funds, it is unlikely that credit constraints could explain a preference for subscription payments. If WTP differs with subscription pricing vs. a one-time check due to a behavioral bias such as focusing bias (Koszegi and Szeidl 2013), it is again not clear what WTP to respect for welfare analysis.

Fourth, Beauchamp *et al.* (2015) demonstrate a compromise effect in multiple price lists — that is, that people tend to favor the middle option of an MPL. Because our phone MPL questions were given sequentially, however, this concern does not apply to our phone MPL. As mentioned above, the mean WTPs for the phone and mail MPLs are indistinguishable for households that responded to both surveys, suggesting that the mail MPL is also unaffected by a compromise effect.

Models of contextual inference, such as Kamenica (2008), suggest two reasons why our mail MPL would not be biased by a compromise effect. First, there is little imperfect information: the MPL asks simple questions about a familiar good and, unlike Beauchamp *et al.* (2015), there are no risky prospects that could increase cognitive complexity. Second, consumers were unlikely to infer that they are "middlebrow" relative to the bounds of the MPL: the distribution of responses suggests that the first two questions had relatively obvious answers (very few people were willing to pay significant amounts to avoid HERs) while the last two questions did not (many people were in the top two WTP ranges).

IV Empirical Analysis

In this section, we estimate parameters needed for the welfare analysis prescribed by Equation (7). We begin by estimating the treatment effects on energy use, which determine the externality benefits and profit losses. We then calculate average WTP, which will be our measure of the consumer welfare effects.

Before estimating treatment effects and calculating WTP, it is important to clarify the target population and time period for which we want the estimates to be relevant. The survey elicits willingness-to-pay for the second year of HERs, and we thus want to evaluate the welfare effects of the program's second year in the population of households that would normally (in the absence of our experiment) receive reports in that year. We denote this target population as $\mathcal{P}_n$. As reported above, 16 households had opted out before the second year began, and another 146 households were dropped from the second year due to their survey responses. $\mathcal{P}_n$ excludes the former 16 but includes the latter 146. We will also present alternative estimates valid for the smaller subset of households that responded to the survey and did not opt out, which we denote as $\mathcal{P}_s$.

At times, constructing estimates for one or both of these target populations requires extrapolation, for example extrapolating WTP from the subset of survey respondents to $\mathcal{P}_n$. Our primary approach in these cases is to use inverse probability weights (IPWs) to re-weight a sample to match a target on observable characteristics. Specifically, we use probit regressions presented in Appendix Table A7 to estimate $\Pr(H_i = 1|\mathbf{X}_i; \mathcal{P})$ using data from target population $\mathcal{P}$, where H_i is an indicator for whether observation i is in the sample, and then construct sample weights $\left[\hat{\Pr}(H_i = 1|\mathbf{X}_i)\right]^{-1}$. Of course, we are not able to correct for unobservable differences between sample and target populations, and we discuss this issue further below.

IV.A Effects on Energy Use

To estimate the effect of Home Energy Reports on energy use, we limit the sample to post-treatment data and control for pre-treatment usage. The first HERs were generated on October 13th, 2014, and first HERs had been generated for 61 percent of households by November 3rd, and 98 percent by December 8th. Post-treatment is defined as any meter read after November 1, 2014.

Y_{it} is household i's average natural gas use (in ccf/day) over the billing period ending on date t, and T_i is an indicator for whether household i was randomly assigned to the initial recipient group. We define S_{st} as indicators for whether date t falls within season of the year s, and we allow treatment effect τ_s to vary by season s. We define the baseline period as the earliest 365 days in the data: July 1, 2013 through June 30, 2014. $\tilde{Y}_{it}$ is "baseline usage" — more specifically, the average daily usage from the meter read in the baseline period that most closely corresponds to billing date t. For example, if t is October 14, 2015, $\tilde{Y}_{it}$ is the average daily usage from the meter read date closest to October 14, 2013. Because meters are read on a very regular bi-monthly basis, we have fairly precise matches that help account for seasonality.[15] ν_m allows separate coefficients on $\tilde{Y}_{it}$ by the month of sample that contains date t, and ω_m is a vector of month of sample indicators. The estimating equation is:

$$Y_{it} = \sum_s \tau_s S_{st} T_i + \nu_m \tilde{Y}_{it} + \omega_m + \varepsilon_{it}. \tag{8}$$

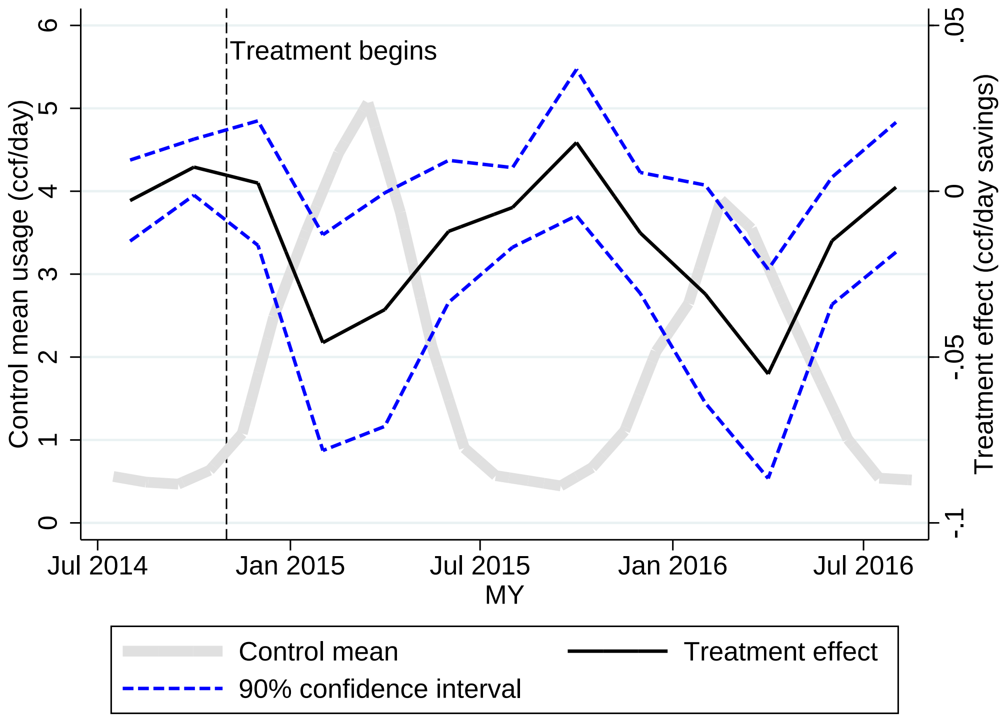
Standard errors are clustered by household to allow for arbitrary serial correlation.

In Figure 6, the thick grey line plots control group mean usage in each month of the sample, illustrating considerable seasonality. Usage is lowest during the bimonthly billing periods ending in July through October, and usage is about five times higher during the bimonthly billing periods ending in November through June. The thin black line and confidence intervals are estimates of treatment effects τ_s, where s indexes each pair of months after the baseline period ends on June 30, 2014. The several months of pre-treatment observations allow us to test for spurious pre-treatment effects, and there are indeed zero statistical effects for meters read in July through October 2014.

[15] Since it is primarily used for heating, natural gas use is highly seasonal, as illustrated in Figure 6. Thus, controlling for seasonal fluctuations is crucial for improving statistical efficiency. Note that estimating in logs and transforming the percent savings back into levels is not a consistent estimator of the level of average savings due to Jensen's Inequality. For this reason, Allcott (2011, 2015) and Allcott and Rogers (2014) estimate effects in levels.

There are also zero statistical effects for meters read in November and December. We then see strong seasonality in the treatment effects: as much as a 0.05 ccf per day reduction in the winter periods, and zero statistical effects in any of the summer billing periods ending in July through October. This seasonality is standard in natural gas energy conservation programs: households cannot conserve much natural gas when they are not using much in the first place.

Figure 6: **Effects of Home Energy Reports on Natural Gas Use**



Notes: This figure presents estimates of Equation (8), allowing the treatment effect to vary across two-month periods. Dependent variable is natural gas use in ccf/day, where "ccf" means hundred cubic feet. For context, the average marginal retail price is \$0.99/ccf during the program's first winter and \$0.80/ccf during the program's second winter. Observations weighted by billing period duration. Robust standard errors, clustered by household.

Table 4 presents estimates of Equation (8). In all columns, we weight each observation of daily usage by the duration of the billing period, which gives average treatment effects in ccf/day. In columns 3 and 4, we multiply this duration weight by additional household weights for extrapolation, as discussed below. As suggested by the graphical results in Figure 6, we estimate separate treatment effects for four seasons $s = \{1, 2, 3, 4\}$: winter (November-June) of 2014-2015, summer (July-October) of 2015, winter of 2015-2016, and summer of 2016, respectively.

Table 4: **Effects of Home Energy Reports on Natural Gas Use**

	(1)	(2)	(3)	(4)
Specification:	OLS	IV	IV	IV
Assigned to treatment ×	-0.0227	-0.0227	-0.0229	-0.0248
winter 2014-2015	(0.0117)*	(0.0117)*	(0.0118)*	(0.0128)*
Assigned to treatment×	0.00635	0.00635	0.00642	0.00856
summer 2015	(0.00803)	(0.00803)	(0.00806)	(0.00933)
Assigned to treatment ×	-0.0263			
winter 2015-2016	(0.0115)**			
Assigned to treatment×	0.00593			
summer 2016	(0.0111)			
2nd-year recipient ×		-0.0268	-0.0270	-0.0315
winter 2015-2016		(0.0117)**	(0.0117)**	(0.0121)***
2nd-year recipient ×		0.00605	0.00597	0.00407
summer 2016		(0.0114)	(0.0114)	(0.0110)
Observations	200,528	200,528	200,528	200,528
R^2	0.853	0.853	0.853	0.859
Weights	Duration	Duration	Duration × IPW for $\mathcal{P}_n$	Duration × IPW for $\mathcal{P}_s$

Notes: This table presents estimates of Equation (8), using post-treatment data only. Dependent variable is natural gas use in hundred cubic feet (ccf) per day. For context, control group sample mean usage is 2.07 ccf/day, and the average marginal retail price is \$0.99/ccf during the program's first winter and \$0.80/ccf during the program's second winter. Columns 2-4 are IV regressions, where we instrument for 2nd-year recipient × winter 2015-2016 and 2nd-year recipient × summer 2016 with Assigned to treatment × winter 2015-2016 and Assigned to treatment × summer 2016. Columns 1 and 2 weight by billing period duration. Column 3 weights by duration times a household weight that matches the compliers to the target population $\mathcal{P}_n$ of treatment group households that did not opt out before the second year. Column 4 weights by duration times a household weight that matches the compliers to the target population $\mathcal{P}_s$ of treatment group households that did not opt out and returned a survey with valid willingness-to-pay. Robust standard errors, clustered by household, in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Column 1 presents intent-to-treat effects: the average effect over time for households *assigned* to the treatment group. As discussed above, we are also interested in estimating the causal impact of the second year of HERs on target populations $\mathcal{P}_n$ and $\mathcal{P}_s$. This requires the additional assumption (which we call the "no persistence" assumption) that HERs delivered during program's first winter do not have persistent effects in the second year.[16] Define R_i as an indicator for whether

[16]As we discuss below in Section V.B, the ideal way to estimate this would be to randomly assign the treatment group to receive either one year or two years of HERs, but this was not feasible. Note that if the "no persistence" assumption does not hold, then treatment effects observed during the second year also reflect effects caused by delivering the first winter of HERs to the entire treatment group. The "no persistence" assumption is consistent with

household i was sent HERs in the program's second year. $R_i = T_i$, except that $R_i = 0$ for the 162 households that opted out or were dropped due to survey responses. Column 2 presents results of an instrumental variables (IV) regression where instead of the second year winter and summer treatment assignment indicators S_3T_i and S_4T_i in Equation (8), we substitute S_3R_i and S_4R_i and then instrument with S_3T_i and S_4T_i. Under the "no persistence" assumption, this IV regression delivers the local average treatment effects of the second year of HERs. Because there are no always-takers (i.e., no households would receive HERs in the absence of the program), $R_i = 1$ is an indicator for being a complier.

In column 3, we re-weight compliers have the same observable characteristics as target population $\mathcal{P}_n$, the households that would normally receive reports in the program's second year.[17] In column 4, we re-weight compliers to match $\mathcal{P}_s$, the subset of households that responded to the survey and did not opt out.[18]

The estimates are almost exactly the same in the first three columns: zero statistical effects in the summers, and reductions of 0.026 to 0.027 ccf per day in winter 2015-2016. The re-weighting and IV estimation hardly change the estimates because 98.4 percent of households are compliers, and only 0.08 percent of households opted out. The estimated energy savings are slightly — although not statistically significantly — larger in column 4, which suggests that survey respondents have somewhat larger energy savings, perhaps because they are more engaged with the HERs.

In column 3, our estimate of the average treatment effect for winter 2015-2016 in target population $\mathcal{P}_n$ is 2.70 ccf/day. There were 243 days between November 1, 2015 and June 30, 2016, and the average marginal price was \$0.80/ccf, so this sums to \$5.23 of retail natural gas cost savings. Control group usage averages 2.35 ccf/day in winter 2015-2016, so the treatment effects in columns 1-3 amount to about 1.1 percent of counterfactual usage.[19]

results of Allcott and Rogers (2014), who find that the effects of the first four HERs quickly decay. Allcott and Rogers (2014) also find persistent effects of HERs on electricity use after discontinuing treatment, but this is in samples of households that had received between eight and 24 HERs over two years, which is much more extensive than the first year of the Central Hudson program. Figure 6's evidence of zero statistical effects in summer 2015 (between the program's first and second year) is also consistent with the no persistence assumption, although certainly not dispositive.

[17]Specifically, we weight observations by billing period duration times a household weight, where the household weight is the inverse predicted probability of being a complier for households in the population $\mathcal{P}_n$, $\left[\hat{\Pr}(R_i = 1|\mathbf{X}_i; \mathcal{P}_n)\right]^{-1}$. This uses probit estimates from column 9 of Appendix Table A7.

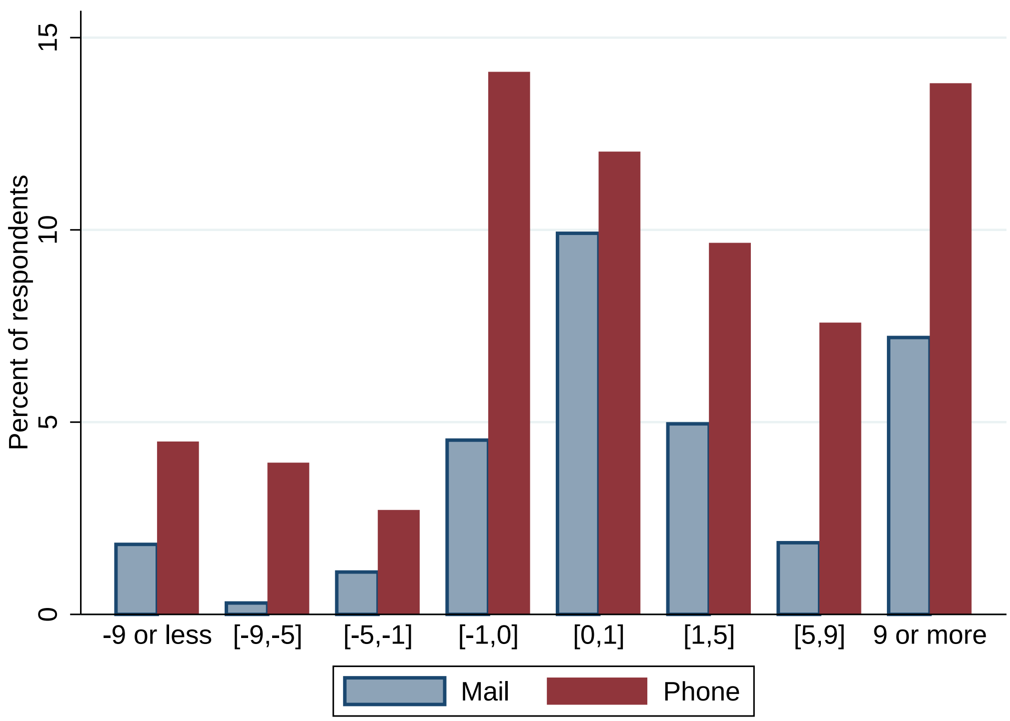
[18]Specifically, we weight households by the ratio of the predicted probability of responding to the survey with valid WTP to the predicted probability of being a complier, $\frac{\hat{\Pr}(H_i=1|\mathbf{X}_i;\mathcal{P}_n)}{\hat{\Pr}(R_i=1|\mathbf{X}_i;\mathcal{P}_n)}$, where H_i is an indicator for whether the household responded to the survey and has valid WTP. The numerator of this weight is predicted from estimates in column 7 of Appendix Table A7, while the denominator is from column 9.

[19]In percent terms, this is somewhat less than the typical effect of HERs on electricity use (Allcott 2015), but Opower's natural gas-focused programs typically have smaller percent effects than their electricity-focused programs.

IV.B Willingness-to-Pay

Figure 7 presents the distribution of WTP, with separate bars for the mail and phone survey responses. Fewer households responded via mail, so all mail bars are shorter. Mail respondents also have slightly higher willingness-to-pay, with relatively less density in the negative range and more in the positive range. Across all respondents, 34 percent reported weakly negative WTP, although most of that group is close to indifferent: 56 percent of negative WTPs are between $0 and $-1. This dispersion in WTP, and in particular the result that a meaningful share of the population is willing to pay to avoid being nudged, will motivate the analysis of opt-in programs and targeting in Section VI.

Figure 7: **Willingness-to-Pay for Home Energy Reports**



Notes: This figure presents the histogram of willingness-to-pay for four more Home Energy Reports, with all survey responses weighted equally.

Table 5 presents correlates of WTP. To simplify the presentation of the many **X** covariates, column 1 presents the post-Lasso estimator — that is, we use Lasso for variable selection, then present the OLS regression of WTP on the selected covariates; see Belloni and Chernozhukov (2013). The correlations are sensible: retirees have lower WTP, perhaps because of lower cash flow or less environmental concern, as do renters, likely because they do not have the ability or incentive to make energy-saving capital stock changes in response to HERs.

To give intuition for how our re-weighting on observables using IPWs affects estimated WTP, column 2 of Table 5 presents marginal effects of probit estimates of how the WTP predictors from column 1 are associated with whether a household responds and has valid WTP. The fact that four out of the six coefficients have the same signs in columns 1 vs. 2 suggests that survey responders may be slightly positively selected on observables, although one mechanism that works against this is that retirees have lower WTP but are more likely to respond to the survey.

Table 5: **Correlates of WTP and Their Correlation with Response**

	(1)	(2)
Dependent variable:	WTP	Have WTP
ln(Income)	0.0603	0.0295
	(0.244)	(0.0225)
Retired	-1.588	0.182
	(0.812)*	(0.0751)**
Married	0.683	-0.00765
	(0.414)*	(0.0368)
Rent	-0.780	-0.114
	(0.443)*	(0.0399)***
Single family home	0.322	0.0629
	(0.424)	(0.0382)*
Buyer score	0.342	0.0500
	(0.219)	(0.0199)**
Observations	2137	9439

Notes: Column 1 presents estimates from a post-Lasso estimator, in which OLS is run on covariates selected by Lasso, using equally-weighted observations. For the Lasso estimates only, each variable is normalized to standard deviation one. Column 2 presents marginal effects probit estimates from a model where the same selected covariates are used to predict whether a household responds to a survey and has valid WTP. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table 6 presents estimates of mean WTP, with standard errors in parentheses. Column 1 presents unweighted estimates, while column 2 uses row-specific IPWs to weight each row's sample to match target population $\mathcal{P}_n$ on observables. Mail survey responses are divided in two different ways: households randomly assigned to the base vs. follow-up groups and households that actually returned the first survey vs. the follow-up survey. The bottom row of Panel A reports that the unweighted mean WTP for the 24.5 percent of households that returned the survey is \$2.97. When re-weighted on observables to match $\mathcal{P}_n$, the mean falls to \$2.81, confirming that respondents are slightly positively selected on observables. We use this row of estimates as the base case for welfare analysis.

Table 6 shows that respondents to the first mail survey are positively selected. Unweighted mean WTP is marginally significantly higher for the randomly-assigned base group vs. follow-up group (\$4.32 vs. \$3.22, $p \approx 0.117$), and mean WTP is much higher for households in either group that returned the first mail survey vs. those that returned only the follow-up survey (\$4.36 vs. \$2.58, $p \approx 0.001$). This positive selection is almost mechanical: people who do not open and read HERs likely have WTP closer to zero than people who do, and the former group would not have even seen the first mail survey.

Table 6: **Estimates of Mean Willingness-to-Pay**

	(1)	(2)
	Unweighted	Weighted
Panel A: Mean WTP		
Mail	3.40	3.27
(standard error)	(0.26)	(0.3)
Base group	4.32	3.66
	(0.57)	(0.62)
Follow-up group	3.22	3.11
	(0.29)	(0.33)
Returned first survey	4.36	3.97
	(0.35)	(0.42)
Returned follow-up survey	2.58	2.63
	(0.37)	(0.41)
Phone	2.79	2.67
	(0.18)	(0.19)
Combined	2.97	2.81
	(0.16)	(0.16)
Panel B: p-Values of Differences		
Base vs. follow-up mail	0.117	0.490
Returned first vs. returned follow-up mail	0.001	0.024
Mail vs. phone	0.059	0.081
Base group vs. phone	0.026	0.160
Follow-up group vs. phone	0.213	0.224
Returned first survey vs. phone	0.000	0.004
Returned follow-up survey vs. phone	0.606	0.925

Notes: Samples exclude households that opted out before the program's second year. Estimates in column 2 are weighted to match the target population of treatment group households that did not opt out before the program's second year.

By contrast, the phone survey and follow-up mail survey, which was sent from a different outbound address and was not part of an HER, are not subject to this form of positive selection. Indeed, unweighted mean WTP is statistically and economically very similar for phone survey vs. follow-up mail survey respondents (\$2.79 vs. \$2.58, $p \approx 0.606$), and the weighted means are almost identical (\$2.67 vs. \$2.63, $p \approx 0.925$). This implies that these two samples are either not selected from non-respondents or that they have the same sample selection bias despite coming from two different forms of contact (mail vs. phone). Appendix Table A9 presents suggestive evidence in favor of the former explanation, showing that WTP does not vary statistically or economically for households that responded on earlier vs. later phone survey attempts. Extending this logic suggests that phone survey non-responders, who would in theory have responded on some eventual phone survey attempt, would have similar mean WTP. (This argument draws on the intensive follow-up approach used by DiNardo, McCrary, and Sanbonmatsu (2006) and others.) These results build confidence that we can extrapolate from the phone survey and the follow-up mail survey to the target population $\mathcal{P}_n$.

If respondents to the first mail survey are positively selected on unobservables but the remainder of mail and phone survey respondents are selected only on observables, then an unbiased estimate of mean WTP for the full HER recipient population can be constructed by giving first mail survey respondents weight of one (representing themselves only), and re-weighting phone and follow-up mail respondents to match the remaining HER recipients on observables.[20] We do this by repeating the previous IPW exercise but fixing the weights of first mail survey respondents to one. This gives a predicted population mean WTP of $2.68, not far from the full-sample weighted estimate of $2.81.

While we do not take lightly the extrapolation from survey respondents to the full target population $\mathcal{P}_n$, this discussion suggests that we have several reasonable approaches that generate similar results. If the reader believes that survey respondents are positively selected on unobservables, this would only reinforce the paper's main argument that traditional evaluation approaches overstate welfare gains. Readers who believe that respondents could be either positively or negatively selected can focus on the welfare evaluation for the subsample of survey respondents $\mathcal{P}_s$.

IV.B.1 Measuring Moral Utility

Our model in Section I includes a moral utility term that does not appear explicitly in most models. Does moral utility have any empirical content? And if so, are social comparisons a moral tax on "bad" behavior, as suggested by the concerns of Glaeser (2006) and others? The model generates four predictions that allow us to shed light on these questions.

First, if there is no moral utility, then a nudge that does not affect behavior will not affect consumer welfare. To see this, recall from Equation (5) that consumer welfare change is $\Delta V = -\Delta\tilde{e} \cdot p_e + \Delta f + \Delta M$. If there is no behavior change, then $-\Delta\tilde{e} \cdot p_e + \Delta f = 0$. If $\Delta M = 0$ also, then $\Delta V = 0$, so WTP should be zero. 39 percent of respondents to question 3 on the phone survey

[20]More precisely, denote $\mathcal{S}_1$ as the set of households that responded to the first mail survey, and denote H_i as an indicator for whether household i responded to any survey and has valid WTP. Further denote $\hat{\Pr}(H_i|\mathbf{X}_i; \mathcal{P}_n \setminus \mathcal{S}_1)$ as the conditional probability that a household in the population $\mathcal{P}_n$ excluding $\mathcal{S}_1$ has valid WTP. We can fit this probability using estimates in column 8 of Appendix Table A7. If w_i is WTP for household i and $N_n = 9948$ is the number of households in target population $\mathcal{P}_n$, the predicted target mean WTP is $\left(\sum_{i\in\mathcal{S}_1} w_i + \sum_{i\in\{\mathcal{P}_n\setminus\mathcal{S}_1\}} \frac{w_i}{\hat{\Pr}(H_i|\mathbf{X}_i;\mathcal{P}_n\setminus\mathcal{S}_1)}\right)/N_n$.

predicted that future HERs would not help them reduce their natural gas use "by even a small amount." These consumers have wide dispersion in WTP, with observations in all eight ranges and standard deviation just as large as for respondents predicting non-zero savings. Moral utility, or some other unmodeled factor unrelated to financial gain or consumption utility, is needed to explain this non-zero WTP for consumers predicting zero behavior change.

Second, if HERs act *only* as a moral tax, i.e. they increase the moral price μ but have no other effect, then $\Delta V < 0$. As we saw above, however, average WTP is positive. HERs almost certainly have a meaningful informational component, and we saw above that 73 percent of phone survey respondents agree that HERs give useful energy conservation information. Thus, it is clear that HERs do not act only as a moral tax.

Third, if HERs increase the moral price μ, this should tend to decrease moral utility more for heavy users. Intuitively, a moral price increase hurts heavy users more because it accrues over more inframarginal units, just as an actual price change affects expenditures more for high-demand consumers.[21] Testing this requires us to measure ΔM. The phone survey questions asking consumers if HERs made them feel inspired, pressured, proud, or guilty were designed to help proxy for positive and negative aspects of moral utility. Define $\mathbf{A}_i$ as a vector of four indicator variables capturing individual i's responses to those four affect questions, and define E_i as predicted savings from question 3. We regress WTP w_i on $\mathbf{A}_i$ and E_i in the sample of phone survey respondents:

$$w_i = \beta_0 + \beta_E E_i + \beta_A \mathbf{A}_i + \epsilon_i. \tag{9}$$

This is a rough empirical analogue to Equation (5), in which w_i proxies for ΔV, $\beta_E E_i$ proxies for $-\Delta\tilde{e}\cdot p_e + \Delta f$ (under the assumption that Δf scales proportionally with savings $-\Delta\tilde{e}\cdot p_e$), and $\beta_A \mathbf{A}_i$ proxies for ΔM_i. Estimates show that predicted savings E_i is strongly positively associated with WTP, while feeling inspired and pressured, respectively, are positively and negatively conditionally associated with WTP with greater than 90 percent confidence. See Appendix Table A10 for formal results. Using these estimates, we fit $\widehat{\Delta M}_i = \hat{\beta}_A \mathbf{A}_i$.

The first row of Table 7 presents results of univariate regressions of seven different variables on mean usage in winter 2014-2015, measured in ccf/day. In the context of the model, this usage variable is $\tilde{e}(\theta_1)$. Column 1 reports that a one ccf/day increase in usage is unconditionally associated with a \$0.188 increase in WTP. Heavier users have higher predicted savings, are more likely to report negative affect (feeling pressured or guilty), and are less likely to report positive affect (feeling inspired or proud). Column 7 reports that heavier usage is associated with reduced moral utility $\widehat{\Delta M}_i$. This suggests that the HERs do increase μ. Results are similar when regressing the

[21] This requires a bound on the usage decrease for heavier users relative to lighter users. Intuitively, if the existing moral price was positive and heavy users decrease usage by much more than light users, heavy users could gain moral utility relative to light users by reducing inframarginal moral utility losses. Formally, decompose ΔM into $\Delta M = \Delta m - \Delta\mu \cdot \tilde{e}(\theta_1) - \mu_0 \cdot \Delta\tilde{e}$ and take $\frac{d\Delta M}{d\tilde{e}(\theta_1)} = -\Delta\mu - \mu_0 \frac{d\Delta\tilde{e}}{d\tilde{e}(\theta_1)}$. We think of the moral price μ as being weakly positive. If $\mu_0 > 0$, then $\frac{d\Delta M}{d\tilde{e}(\theta_1)} < 0$ if $\frac{d\Delta\tilde{e}}{d\tilde{e}(\theta_1)} > \frac{\Delta\mu}{\mu_0}$, i.e. if behavior change does not increase too much in $\tilde{e}(\theta_1)$. If $\mu_0 = 0$, $\frac{d\Delta M}{d\tilde{e}(\theta_1)} < 0$ holds unambiguously.

same outcomes on baseline usage instead of post-treatment usage.

In Section I, we remarked that our model nests a model in which moral utility depends on the perceived social norm s: $M = m_s - \mu(e - s)$. The variable "Mean comparison" is an empirical analogue of $(e - s)$: it is the average difference between own natural gas usage and mean neighbor usage on the first winter of HERs. Households with higher "Mean comparison" were informed that they were relatively heavy users. Substituting $(e - s)$ for e in the model generates the analogous prediction that if $\Delta\mu > 0$, then $(e - s)$ should be negatively correlated with ΔM. The second row of Table 7 confirms that this is the case empirically: relatively heavier users report more negative affect, less positive affect, and have lower fitted moral utility $\widehat{\Delta M}_i$. Results are similar when regressing the same outcomes on $(e - s)$ from only the first HER.

Table 7: **Measuring Moral Utility**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	WTP	Predicted savings	Inspired	Pressured	Proud	Guilty	$\widehat{\Delta M}$
Mean usage	0.188	0.658	-0.0130	0.0128	-0.0411	0.0180	-0.0417
	(0.0773)**	(0.257)**	(0.00478)***	(0.00366)***	(0.00451)***	(0.00401)***	(0.0156)***
Mean comparison	0.246	1.208	-0.0369	0.0364	-0.0974	0.0402	-0.120
	(0.157)	(0.547)**	(0.00985)***	(0.00792)***	(0.0108)***	(0.00830)***	(0.0316)***

Notes: This table presents results of univariate regressions of the dependent variable in each column on the independent variable in each row. "Mean usage" is mean natural gas usage in ccf/day in winter 2014-2015. "Mean comparison" is the average difference (in 1000s of ccf) between own natural gas usage and mean neighbor usage on the HERs in winter 2014-2015. Observations are weighted to match the target population of treatment group households that did not opt out before the program's second year. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

We also find that WTP is \$0.69 lower ($p \approx 0.076$) for the randomly-assigned "Comparison" survey version that reminds people that the HERs compare their energy use to their neighbors' use. This is consistent with the hypothesis that social comparisons are the part of the HERs that reduce moral utility. The "Environmental" version does not statistically significantly affect WTP. See Appendix Table A11 for formal results.

A fourth prediction is that if $\Delta\mu > 0$ but $\Delta m = 0$, then $\Delta M < 0$. In words, if a nudge increases the moral price but provides no other utility windfall, then it will decrease moral utility. Alternatively, however, a nudge can both increase the moral price and provide some additional utility Δm. In fact, the mean $\widehat{\Delta M}_i$ fitted from above is \$0.95, because as shown in Figure 5, more people reported positive affect than negative affect. This suggests that $\Delta m > 0$.

In simple terms, these results show that more people report positive affect than negative affect, but heavy users are relatively less likely to report positive affect. In the context of our model, these results imply that the HERs act through multiple channels: providing information, increasing the moral price μ, and providing a "windfall" of positive affect $\Delta m > 0$ for the average consumer.

V Welfare

In this section, we use the empirical estimates of energy savings and WTP to calibrate the welfare formula from Equation (7). Before doing this, we make explicit our key revealed preference assumption and calibrate four additional parameters.

V.A Revealed Preference Assumption

Our key assumption is that consumer i's WTP w_i equals the consumer welfare change ΔV_i from the second year of the HER program:

$$\Delta V_i = w_i. \tag{10}$$

This assumption is only plausible in situations where consumers are well-informed about what the nudge is and, if the nudge addresses behavioral biases, are "sophisticated" about those biases. For example, the assumption would not hold for naive hyperbolic discounters evaluating a commitment device or for individuals who are uninformed about the benefits and costs of a choice that is being nudged. On the other hand, the assumption would hold for people evaluating information in a rational information acquisition model, or for sophisticated hyperbolic discounters evaluating a commitment device.

We chose HERs as our application because we believe that this assumption is particularly plausible in this context. After receiving several HERs, each of which is different but follows a similar structure, consumers are well-informed about what HERs are and have a good sense of how future HERs would further inform or motivate them. Unlike other settings where we might expect experts to be better informed about welfare (see, e.g., Bronnenberg *et al.* (2015)), HER customers are the best equipped to know the personal value they receive from the reports.[22]

As noted above, because WTP w_i is for the second year of HERs, our welfare analysis is relevant only to the program's second year, which is particularly relevant to study, as discussed in Section II.

[22] Appendix D.A provides additional evidence on two biases that might be relevant in this context. There is suggestive evidence that consumers overestimate the energy savings caused by HERs, which could bias WTP upward. There is also suggestive evidence that consumers are overoptimistic, by which we mean that they tend to underestimate their own energy use before the arrival of the first HER. However, there is no evidence that this optimism affects WTP.

V.B Implementation Cost, Externality, Markup, and Energy Savings Parameters

The welfare analysis requires four additional parameters: implementation cost C_n, externality ϕ_e, retail markup π_e, and energy savings $\Delta\tilde{e}$.

We calculate average cost C_n using an accounting analysis detailed in Appendix D.B. The second year of an HER program entails both fixed costs F_n and per-household marginal costs c_n, so $C_n = F_n/N_n + c_n$, where $N_n = 9948$ is the number of households in target population $\mathcal{P}_n$, the set of recipient households that did not opt out before the program's second year. The per-household marginal cost of the program's second year is $c_n \approx \$2.06$, almost entirely for printing and mailing HERs. Opower and Central Hudson also incur an estimated \$16,339 per year in costs to manage ongoing programs. Central Hudson has three HER programs in addition to the one we study, for a total of four programs and about 100,000 recipient households.[23] Some of the ongoing management costs are effectively fixed costs per program, whereas others do not depend on the number of programs. In our primary estimates, we allocate the \$16,339 equally to each of Central Hudson's 100,000 recipient households, giving $F_n/N_n \approx \$0.16$/household. We also present an alternative calculation in which these costs are allocated equally to each of the four programs. This gives $F_n \approx \$4085$ per program, or $F_n/N_n \approx \$0.41$ per household in the program we study.

To calculate externality ϕ_e, we include local air pollution and carbon dioxide externalities from natural gas combustion as well as methane externalities from the natural gas supply chain. For local air pollutants, we consider nitrogen oxides, particulate matter, and sulfur dioxide. We use the EPA (1995) AP-42 emission factors and marginal damages from Holland *et al.* (2015), whose key assumptions are a \$6 million value of a statistical life and a fine particulate dose response function from Pope *et al.* (2002). Holland *et al.* provided us with county-specific marginal damages relevant for ground-level emissions (i.e., homes instead of power plant smokestacks), and we take the mean across counties, weighting by the number of households in the HER experiment. Local air pollutant damages amount to \$0.045/ccf. Using results from the U.S. Government Interagency Working Group on the Social Cost of Carbon (2013), we use a \$40 social cost of carbon, which translates to \$0.264/ccf damages from natural gas combustion. Drawing on Howarth *et al.* (2012) and Abrahams *et al.* (2015), we assume that three percent of natural gas escapes during drilling and transportation before arriving in homes. We translate this to carbon dioxide equivalents using a methane global warming potential of 34 from the Intergovernmental Panel on Climate Change (Myhre *et al.* 2013), giving an additional \$0.10/ccf externality. Thus, the total environmental externality ϕ_e is \$0.045+\$0.264+\$0.10≈\$0.411 per ccf.

Central Hudson uses decreasing block pricing. Marginal prices consist of a constant marginal gas

[23] Different programs are well-defined in the sense that they have different specific customer sub-populations in recipient and control groups. Different programs start at different times, may focus on different fuels (e.g., households that purchase electricity but not natural gas), and have custom-designed elements on the HERs.

supply charge, which passes through Central Hudson's cost to acquire gas from wholesale pipelines, plus constant marginal fees and taxes such as the "system benefit charge" used to fund energy efficiency programs, plus decreasing marginal delivery charges, which allow Central Hudson to recover additional fixed costs such as maintenance, customer service operations, meter reading, and billing.[24] During the program's second year (winter 2015-2016), Central Hudson's usage-weighted marginal acquisition cost was $c_e \approx \$0.37/\text{ccf}$, while the usage-weighted marginal retail price was $p_e \approx \$0.80/\text{ccf}$. Thus, marginal retail prices exceed marginal acquisition costs by an average of $\pi_e \approx \$0.43/\text{ccf}$.[25] When households use less gas due to HERs or any other conservation program, they reduce their contribution to Central Hudson's fixed costs by that $\pi_e \approx \$0.43/\text{ccf}$, and this will eventually be made up through higher prices.[26] This pricing approach is not unusual: Central Hudson's retail markup is almost exactly identical to the 40 percent average markup for residential and commercial natural gas consumers nationwide, as calculated by Davis and Muehlegger (2010).

Comparing the environmental externality and retail markup, Central Hudson's retail marginal gas price is 2.1 percent ($0.016/ccf) *above* social marginal cost. As Davis and Muehlegger (2010) point out, although this is sensitive to the social cost of carbon and other externality damage parameters, it calls into question the argument that energy efficiency programs are needed as second best substitutes for getting prices right. Instead, natural gas conservation programs in this context are justified only to the extent that they address market failures such as imperfect information or otherwise increase consumer welfare. Any welfare gains of the nudge will need to be driven by private gains to nudge recipients rather than by uninternalized social benefits.

In this context, $\Delta\tilde{e}$ represents the causal effect of the second year of HERs on natural gas use. Conceptually, the ideal way to estimate this would be to compare natural gas use at households that were randomly assigned to receive a second year of HERs to households that received only the first year, using a design similar to Allcott and Rogers (2014). This was not feasible due to regulatory constraints — and, in any event, such estimates would not have been very precise in a program of this size. Instead, we maintain the "no persistence" assumption that HERs sent during a given winter only affect energy use in that same winter. Under this assumption, $\Delta\tilde{e}$ equals the treatment effects on energy use in winter 2015-2016 in columns 3 and 4 of Table 4, for target populations $\mathcal{P}_n$

[24]To be clear, "fixed costs" means costs that are fixed with respect to the volume of natural gas consumed, not necessarily fixed with respect to the number of residential consumers or some other factor. In addition to the gas supply charge, we include the relatively small "merchant function charge" as part of marginal acquisition costs. If this charge is instead classified as a fixed cost, this slightly increases π_e, which would slightly worsen the program's welfare effects.

[25]Because the extensive margin (natural gas connections) is highly inelastic, while the intensive margin (natural gas use) is more moderately inelastic, the Ramsey-Boiteux framework suggests that it would be more economically efficient to pass through fixed costs as fixed monthly charges. There are various justifications for amortizing fixed costs into marginal prices, including horizontal and vertical equity (Borenstein and Davis 2012), and the allocative impact of this distortion is mitigated if consumers respond to average instead of marginal prices (Ito 2014). Regardless of whether this rate structure is desirable, decreased net revenue still enters the welfare calculation.

[26]Central Hudson's profits are regulated by the New York Public Service Commission. If profits fall short of the allowed amount, Central Hudson is allowed to make this up in future years through higher retail prices.

and $\mathcal{P}_s$ respectively, multiplied by 243 to translate from ccf per day to total ccf over the winter.

Even if this "no persistence" assumption does not hold exactly, we emphasize that it does not matter for our social welfare estimates: these estimates are not very sensitive to the value of $\Delta\tilde{e}$ because retail marginal prices are so close to social marginal cost. In other words, the term $(\pi_e - \phi_e)$ that multiplies $\Delta\tilde{e}$ in the social welfare equation is approximately zero, so our welfare calculations would change very little under any plausible alternative assumptions for $\Delta\tilde{e}$.

V.C Results

Table 8 presents the welfare analysis of the program's second year. Columns 1 and 2 present results after reweighting the samples to match $\mathcal{P}_n$, the target population of treatment group households that did not opt out before the program's second year. These columns use energy savings from column 3 of Table 4 and WTP from column 2 of Table 6. Columns 3 and 4 present results for $\mathcal{P}_s$, the target population of MPL respondents with valid WTP that did not opt out. These columns use energy savings from column 4 of Table 4 and WTP from column 1 of Table 6. Columns 1 and 2 are noteworthy because they evaluate the full policy, while Columns 3 and 4 are noteworthy because we do not have to re-weight observations in calculating average WTP.

Table 8: **Social Welfare Effects of a Second Year of Home Energy Reports**

	(1)	(2)	(3)	(4)
Target population:	All HER Recipients		MPL Respondents	
Panel A: Benefits and Costs Other than Consumer Welfare ($/recipient)				
Implementation cost: C_n	2.22		2.22	
Retail gas savings:$-\Delta\tilde{e} \cdot p_e$	5.23		6.10	
Gas acquisition cost savings: $-\Delta\tilde{e} \cdot c_e$	2.43		2.83	
Utility net revenue loss: $-\Delta\tilde{e} \cdot (p_e - c_c)$	2.80		3.27	
Externality reduction: $-\Delta\tilde{e} \cdot \phi_e$	2.70		3.15	
Δwelfare, excluding WTP: $-C_n + (\pi_e - \phi_e)\Delta\tilde{e}$	-2.33		-2.34	
Panel B: Mean WTP and Social Welfare Effect ($/recipient)				
Assumption	Mean WTP	ΔWelfare	Mean WTP	ΔWelfare
Base case	2.81	0.49	2.97	0.63
Uniform WTP at MPL endpoints	2.59	0.26	2.74	0.39
WTP = {-12,12} at MPL endpoints	2.34	0.01	2.48	0.13
WTP = {-15,15} at MPL endpoints	2.76	0.43	2.92	0.58
Non-respondents have WTP = 0	0.60	-1.73		
Weight = 1 for first mail respondents	2.68	0.35		
Fixed costs equally allocated	2.81	0.24	2.97	0.38

Notes: Columns 1 and 2 present results after reweighting the samples to match $\mathcal{P}_n$, the target population of treatment group households that did not opt out before the program's second year. These columns use energy savings from column 3 of Table 4 and WTP from column 2 of Table 6. Columns 3 and 4 present results for $\mathcal{P}_s$, the target population of MPL respondents that did not opt out. These columns use energy savings from column 4 of Table 4 and WTP from column 1 of Table 6. ΔWelfare in columns 2 and 4 is $\Delta W = \int \Delta V - C_n + (\pi_e - \phi_e)\Delta\tilde{e} \, dF(\Theta)$ from Equation (7), where Mean WTP in columns 1 and 3 is our measure of consumer welfare gain ΔV.

Panel A presents benefits and costs other than consumer welfare. For target population $\mathcal{P}_n$, natural gas savings amount to $5.23 and $2.43 per recipient household at retail price and acquisition cost, respectively. The difference between these two figures is the utility net revenue loss: $2.80 per household-year. Environmental externalities drop by $2.70 per recipient household. The social welfare effect excluding WTP is externality reduction minus utility net revenue loss minus implementation cost, or $2.70-$2.80-$2.22≈$-2.33 for target population $\mathcal{P}_n$.

Panel B completes the social welfare estimates by adding in WTP. Columns 1 and 3 present WTP, while columns 2 and 4 present the resulting social welfare estimate using Equation (7). In the base case, WTP is $2.81 and $2.97 for target populations $\mathcal{P}_n$ and $\mathcal{P}_s$, respectively, as we found in Table 6. The social welfare effects are $0.49 and $0.63 per household for the full population and for MPL respondents, respectively.

The next three rows of Panel B implement alternative assumptions for mean WTP at the endpoints of the MPL (i.e., mean WTP for those consumers with WTP below -$9 or above $9). The first assumes a uniform distribution of WTP beyond the endpoints, with density equal to the density on the adjacent WTP bin. This gives mean WTPs of \$13.08 and -\$11.48 for the upper and lower endpoints, respectively. The next two rows use \$12 or \$15 as heuristic benchmarks. All three of these alternative assumptions give lower mean WTP, so less positive welfare effects. Because only 27 percent of respondents have WTP at one of the endpoints, this alone does not significantly change mean WTP.

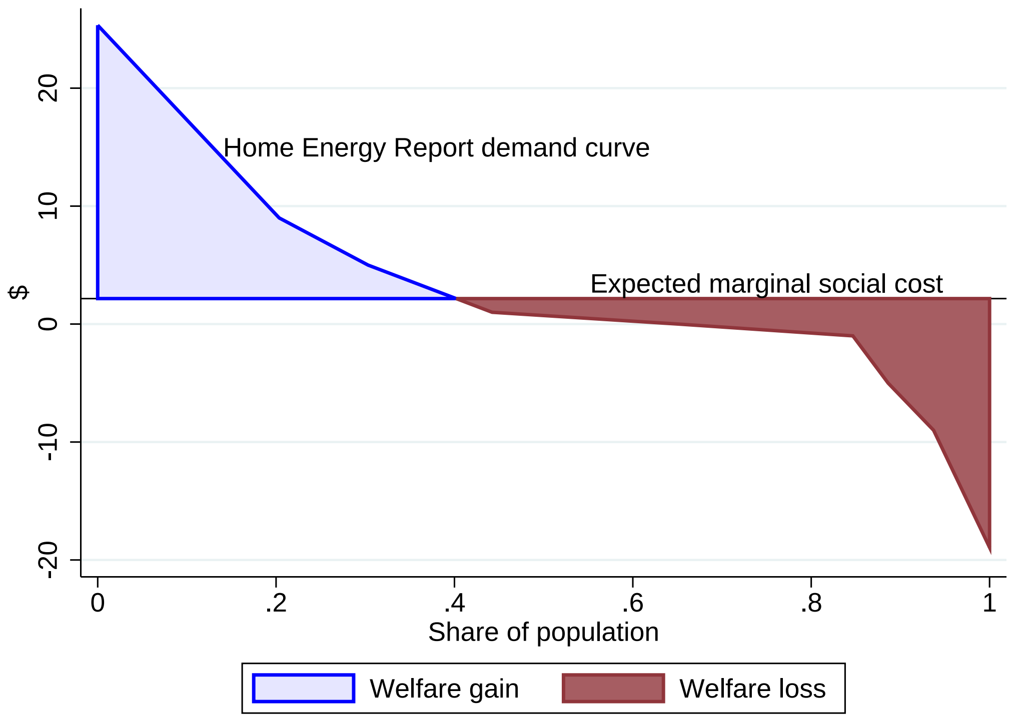
The next two rows of Panel B consider alternative adjustments for survey non-response when extrapolating WTP to the full HER recipient population. In Section IV.B, we speculated that if there is a non-response bias, it is likely positive. Under the extreme assumption that non-respondents have zero WTP, welfare effects would be \$-1.73 per HER recipient. We view this as an unrealistically conservative assumption.[27] When we assume that mean WTP is \$2.68, as calculated by the alternative weighting procedure in which respondents to the first mail survey have weights fixed to one, welfare gains are \$0.35 per recipient.

The final row uses the higher average implementation cost C_n if the fixed costs of continuing programs are allocated equally to each of Central Hudson's four ongoing programs. This penalizes small programs and benefits large ones. While this cost allocation assumption is likely too extreme, it's certainly true that at some point an HER program would not be large enough to generate enough social surplus to outweigh the program-level fixed costs. If implementation costs were 21 percent higher, externality damages were 18 percent lower, or WTP were 17 percent lower, the base case social welfare point estimate would be negative.

Because $(\pi_e - \phi_e) \approx 0$, i.e. marginal retail prices are very close to true social marginal cost, the social welfare effect depends very little on estimated energy savings $\Delta\tilde{e}$. Thus, the sampling error in our energy savings estimates does not make much difference on our welfare estimates. Applying the Delta method to the energy savings estimates in Table 4, the 95 percent confidence interval on welfare effects for target $\mathcal{P}_n$ extends $1.96 \cdot \hat{SE}(\tau_4) \cdot 243 \cdot (\phi_e - \pi_e) \approx \$0.08$ in either direction. WTP estimates in Table 6 are relatively precisely estimated, with a 95 percent confidence interval that extends \$0.31 in either direction for both the weighted and unweighted estimates.

[27] Although EPA (2006) reports that 44 percent of unsolicited mail is not read, HERs arrive in utility branded envelopes. Since utilities typically send bills or other important communications, open rates are likely to be much higher than standard unsolicited mail. Just under five percent of phone survey respondents reported not remembering HERs.

Figure 8: **Social Welfare Analysis: Graphical**



Notes: This figure presents a graphical version of the base case welfare analysis weighted for target population of HER recipients that did not opt out before the program's second year, corresponding to columns 1 and 2 of Table 8.

Figure 8 illustrates our base case welfare analysis, weighted for the HER recipient population. The demand curve is drawn to be consistent with the assumptions used to code WTP from the MPL responses: WTP is distributed triangular on the highest and lowest ranges and uniform on the six interior ranges of the MPL. "Expected marginal social cost," i.e. $\int c_n - (\pi_e - \phi_e)\Delta\tilde{e}\ dF(\Theta)/N_n = c_n - (\pi_e - \phi_e) \cdot 243 \cdot \hat{\tau}_4$, is approximately \$2.16 per household. The net social welfare effect is the area between the demand curve and expected marginal social cost, i.e. the light-shaded blue area minus the dark-shaded red area, minus fixed cost F_n. Leaving aside the variation in $\Delta\tilde{e}$ across households, which is less relevant because $(\pi_e - \phi_e) \approx 0$, the social welfare effect trades off the gains to the 40 percent of consumers willing to pay more than \$2.16 with the losses to the 60 percent of consumers that are not. The consumer surplus in the lightly shaded area is large: more than 30 percent of people are willing to pay twice the social marginal cost. This figure motivates the opt-in and targeting analysis in Section VI: perhaps the nudge policy can be modified to avoid the loss in the darkly shaded area.

V.D Discussion: Why Measuring Consumer Welfare Matters

Using the consumer welfare formula in Equation (5), the difference between mean WTP (\$2.81) and retail energy savings (\$5.23) implies that consumers incur an average of \$2.42 in net utility costs, which we call "non-energy costs" for shorthand.[28] This benefit/cost ratio of \$5.23/\$2.42≈2.16 implies that leaving aside implementation costs C_n, HERs generate highly privately-beneficial energy savings for recipients. However, these energy savings do not accrue to consumers for free.

This is important because HERs and other behavior-based energy efficiency programs are evaluated for regulatory compliance purposes using institutionalized program evaluation approaches that ignore non-energy costs. Specifically, these programs are evaluated using what's called a "program administrator cost" metric, which considers energy savings and program implementation costs but does not consider non-energy costs incurred by consumers. In other words, the energy industry evaluates HERs and similar behavior-based programs as if they allow consumers to somehow achieve energy savings with no effort or cost whatsoever. In the context of a smoking cessation program, this is analogous to assuming that the only effect on consumer welfare is to save people money on buying cigarettes.

How does ignoring non-energy costs affect the social welfare calculation in Table 8? If we set $\Delta V = -\Delta\tilde{e} \cdot p_e$, the welfare gain is \$2.91 per recipient — almost six times larger than our base case estimate of \$0.49 per household. In other contexts, it is easy to imagine that this could change whether or not a nudge is determined to be welfare enhancing.

Evaluating only the program's second year leaves open the question of whether the full program (from beginning to end) is welfare enhancing. In particular, there are fixed costs to begin a program that do not enter F_n, the fixed cost of continuing an existing program. Furthermore, there have been many different Home Energy Report programs with very different energy savings effects. In Appendix D.C, we provide a speculative, back-of-the envelope calculation under the assumptions that Opower's price reflects the cost of a full program and that non-energy costs are \$2.42/\$5.23≈46% of total retail energy savings. We consider the full life of a typical Opower program, using energy savings estimates from Allcott and Rogers (2014). Our estimates suggest that the typical full program is welfare enhancing, but ignoring non-energy costs overstates welfare gains by a factor of 2.2.

VI Allocating Nudges: Opt-In vs. Smart Defaults

Figure 7 shows that WTP for HERs is highly heterogeneous. The effect of HERs on energy use may be heterogeneous as well. Can better allocation of this nudge improve its social welfare effects?

[28]Natural gas prices dropped sharply in April 2015, after most of the winter 2014-2015 heating season but before our surveys were conducted. Instead of using realized savings, we could instead use consumers' predictions of future retail energy cost savings from the phone survey. Mean expected savings is larger than the observed \$5.23, so this would imply larger non-energy costs, which further reinforces our arguments in this section.

We consider two approaches: an opt-in program and a machine learning algorithm that targets the nudge to maximize social welfare.

VI.A Opt-In Programs

A natural reaction to heterogeneous valuations of a good or service is that it should be priced at social marginal cost, and consumers should be allowed to buy or not buy as they wish. We begin by evaluating that idea. For simplicity, we assume that the average energy savings of consumers that opt into HERs equals the estimated average treatment effect $\hat{\tau}$ from Table 4. We then set the price at expected social marginal cost $c_n - (\pi_e - \phi_e)\hat{\tau}D \approx \$2.16$.

Table 9 presents results. Column 1 presents the percent of population receiving HERs, while Columns 2-4 present the mean natural gas use change, WTP, and social welfare change per recipient household, respectively. Column 5 presents the aggregate social welfare effect across all 19,929 households, which is (column 1)/100 $\times$ (column 4) $\times$ (19,929/1000). Row 1 presents the existing opt-out program as a benchmark.

Row 2 presents the welfare effects of an opt-in program assuming zero switching cost — that is, we assume that all consumers opt into the second year of HERs if they are willing to pay more than the \$2.16 price. Under this assumption, 40.1 percent of consumers opt in, and they have mean WTP of \$9.95. The total social welfare gain in column 5 is 12 times larger than for the existing program, even though fewer households are included. This dramatic improvement arises because a significant number of consumers with low or negative WTP no longer are nudged.

Opower has run one opt-in program in the U.S., at a large utility called American Electric Power in Ohio. They aggressively marketed free HERs to 250,000 customers, of whom only 1.5 percent opted in. Although the Ohio population could be different, the low opt-in rate suggests that default effects are very powerful in this context. In other words, switching costs or other forms of inertia prevent many people who value the nudge at more than its price from opting in. Given results from Madrian and Shea (2001), Kling *et al.* (2012), Handel (2013), Ericson (2014), and others showing the power of inertia in high-stakes choices such as retirement savings plans and health insurance, it is very plausible that inertia could be powerful in low-stakes decisions such as whether to receive Home Energy Reports. This implies that the zero switching cost assumption in row 2 is unrealistic.

We explore the importance of switching costs under three assumptions. First, 1.5 percent of consumers opt in, as in Ohio. Second, consumers opt in if and only if their WTP is larger than the switching cost, so the 1.5 percent that opt in will be drawn from the right tail of the WTP distribution. Third, the switching cost is not welfare-relevant — in other words, an implied switching cost arises from factors such as imperfect information, not because of a material transaction cost. These latter two assumptions give a best-case scenario for welfare gains for a given switching cost.

Row 3 shows that even under this best-case scenario, the welfare gains from an opt-in program are \$5,100 — only slightly more than for the current opt-out program in row 1. Even though mean WTP of nudge recipients is high, substantial potential consumer welfare gains are lost because many high-WTP consumers do not opt in. Furthermore, the fixed implementation cost F_n is spread across a small number of recipients.

Table 9: **Opt-In and Smart Defaults: Results**

		(1)	(2)	(3)	(4)	(5)
Row	Policy	Percent of population receiving HERs	Mean gas use change (ccf/ recipient-day)	Mean WTP (\$/ recipient)	Welfare effect (\$/ recipient)	Total welfare effect (\$000s)
1	Existing opt-out program	50	-0.026	2.81	0.49	4.9
2	Opt-in; zero switching cost	40.1	-0.026	9.95	7.59	60.7
3	Opt-in; 1.5% opt-in rate	1.5	-0.026	24.5	16.94	5.1
4	Targeted on energy savings	50	-0.052	3.21	0.78	7.8
5	Targeted on WTP	50	-0.041	3.51	1.12	11.2
6	Targeted on welfare	50	-0.029	3.59	1.26	12.6
7	Drop recipients; maximize welfare	40	-0.024	3.25	0.89	7.1

Notes: Column 5 presents the aggregate social welfare effect across all 19,929 households, which is (column 1)/100 × (column 4) × (19,929/1000).

VI.B Targeted Opt-Out Programs

The importance of both heterogeneity and inertia suggests a different policy approach: an opt-out program that targets consumers who would generate large welfare gains and excludes consumers who would not.

Formally, we want to derive a statistical decision rule $\delta : \mathcal{X} \to \{0, 1\}$ that maps household covariates from space $\mathcal{X}$ to treatment assignment $\{0, 1\}$ in order to maximize objective $L(\delta)$. Initially, we hold the number of recipient households constant at 50 percent of the 19,929-household population and compare the results of maximizing three different objectives: energy conservation, where $L_\tau(\delta) = \sum_i -\tau_i \delta(\mathbf{X}_i)$, consumer welfare, where $L_{CW}(\delta) = \sum_i w_i \delta(\mathbf{X}_i)$, and social welfare, where $L_W(\delta) = -F_n + \sum_i (w_i + (\pi_e - \phi_e)\tau_i - c_n)\, \delta(\mathbf{X}_i)$.

This is a standard prediction problem where additional covariates provide more information, but using too many covariates leads to overfitting, which worsens out-of-sample prediction. From a computational perspective, the only unusual feature of this problem is that different parts of the social welfare objective function $L_W(\delta)$ are estimated from different datasets: WTP w_i is from the MPL surveys, while energy savings τ_i is from the full sample of billing data. Standard pre-packaged procedures thus cannot be used. For simplicity, we select variables using forward stepwise regression and consider only linear combinations of the 20 $\mathbf{X}$ variables. We use five-fold cross validation to avoid overfitting. Appendix E presents details.[29]

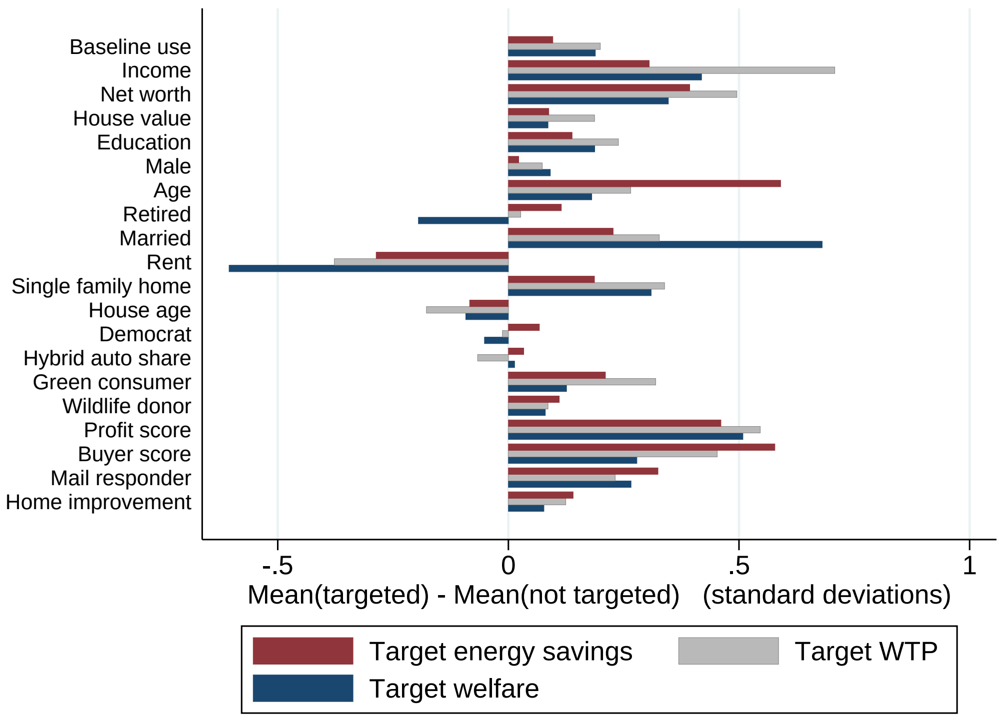
Rows 4-6 present results when maximizing energy savings, WTP, and welfare, respectively. The table clearly shows how targeting can improve performance on whatever objective the algorithm is trained to maximize. Remarkably, there is little tradeoff between targeting on energy savings and targeting on WTP: maximizing WTP generates only slightly lower energy savings than maximizing energy savings, and vice versa.

Figure 9 presents differences in mean **X** variables (normalized into standard deviations) between targeted and non-targeted households for each of the three maximands in rows 4-6. The figure shows that all three algorithms target similar households: for most variables, all three bars extend in the same direction and even have comparable magnitudes. This explains why there is little tradeoff between maximizing WTP and maximizing energy savings. The fact that WTP and energy savings are positively correlated with the same observables implies that WTP and energy savings are themselves positively correlated, unless they have strong opposite correlations with unobservables.

This positive correlation between WTP and energy savings has two interesting implications. First, this is again consistent with the idea that the informational channel outweighs the moral tax channel in generating behavior change: as discussed in Section IV.B.1, if the moral tax channel were more active, the households with the largest behavior change would likely have the lowest WTP. Second, at least in this population, existing policies, such as Energy Efficiency Portfolio Standards, that encourage utilities to target households that generate the largest energy savings also tend to encourage targeting that is beneficial from a social welfare perspective. Of course, this result is purely accidental, and ideally policies would be written to explicitly encourage targeting to maximize welfare.

[29] One alternative approach would be to use a tree method akin to those in Athey and Imbens (2015), splitting the WTP data and energy use data at the same nodes. In order to reduce residual variance, we would first residualize energy use of the baseline usage controls in Equation (8).

Figure 9: **Demographic Differences Between Targeted and Non-Targeted Households**



Notes: We use the machine learning algorithm to target 50 percent of the Central Hudson program population, maximizing energy savings, willingness-to-pay, or welfare. This figure presents the normalized difference in means between targeted and non-targeted households for each of these three maximands, in standard deviation units.

The algorithm can also be used to predict which current recipients generate welfare losses and drop them from the program's second year. To do this, we train the algorithm to maximize social welfare $L_W(\delta)$ while allowing it to target any subset of the existing treatment group. Row 7 presents results; this should be compared to the results for the current recipient group in row 1. By nudging 40 percent instead of 50 percent of the entire population — i.e. dropping about 20 percent of the current recipient group — the algorithm increases the total welfare gain by 45 percent.

When comparing opt-in and targeted opt-out policies, the typical comparative static is that more consumer inertia favors a targeted policy, while poor ability to predict welfare favors an opt-in policy. The remarkable feature of these results is that even with generous assumptions about the welfare gains from an opt-in policy, inertia is such a large barrier that a targeted opt-out policy is preferred.

VII Conclusion

Many economists recognize the importance of evaluating nudge-style interventions on the basis of social welfare, not just behavior change. Nevertheless, it is often difficult to actually quantify the full consumer welfare effects of a given nudge. Our main contribution is to develop and implement an experimental design that allows for an empirical social welfare analysis in a case study of one prominent nudge.

There are three main takeaways. First, we find significant individual-level heterogeneity in willingness to pay for the nudge, including a significant minority of consumers who prefer not to be nudged. This implies large welfare gains from using prediction for "smart defaults." Second, despite the worries of Glaeser (2006) and others, social comparison nudges need not only act as an emotional tax on "bad" behavior. We find evidence that in addition to increasing the moral price, HERs work by providing both information and additional windfall utility through positive affect. Third, the nudge we study increases welfare. However, this welfare gain comes with costs to consumers that typically go unmeasured, and ignoring these "non-energy costs" would cause the analyst to overstate social welfare gains by a factor of six. These results highlight the importance of measuring the full welfare effects of nudges.

References

[1] Abrahams, Leslie S., Constantine Samaras, W. Michael Griffin, and H. Scott Matthews. 2015. "Life Cycle Greenhouse Gas Emissions From U.S. Liquefied Natural Gas Exports: Implications for End Uses." *Environmental Science and Technology* 49: 3237-3245.

[2] ACEEE (American Council for an Energy-Efficient Economy). 2015. "State Energy Efficiency Resource Standards (EERS): April 2015." Available from http://aceee.org/sites/default/files/eers-04072015.pdf. Accessed August 5, 2015.

[3] Allcott, Hunt. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* 95 (9-10): 1082–95.

[4] Allcott, Hunt. 2013. "The Welfare Effects of Misperceived Product Costs: Data and Calibrations from the Automobile Market." *American Economic Journal: Economic Policy* 5 (3): 30-66.

[5] Allcott, Hunt. 2015. "Site Selection Bias in Program Evaluation." *Quarterly Journal of Economics* 130 (3): 1117-1165.

[6] Allcott, Hunt, and Michael Greenstone. 2015. "Measuring the Welfare Effects of Energy Efficiency Programs." Working Paper, New York University.

[7] Allcott, Hunt, and Todd Rogers. 2014. "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review* 104 (10): 3003-37.

[8] Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi. 2014. "The Effect of Financial Education on the Quality of Decision Making." NBER Working Paper No. 20618 (October).

[9] Andreoni, James, Justin M. Rao, and Hannah Trachtman.. 2011. "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving." NBER Working Paper No. 17648.

[10] Ashby, Kira, Hilary Forster, Bruce Ceniceros, Bobbi Wilhelm, Kim Friebel, Rachel Henschel, and Shahana Samiullah. 2012. "Green with Envy: Neighbor Comparisons and Social Norms in Five Home Energy Report Programs." http://www.aceee.org/files/proceedings/2012/data/papers/0193-000218.pdf (accessed September 2013).

[11] Ashley, Elizabeth M., Clark Nardinelli, and Rosemarie A. Lavaty. 2015. "Estimating the Benefits of Public Health Policies that Reduce Harmful Consumption." *Health Economics* 24 (5): 617-624.

[12] Athey, Susan, and Guido Imbens. 2015. "Machine Learning Methods for Estimating Heterogeneous Causal Effects." Working Paper, Stanford University.

[13] Attari, Shahzeen, Michael DeKay, Cliff Davidson, and Wandi Bruine de Bruin. 2010. "Public Perceptions of Energy Consumption and Savings." *Proceedings of the National Academy of Sciences* 107 (37): 16054–16059.

[14] Ayres, Ian, Sophie Raseman, and Alice Shih. 2013. "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." *Journal of Law, Economics, and Organization* 29 (5): 992–1022.

[15] Beauchamp, Jonathan P., Daniel J. Benjamin, Christopher Chabris, and David I. Laibson. 2015. "Controlling for Compromise Effects Debiases Estimates of Preference Parameters." Working Paper, Harvard University.

[16] Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least Squares After Model Selection in High-Dimensional Sparse Models." *Bernoulli* 19 (2): 521-547.

[17] Bernheim, B. Douglas, Andrey Fradkin, and Igor Popov. 2015. "The Welfare Economics of Default Options in 401(k) Plans." *American Economic Review*, 105 (9): 2798-2837.

[18] Borenstein, Severin, and Lucas W. Davis. 2012. "The Equity and Efficiency of Two-Part Tariffs in U.S. Natural Gas Markets." *Journal of Law and Economics*, 55 (1): 75-128.

[19] Bronnenberg, Bart, Jean-Pierre Dube, Matthew Gentzkow, and Jesse Shapiro. 2015. "Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium." *Quarterly Journal of Economics*, 130 (4): 1669-1726.

[20] Brunnermeier, Markus K., and Jonathan A. Parker. 2005. "Optimal Expectations." *American Economic Review* 95 (4): 1092-1118.

[21] Caplin, Andrew (2003). "Fear as a Policy Instrument." In George Loewenstein, Daniel Read, and Roy Baumeister, Eds, Time and Decision: Economic and Psychological perspectives in Intertemporal Choice. New York, NY: Russel Sage.

[22] Carroll, Gabriel, James Choi, David Laibson, Brigitte Madrian, and Andrew Metrick. 2009. "Optimal Defaults and Active Decisions: Theory and Evidence from 401(k) Saving." *Quarterly Journal of Economics* 124 (4): 1639-1674.

[23] Chaloupka, Frank J., Kenneth E. Warner, Daron Acemoglu, Jonathan Gruber, Fritz Laux, Wendy Max, Joseph Newhouse, Thomas Schelling, and Jody Sindelar. 2014. "An Evaluation of The FDA's Analysis of the Costs and Benefits of the Graphic Warning Label Regulation." *Tobacco Control* 0: 1-8.

[24] Chaloupka, Frank J., Gruber, Jonathan & Warner, Kenneth E. 2015. "Accounting for "Lost Pleasure" in a Cost–Benefit Analysis of Government Regulation: The Case of the FDA's Proposed Cigarette Labeling Regulation." *Annals of Intern Medicine*, 162 (1): 64.

[25] Costa, Dora L., and Matthew E. Kahn. 2013. "Energy Conservation 'Nudges' and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." *Journal of the European Economic Association* 11 (3): 680–702.

[26] Cutler, David M., Amber Jessup, Donald Kenkel, and Martha A. Starr. 2015. "Valuing Regulations Affecting Addictive or Habitual Goods." *Journal of Benefit-Cost Analysis* 6 (2): 247-280.

[27] Davis, Lucas, and Erich Muehlegger. 2010. "Do Americans Consume Too Little Natural Gas? An Empirical Test of Marginal Cost Pricing." *RAND Journal of Economics* 41 (4): 791-801.

[28] DellaVigna, Stefano, John List, and Ulrike Malmendier. 2012. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics* 127 (1): 1-56.

[29] Delmas, Magali A., Miriam Fischlein, and Omar I. Asensio. 2013. "Information Strategies and Energy Conservation Behavior: A Meta-Analysis of Experimental Studies from 1975 to 2012." *Energy Policy* 61: 729-739.

[30] DiNardo, John, Justin McCrary, and Lisa Sanbonmatsu. 2006. "Constructive Proposals for Dealing with Attrition: An Empirical Example." Working Paper.

[31] Dolan, Paul and Robert Metcalfe. 2013. "Neighbors, Knowledge, and Nuggets: Two Natural Field Experiments on the Role of Incentives on Energy Conservation." Centre for Economic Performance Discussion Paper No. 1222.

[32] EOP (Executive Office of the President). 2015. "Executive Order – Using Behavioral Science Insights to Better Serve the American People." Available from https://www.whitehouse.gov/the-press-office/2015/09/15/executive-order-using-behavioral-science-insights-better-serve-american. Accessed September 26, 2015.

[33] EPA (U.S. Environmental Protection Agency). 1995. "AP 42, Fifth Edition: Compilation of Air Pollutant Emission Factors, Volume 1: Stationary Point and Area Sources. Chapter 1: External Combustion Sources." Available from http://www.epa.gov/ttn/chief/ap42/ch01/index.html. Accessed June 12, 2015.

[34] Ericson, Keith M. Marzilli. 2014. "Consumer Inertia and Firm Pricing in the Medicare Part D Prescription Drug Insurance Exchange." *American Economic Journal: Economic Policy*, 6(1): 38-64.

[35] Farhi, Emmanuel, and Xavier Gabaix. 2015. "Optimal Taxation with Behavioral Agents." Working Paper, New York University (August).

[36] FDA (U.S. Food and Drug Administration). 2011. "Required Warnings for Cigarette Packages and Advertisements, Final Rule." *Federal Register*, 76, 36628–36777.

[37] Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2015. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." Working Paper, University of Chicago (June).

[38] Glaeser, Edward. 2006. Paternalism and Psychology. *University of Chicago Law Review.* 73:133-156.

[39] Glaeser, Edward. 2014. The Supply of Environmentalism: Psychological Interventions and Economics. *Review of Environmental Economics and Policy.* 8(2):208-229.

[40] Handel, Benjamin R. 2013. "Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts." *American Economic Review* 103 (7): 2643-2682.

[41] Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47 (1): 153-161.

[42] Herberich, David H., John A. List, and Michael K. Price. 2012. "How Many Economists Does It Take to Change a Light Bulb? A Natural Field Experiment on Technology Adoption." Working Paper.

[43] Holland, Stephen, Erin Mansur, Nicholas Muller, and Andrew Yates. 2015. "Measuring the Spatial Heterogeneity in Environmental Externalities from Driving: A Comparison of Gasoline and Electric Vehicles." NBER Working Paper No. 21291 (June).

[44] Howarth, Robert, Drew Shindell, Renee Santoro, Anthony Ingraffea, Nathan Phillips, and Amy Townsend-Small. 2012. "Methane Emissions from Natural Gas Systems." Available from http://www.eeb.cornell.edu/howarth/publications/Howarth_et_al_2012_National_Climate_Assessment.pdf. Accessed August 15, 2015.

[45] Integral Analytics. 2012. "Sacramento Municipal Utility District Home Energy Report Program." http://www.integralanalytics.com/ia/Portals/0/FinalSMUDHERSEval2012v4.pdf (accessed September 2013).

[46] Interagency Working Group on the Social Cost of Carbon, United States Government. 2013. "Technical Support Document: Technical Update of the Social Cost of Carbon for Regulatory Impact Analysis Under Executive Order 12866." Available from

https://www.whitehouse.gov/sites/default/files/omb/inforeg/social_cost_of_carbon_for_ria_2013_update.pdf. Accessed July 26, 2015.

[47] Ito, Koichiro. 2014. "Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing." *American Economic Review* 104 (2): 537-563.

[48] Ito, Koichiro, Tanakori Ida, and Makoto Tanaka. 2015. "The Persistence of Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand." Working Paper, University of Chicago (April).

[49] Jin, Lawrence, Don Kenkel, Feng Liu, and Hua Wang. 2015. "Retrospective and Prospective Benefit-Cost Analyses of U.S. Anti-Smoking Policies." *Journal of Benefit-Cost Analysis* 6 (1): 154-186.

[50] Kamenica, Emir. 2008. "Contextual Inference in Markets: On the Informational Content of Product Lines." American Economic Review 98 (5): 2127-2149.

[51] Kantola, S. J., G. J. Syme, and N. A. Campbell. 1984. "Cognitive Dissonance and Energy Conservation." *Journal of Applied Psychology* 69 (3): 416-421.

[52] Keuring van Elektrotechnische Materialen te Arnhem (KEMA). 2012. "Puget Sound Energy's Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation." Madison, Wisconsin: DNV KEMA Energy and Sustainability.

[53] Kling, Jeffrey R., Sendhil Mullainathan, Eldar Shafir, Lee C. Vermeulen, and Marian V. Wrobel. 2012. "Comparison Friction: Experimental Evidence from Medicare Drug Plans." *Quarterly Journal of Economics* 127 (1): 199-235.

[54] Koszegi, Botond, and Adam Szeidl. 2013. "A Model of Focusing in Economic Choice." *Quarterly Journal of Economics* 128 (1): 53–104.

[55] Kushler, Martin, Seth Nowak, and Patti Witte. 2012. "A National Survey of State Policies and Practices for the Evaluation of Ratepayer-Funded Energy Efficiency Programs." American Council for an Energy-Efficient Economy Report Number U122.

[56] Larrick, Richard, and Jack Soll. 2008. "The MPG Illusion." *Science* 320 (5883): 1593-1594.

[57] Madrian, Brigitte C., and Dennis F. Shea. 2001. The Power of Suggestion: Inertia in 401(k) Participation and Savings Behavior." *Quarterly Journal of Economics* 116 (4): 1149-1187.

[58] Myhre, G., D. Shindell, F.-M. Bréon, W. Collins, J. Fuglestvedt, J. Huang, D. Koch, J.-F. Lamarque, D. Lee, B. Mendoza, T. Nakajima, A. Robock, G. Stephens, T. Takemura and H. Zhang. 2013. "Anthropogenic and Natural Radiative Forcing." In: Climate Change 2013:

The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[59] Nadel, Steven, and Kenneth Keating. 1991. "Engineering Estimates vs. Impact Evaluation Results: How Do They Compare and Why?" American Council for an Energy-Efficient Economy Report Number U915.

[60] Nolan, Jessica M., Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2008. "Normative Influence is Underdetected." *Personality and Social Psychology Bulletin* 34 (7): 913–923.

[61] Opinion Dynamics. 2012. "Massachusetts Three Year Cross-Cutting Behavioral Program Evaluation Integrated Report." Waltham, MA: Opinion Dynamics Corporation.

[62] Opower. 2014. "Prospectus." Available from http://www.nasdaq.com/markets/ipos/filing.ashx?filingid=94658 Accessed September 26, 2015.

[63] Oster, Emily, Ira Shoulson, and E. Ray Dorsey. 2013. "Optimal Expectations and Limited Medical Testing: Evidence from Huntington Disease." *American Economic Review* 103 (2): 804-830.

[64] Perry, Michael, and Sarah Woehleke. 2013. "Evaluation of Pacific Gas and Electric Company's Home Energy Report Initiative for the 2010–2012 Program." San Francisco: Freeman, Sullivan, and Company.

[65] Pope, C. Arden, Richard T. Burnett, Michael J. Thun, Eugenia, E. Calle, Daniel Krewski, Kazuhiko Ito, and George D. Thurston. 2002. "Lung Cancer, Cardiopulmonary Mortality, and Long-term Exposure to Fine Particulate Air Pollution." *Journal of the American Medical Association* 287 (9): 1132-1141.

[66] Schultz, P. Wesley, Jessica M. Nolan, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science* 18 (5): 429–434.

[67] Sudarshan, Anant. 2014. "Nudges in the Marketplace: Using Peer Comparisons and Incentives to Reduce Household Electricity Consumption." Working Paper. http://www.anantsudarshan.com/uploads/1/0/2/6/10267789/nudges_sudarshan_2014.pdf.

[68] Thaler, Richard, and Cass R. Sunstein. 2008. Nudge: Improving Decisions about Health, Wealth, and Happiness. New Haven: Yale University Press.

[69] Trachtman, Hannah, Andrew Steinkruger, Mackenzie Wood, Adam Wooster, James Andreoni, James J. Murphy, and Justin M. Rao. 2015. "Fair Weather Avoidance: Unpacking the Costs and Benefits of "Avoiding the Ask."" *Journal of the Economic Science Association* 1 (1): 8-14.

[70] Violette, Daniel, Provencher, Bill, and Mary Klos. 2009. "Impact Evaluation of Positive Energy SMUD Pilot Study." Boulder, CO: Summit Blue Consulting, LLC.

[71] Watson, David, Lee Anna Clark, and Auke Tellegen. 1988. "Development and Validation of Brief Measures of Positive and Negative Affect: The PANAS Scales." *Journal of Personality and Social Psychology* 54: 1063-1070.

[72] Weimer, David L., Aidan R. Vining, and Randall K. Thomas. 2009. "Cost-Benefit Analysis Involving Addictive Goods: Contingent Valuation to Estimate Willingness-to-Pay for Smoking Cessation." *Health Economics* 18 (2): 181-202.

[73] Whitehead, Mark, Rhys Jones, Rachel Howell, Rachel Lilley, and Jessica Pykett. 2014. "Assessing the Global Impact of the Behavioral Sciences on Public Policy." https://changingbehaviours.files.wordpress.com/2014/09/nudgedesignfinal.pdf

Online Appendix: Not for Publication

The Welfare Effects of Nudges: A Case Study of Energy Use Social Comparisons

Hunt Allcott and Judd B. Kessler

A Phone Survey Questionnaire

Below is the phone survey questionnaire. Programming notes and comments are in italics. Bolded headers are for organizational purposes and were not read.

Introduction

Hi. I am calling on behalf of Central Hudson Gas and Electric, your local utility. Central Hudson has been sending you Home Energy Reports since last fall, and we want to know what you think about them. Do you have about two minutes to answer some questions? If yes, Central Hudson will send you a check for up to $10.

If asked, "What is a Home Energy Report?", say: "Home Energy Reports are one-page letters that compare your natural gas use to your neighbors' use and provide energy conservation tips. Central Hudson sent up to four of these reports to the address on the account associated with this phone number between late fall 2014 and early spring 2015. Do you recall receiving any Home Energy Reports in the past nine months?"

- *If "Yes", continue to Question 1.*
- *If "No", or if the customer otherwise says "I don't remember receiving any Home Energy Reports," say:* "Is there someone else in the household who may have seen these reports come in the mail? If so, may I speak to him or her?" *If there is no one else who might have seen the reports, terminate call and code response as "Does not remember Home Energy Reports." If there is someone else but not available, record that person's name and attempt to call him/her later.*

If the caller indicates that he/she has already answered these questions in a mail survey, then skip questions 1 and 2 and say: "Thank you for responding to our mail survey. We have a couple of follow-up questions that are better to ask by phone." *Then continue to Question 3.*

Question 1

To start, I'm going to ask three questions where you'll choose between some combination of continuing Home Energy Reports and receiving checks for different amounts of money. These are unusual questions, but they're designed to tell us how much you value the Reports. These are real questions: Central Hudson will use a lottery to pick one question and will actually mail you what you chose, so please answer carefully.

Survey Version B only: "Remember that Home Energy Reports compare your energy use to your neighbors' use.

Survey Version C only: "Remember that Home Energy Reports help you to reduce your environmental impact."

a. Which would you prefer: 4 more Home Energy Reports PLUS a $10 check, OR a $1 check?

b. Which would you prefer: 4 more Home Energy Reports PLUS a $10 check, OR a $5 check?

c. Which would you prefer: 4 more Home Energy Reports PLUS a $10 check, OR a $9 check?

d. Which would you prefer: 4 more Home Energy Reports PLUS a $10 check, OR a $10 check?

e. Which would you prefer: 4 more Home Energy Reports PLUS a $9 check, OR a $10 check?

f. Which would you prefer: 4 more Home Energy Reports PLUS a $5 check, OR a $10 check?

g. Which would you prefer: 4 more Home Energy Reports PLUS a $1 check, OR a $10 check?

If consumers have consistent preferences, we would not need to ask all seven MPL questions because answers to some imply answers to others. Questions 1a-1g were asked in the following order:

Ask 1d first

If 1d="HER+$10", then 1f

If 1f="HER+$5", then 1g

If 1f="$10", then 1e

If 1d="$10", then 1b

If 1b="HER+$10", then 1c

If 1b="$5", then 1a

Question 2

Think back to when you received your first Home Energy Report. Did the Report say that you were using more or less energy than you thought?

a. Much less than I thought

b. Somewhat less than I thought

c. About what I thought

d. Somewhat more than I thought

e. Much more than I thought

Question 3

Do you think that receiving four more Home Energy Reports this fall and winter would help you reduce your natural gas use by even a small amount?

a. Yes

b. No

If Yes: How much money do you think you would save on your natural gas bills if you receive four more Reports compared to if you do not receive them?

If necessary: "We just want to know your best guess."

Note to enumerators: Prompt for a dollar value, not a percentage. If necessary: "I'm supposed to ask for your best guess of how many dollars you'd save in total."

Question 4

Since last fall, Central Hudson sent up to four Home Energy Reports to many households like yours. For the average household, how much money do you think these Reports have helped them save on their natural gas bills?

If necessary: "We just want to know your best guess."

Note to enumerators: Prompt for a dollar value, not a percentage. If necessary: "I'm supposed to ask for your best guess of total dollar savings since last fall."

Question 5

How would you like the Reports if they did not have the bar graph comparing your energy use to your neighbors' use?

a. Much less

b. Somewhat less

c. About the same

d. Somewhat more

e. Much more

Question 6

Some people feel either inspired or pressured when they see their Home Energy Reports. Did you feel inspired, pressured, neither, or both?

a. Inspired

b. Pressured

c. Neither

d. Both

Question 7

Some people feel either proud or guilty when they see their Home Energy Reports. Did you feel proud, guilty, neither, or both?

a. Proud

b. Guilty

c. Neither

d. Both

Question 8

To what extent do you agree or disagree with the following statement: "The Home Energy Reports gave useful information that helped me conserve energy."

a. Strongly agree

b. Agree

c. Neither

d. Disagree

e. Strongly disagree

Question 9

Do you have any other comments about the Home Energy Reports that you'd like to share?

Open response, please write down as much as possible.

B Data Appendix

Table A1: **Balance Tests (Page 1)**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Dependent variable:	Baseline use (ccf/day)	ln(Income)	ln(Net worth)	ln(House value)	Education (years)	Male	ln(Age)	Retired	Married	Rent
				Panel A: Home Energy Report Recipient/Control						
Recipient	0.013	-0.011	0.0058	-0.029	-0.032	-0.0026	-0.0044	0.0028	-0.0064	0.0039
	(0.024)	(0.012)	(0.024)	(0.029)	(0.035)	(0.0077)	(0.0051)	(0.0030)	(0.0079)	(0.0069)
Observations	19,921	19,927	15,557	16,741	19,475	16,811	17,282	16,728	15,406	17,561
				Panel B: Survey Group						
Mail follow-up	-0.027	0.0073	-0.054	-0.043	-0.0076	0.0093	-0.0092	-0.012	0.0068	-0.012
	(0.037)	(0.018)	(0.036)	(0.044)	(0.054)	(0.012)	(0.0078)	(0.0050)**	(0.012)	(0.011)
Comparison cue	-0.039	0.00063	-0.043	-0.062	-0.025	-0.014	0.0015	0.0057	-0.00061	-0.011
	(0.043)	(0.021)	(0.042)	(0.051)	(0.063)	(0.014)	(0.0090)	(0.0054)	(0.014)	(0.012)
Environmental cue	0.011	0.0056	0.012	-0.049	0.0016	-0.018	0.015	0.011	0.0058	-0.0100
	(0.043)	(0.021)	(0.042)	(0.051)	(0.063)	(0.014)	(0.0090)*	(0.0055)**	(0.014)	(0.012)
Observations	9436	9439	7466	7965	9226	8036	8251	8004	7255	8371
F-test p-value	0.54	0.97	0.26	0.45	0.97	0.46	0.18	0.023	0.91	0.53

Notes: This table presents tests of balance on observables between randomly-assigned groups. Samples in Panel A include the full HER recipient and control groups, while samples in Panel B are limited to the households that were sent Home Energy Reports and were thus eligible for our surveys. Observation counts differ between columns because regressions include only non-missing observations of the dependent variable. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A2: **Balance Tests (Page 2)**

	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)
Dependent variable:	Single family home	ln(House age)	Democrat	Hybrid auto share	Green consumer	Wildlife donor	Profit score	Buyer score	Mail responder	Home improvement interest
					Panel A: Home Energy Report Recipient/Control					
Recipient	0.0018	-0.032	0.0018	0.021	0.0042	0.0033	-0.0033	-0.022	-0.012	0.00024
	(0.0070)	(0.016)*	(0.0082)	(0.047)	(0.0052)	(0.0036)	(0.014)	(0.016)	(0.0069)*	(0.0051)
Observations	17,734	14,885	18,080	19,728	18,883	16,728	19,784	14,967	17,734	16,728
					Panel B: Survey Group					
Mail follow-up	-0.0097	-0.021	0.0037	0.0013	-0.0053	-0.0094	0.018	-0.017	0.0017	-0.025
	(0.011)	(0.025)	(0.013)	(0.053)	(0.0081)	(0.0058)	(0.022)	(0.025)	(0.011)	(0.0081)***
Comparison cue	0.0048	-0.047	-0.00065	-0.10	0.0058	-0.0035	-0.022	0.0036	-0.012	-0.0037
	(0.012)	(0.029)	(0.015)	(0.071)	(0.0094)	(0.0066)	(0.025)	(0.029)	(0.012)	(0.0092)
Environmental cue	0.011	-0.048	0.0038	-0.18	-0.0056	-0.0045	0.010	-0.014	-0.0079	-0.0015
	(0.012)	(0.029)*	(0.015)	(0.13)	(0.0092)	(0.0066)	(0.026)	(0.029)	(0.012)	(0.0092)
Observations	8464	7109	8617	9340	8977	8004	9377	7143	8464	8004
F-test p-value	0.67	0.23	0.98	0.35	0.59	0.36	0.49	0.84	0.80	0.020

Notes: This table presents tests of balance on observables between randomly-assigned groups. Samples in Panel A include the full HER recipient and control groups, while samples in Panel B are limited to the households that were sent Home Energy Reports and were thus eligible for our surveys. Observation counts differ between columns because regressions include only non-missing observations of the dependent variable. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A3: **Survey Response Counts by Attempt**

	(1)	(2)
Attempt	Mail	Phone
1	402	523
2	497	358
3		229
4		172
5		163
6		83
7		80
8		80
Overall	899	1690

Notes: For the mail survey, attempt 1 refers to the survey included in the final Home Energy Report, and attempt 2 refers to the follow-up survey sent to 2/3 of households. For the phone survey, attempt refers to the number of times that the phone number was called before completing the survey.

Table A4: **Correlations of Willingness-to-Pay with Qualitative Survey Responses**

	(1)	(2)	(3)	(4)	(5)	(6)
Predicted savings	0.11					
	(0.0089)***					
Like without comparisons		-1.05				
		(0.16)***				
Useful info			2.26			
			(0.18)***			
Inspired				3.36		
				(0.38)***		
Pressured				-1.02		
				(0.50)**		
Proud					1.18	
					(0.41)***	
Guilty					1.39	
					(0.49)***	
Positive comment						4.34
						(0.44)***
Observations	1365	1581	1570	1571	1571	2137
R^2	0.094	0.026	0.093	0.047	0.011	0.042

Notes: Data are the unweighted sample of phone survey responses. Dependent variable is willingness-to-pay. The independent variables in columns 1-6 are from questions 3, 5, 8, 6, 7, and 9, respectively. Predicted savings is winsorized at $50. Columns 2 and 3 consider the five-point Likert scale responses to questions 5 and 8, which we code as integers $\{-2, -1, 0, 1, 2\}$. The sample in column 6 includes both mail and phone survey respondents: the phone survey enumerators transcribed responses to question 9, and we also transcribed the 30 unsolicited comments written on the mail survey. The variable "Positive comment" takes value 1 for positive comments about HERs, -1 for negative comments, and 0 for neutral or no comments. Sample sizes vary due to item non-response. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A5: **Correlations of Negative Willingness-to-Pay with Qualitative Survey Responses**

	(1)	(2)	(3)	(4)	(5)	(6)
Predicted savings	-0.0059					
	(0.00057)***					
Like without comparisons		0.059				
		(0.011)***				
Useful info			-0.14			
			(0.011)***			
Inspired				-0.18		
				(0.024)***		
Pressured				0.10		
				(0.034)***		
Proud					-0.11	
					(0.027)***	
Guilty					-0.026	
					(0.033)	
Positive comment						-0.22
						(0.026)***
Observations	1365	1581	1570	1571	1571	2137
R^2	0.070	0.019	0.089	0.037	0.011	0.025

Notes: Data are the unweighted sample of phone survey responses. Dependent variable is an indicator for negative willingness-to-pay. The independent variables in columns 1-6 are from questions 3, 5, 8, 6, 7, and 9, respectively. Predicted savings is winsorized at \$50. Columns 2 and 3 consider the five-point Likert scale responses to questions 5 and 8, which we code as integers $\{-2, -1, 0, 1, 2\}$. The sample in column 6 includes both mail and phone survey respondents: the phone survey enumerators transcribed responses to question 9, and we also transcribed the 30 unsolicited comments written on the mail survey. The variable "Positive comment" takes value 1 for positive comments about HERs, -1 for negative comments, and 0 for neutral or no comments. Sample sizes vary due to item non-response. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A6: **Within-Household Correlations of Survey Responses**

	(1)	(2)	(3)	(4)
Dependent variable	WTP from first mail survey	WTP from phone survey	1(WTP from phone survey<0)	Belief update from phone survey
WTP from second mail survey	0.819			
	(0.080)***			
WTP from mail survey		0.440		
		(0.072)***		
1(WTP from mail survey<0)			0.362	
			(0.071)***	
Belief update from mail survey				0.500
				(0.064)***
Observations	87	224	224	259
R^2	0.584	0.206	0.132	0.217

Notes: The sample for column 1 is households that returned more than one mail survey with valid WTP. The sample for columns 2-4 is households that responded to both mail and phone surveys. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

C Appendix to Empirical Estimates

Table A7: **Inverse Probability Weights (Page 1)**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent var:	Have WTP from Paper	Have WTP, Assigned to Base	Have WTP, Assigned to Follow-up	Have WTP from Base Mail	Have WTP from Follow-up Mail	Have WTP from Phone	Have WTP	Have WTP; Base Mail Excluded	Received Second Year
Baseline use	-0.358	-0.0985	-0.258	-0.308	-0.0556	0.765	0.541	0.795	0.0182
	(0.161)**	(0.0592)*	(0.149)*	(0.110)***	(0.117)	(0.225)***	(0.255)**	(0.242)***	(0.0706)
ln(Income)	-0.306	-0.171	-0.115	-0.0962	-0.210	-0.624	-0.398	-0.137	-0.0492
	(0.457)	(0.175)	(0.419)	(0.309)	(0.336)	(0.661)	(0.744)	(0.716)	(0.216)
ln(Net worth)	0.0841	0.0691	0.0191	0.212	-0.116	-0.0450	-0.0159	-0.285	-0.202
	(0.289)	(0.110)	(0.267)	(0.194)	(0.214)	(0.395)	(0.445)	(0.427)	(0.135)
ln(House value)	-0.180	0.0255	-0.211	-0.0824	-0.0950	-0.117	-0.140	-0.0774	0.0807
	(0.158)	(0.0595)	(0.145)	(0.106)	(0.116)	(0.230)	(0.255)	(0.246)	(0.0770)
Education	0.554	0.133	0.410	0.295	0.240	0.416	0.776	0.533	-0.00590
	(0.110)***	(0.0388)***	(0.102)***	(0.0745)***	(0.0802)***	(0.166)**	(0.185)***	(0.179)***	(0.0474)
Male	-0.240	0.0661	-0.317	-0.250	0.00731	0.290	-0.126	-0.0476	-0.0277
	(0.550)	(0.221)	(0.500)	(0.372)	(0.403)	(0.812)	(0.906)	(0.875)	(0.248)
ln(Age)	1.598	0.328	1.257	0.799	0.747	1.069	2.187	1.438	-0.737
	(1.012)	(0.389)	(0.928)	(0.690)	(0.739)	(1.452)	(1.621)	(1.558)	(0.458)
Retired	0.593	0.207	0.323	-0.00655	0.544	1.049	1.696	2.163	-1.108
	(1.309)	(0.487)	(1.199)	(0.852)	(0.969)	(2.103)	(2.339)	(2.297)	(0.520)**
Married	-0.123	0.160	-0.265	0.0575	-0.153	-0.975	-1.384	-1.443	0.933
	(0.681)	(0.247)	(0.631)	(0.444)	(0.518)	(0.989)	(1.102)	(1.061)	(0.305)***
Rent	0.210	0.138	0.0469	0.149	0.0390	-2.209	-2.097	-2.317	-0.409
	(0.799)	(0.311)	(0.735)	(0.541)	(0.595)	(1.112)**	(1.253)*	(1.202)*	(0.308)

(table continues on next page)

Notes: This table presents probit estimates used to construct inverse probability weights. We report marginal effects, with coefficients multiplied by 100 for readability. In all columns other than column 8, the sample is all households assigned to the initial HER recipient group that did not opt out before the 2nd year. In column 8, the sample is the same except excluding households that returned the first mail survey. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A8: **Inverse Probability Weights (Page 2)**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Dependent var:	Have WTP from Paper	Have WTP, Assigned to Base	Have WTP, Assigned to Follow-up	Have WTP from Base Mail	Have WTP from Follow-up Mail	Have WTP from Phone	Have WTP	Have WTP; Base Mail Excluded	Received Second Year
Single family	0.880	0.294	0.584	0.310	0.566	-0.374	-0.304	-0.488	0.0787
	(0.708)	(0.273)	(0.649)	(0.462)	(0.539)	(1.011)	(1.129)	(1.086)	(0.290)
ln(House age)	-0.803	-0.0767	-0.717	-0.329	-0.460	-0.997	-1.513	-1.191	-0.0472
	(0.305)***	(0.124)	(0.277)***	(0.204)	(0.224)**	(0.453)**	(0.507)***	(0.492)**	(0.153)
Democrat	0.447	0.326	0.127	0.296	0.129	0.902	1.148	0.866	0.260
	(0.461)	(0.181)*	(0.419)	(0.308)	(0.339)	(0.720)	(0.798)	(0.778)	(0.204)
Hybrid auto share	0.116	0.0262	0.0857	0.0873	0.0143	0.445	0.479	0.404	-0.0331
	(0.0829)	(0.0302)	(0.0763)	(0.0505)*	(0.0656)	(0.123)***	(0.139)***	(0.135)***	(0.0342)
Green consumer	-0.470	-0.160	-0.302	-0.696	0.256	0.816	0.440	0.941	0.150
	(0.728)	(0.293)	(0.662)	(0.491)	(0.531)	(1.115)	(1.244)	(1.205)	(0.340)
Wildlife donor	3.425	1.056	2.216	2.549	0.537	3.212	5.825	3.342	0.543
	(1.151)***	(0.442)**	(1.050)**	(0.735)***	(0.870)	(1.823)*	(2.040)***	(2.019)*	(0.548)
Profit score	1.952	0.0559	1.854	0.696	1.209	1.171	2.433	1.896	-0.411
	(0.392)***	(0.138)	(0.363)***	(0.261)***	(0.291)***	(0.574)**	(0.642)***	(0.618)***	(0.176)**
Buyer score	0.858	0.185	0.655	0.586	0.267	-0.164	0.513	-0.0173	0.140
	(0.374)**	(0.141)	(0.345)*	(0.244)**	(0.282)	(0.539)	(0.607)	(0.584)	(0.160)
Mail responder	0.333	0.0517	0.278	-0.0196	0.363	-0.739	-0.359	-0.442	-0.164
	(0.638)	(0.244)	(0.585)	(0.422)	(0.473)	(0.957)	(1.065)	(1.030)	(0.289)
Home improvement	0.322	-0.322	0.620	-0.397	0.725	1.630	1.224	1.564	0.0255
	(0.892)	(0.382)	(0.800)	(0.598)	(0.648)	(1.332)	(1.500)	(1.451)	(0.401)
Observations	9948	9948	9948	9948	9948	9948	9948	9548	9948

Notes: This table presents probit estimates used to construct inverse probability weights. We report marginal effects, with coefficients multiplied by 100 for readability. In all columns other than column 8, the sample is all households assigned to the initial HER recipient group that did not opt out before the 2nd year. In column 8, the sample is the same except excluding households that returned the first mail survey. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A9: **Correlation of Willingness-to-Pay with Phone Survey Responsiveness**

	(1)	(2)
Completed survey attempt number	0.0229	0.0651
	(0.0885)	(0.0901)
Observations	1609	1609
Weights	Equal	IPW for $\mathcal{P}_n$

Notes: Dependent variable is willingness-to-pay, sample is all phone survey respondents. For the phone survey, each respondent was dialed up to eight times; the independent variable is the attempt number on which the survey was completed. Column 2 re-weights observations to match $\mathcal{P}_n$, the target population of treatment group households that did not opt out before the program's second year. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A10: **Fitting Moral Utility**

	(1)
Predicted savings	0.0870
	(0.00930)***
Inspired	2.800
	(0.452)***
Pressured	-1.224
	(0.594)**
Proud	0.119
	(0.473)
Guilty	0.662
	(0.563)
Observations	1350
R^2	0.122

Notes: Dependent variable is willingness-to-pay. Predicted savings is winsorized at $50. Sample includes only phone survey respondents with non-missing data. Observations are weighted to match the target population of treatment group households that did not opt out before the program's second year. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A11: **Effect of Survey Version on Willingness-to-Pay**

	(1)	(2)	(3)	(4)
Comparison version	-0.686	-0.693	-0.661	-0.658
	(0.386)*	(0.384)*	(0.390)*	(0.387)*
Environmental version	-0.211	-0.155	-0.190	-0.143
	(0.386)	(0.388)	(0.394)	(0.399)
Mean comparison			0.106	0.0391
			(0.215)	(0.286)
Comparison version×Mean comparison			0.111	0.184
			(0.317)	(0.319)
Environmental version×Mean comparison			0.0530	0.0403
			(0.347)	(0.354)
Observations	2137	2137	2137	2137
Include **X** covariates	No	Yes	No	Yes

Notes: Dependent variable is willingness-to-pay. "Mean comparison" is the average difference (in 1000s of ccf) between own natural gas usage and mean neighbor usage on the HERs in winter 2014-2015. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

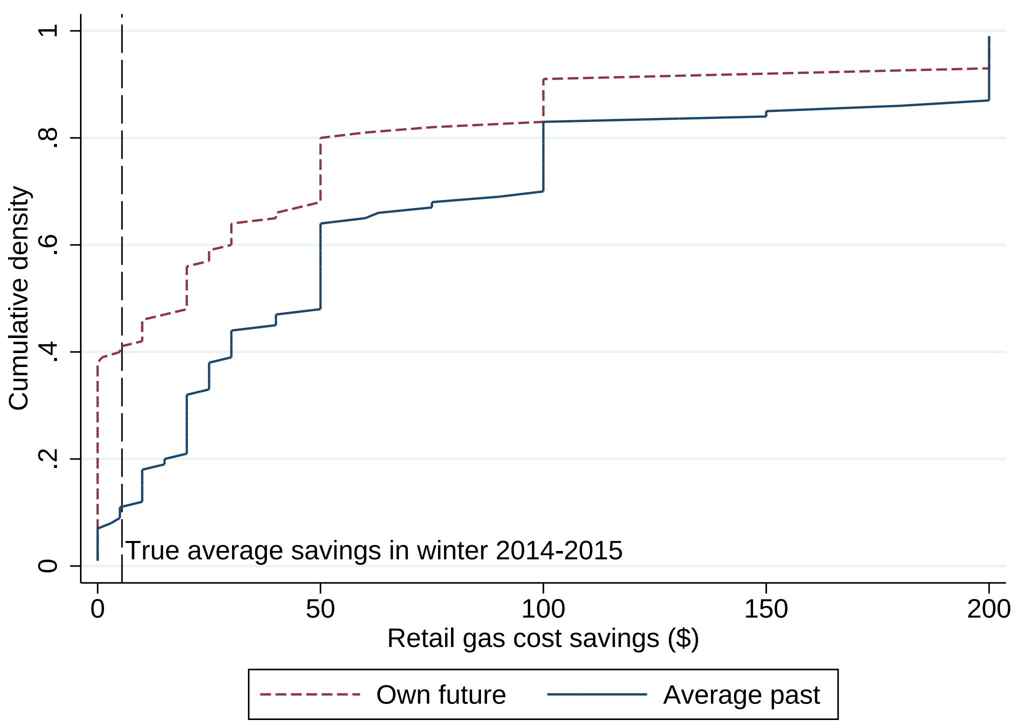
D Appendix to Welfare Estimates

D.A Testing for Biased Beliefs and Overoptimism

For the welfare analysis, we assume that WTP equals consumer utility gain. In this context, we could imagine two reasons why this might fail: biased beliefs and overoptimism.

By biased beliefs, we mean that consumers might systematically underestimate or overestimate the energy cost savings resulting from their conservation efforts. Consumers likely know the monetary and non-monetary costs of their efforts, such as the time to adjust the thermostat or the money to install energy-saving windows, but resulting energy savings can be quite difficult to infer given that gas bills fluctuate substantially across months and years. There is empirical evidence to support this concern: Pronin, Berger, and Molouki (2007) and Nolan *et al.* (2008) find that people underestimate the motivational power of social norm messaging, and Larrick and Soll (2008), Attari *et al.* (2010), and Allcott (2013) explore various belief biases related to energy costs.

Figure A1: **Beliefs About Savings Caused by Home Energy Reports**



Notes: This figure presents the unweighted distribution of responses to the following phone survey questions: "How much money do you think you would save on your natural gas bills if you receive four more Reports?" and "For the average household, how much money do you think these Reports have helped them save on their natural gas bills?" True average savings in winter 2014-2015 was $5.46 per household.

To test this, the phone survey asked respondents how much money they thought they would save on their natural gas bills if they received four more HERs, as well as how much money they thought the average HER recipient had saved since last fall. Figure A1 shows that both the median and mean respondents overstate gas cost savings relative to the true average treatment effect. This suggests that if anything, biased beliefs could bias WTP upward instead of downward. However, we treat this result very cautiously, given that these questions were not incentive-compatible and stated beliefs are highly dispersed.

A second and more controversial reason why WTP might not equal consumer welfare gain has to do with overoptimism bias. Oster, Shoulson, and Dorsey (2013) show that people at high risk of Huntington disease do not get tested despite the fact that knowledge of disease status leads to very different life choices. They propose a model based on Brunnermeier and Parker (2005) in which people optimally choose beliefs while trading off the utility gain from optimistic beliefs with the utility loss from suboptimal actions. Bracha and Brown (2012) develop an alternative model in which overoptimism is constrained by the cost of holding incorrect beliefs. Evaluating information provision in these models requires the analyst to take a stand on whether to recognize overoptimistic beliefs as true utility. In these models, overoptimistic consumers may not experience a utility gain from exogenously-provided information, even though it would lead to more accurate beliefs and (in Brunnermeier and Parker's model) improved decision making. If current Home Energy Report recipients derive utility from incorrectly believing that they use less energy than their neighbors and want to be overoptimistic about their relative energy use in the future, this might reduce their WTP for HERs, and perhaps the utility loss from correcting overoptimism should not be counted as a "true" utility loss.

Even without taking a stand on this issue, we can provide suggestive tests of whether overoptimism affects WTP. On both the mail and phone surveys, we asked people whether their first HER told them they were using more or less energy than they thought. We hypothesize that people who want to be overoptimistic in the future are more likely to have been overoptimistic in the past. The initial belief update should thus be negatively correlated with WTP if overoptimism affects WTP. People gave meaningful responses: the belief update variable is positively correlated with baseline usage, usage relative to neighbors on the first HER, and reporting that they would like the HERs more if they did not have social comparisons. More people report underestimating their energy use than report overestimating. However, Appendix Table A12 shows that the belief update is not associated with WTP, either unconditionally or conditional on X.

Table A12: **Correlation of Willingness-to-Pay with Pre-Treatment Optimism**

	(1)	(2)
Belief update	0.0540	0.0532
	(0.134)	(0.138)
Observations	2102	2102
Include X covariates	No	Yes

Notes: This table presents regressions of WTP on the belief update using unweighted responses from both mail and phone surveys. Belief update is from question 8 on the mail survey and question 2 on the phone survey: "Think back to when you received your first Home Energy Report. Did the Report say that you were using more or less energy than you thought?" Responses are on a five-point Likert-style scale from "much less than I thought" to "much more than I thought," and we code these as integers from -2 to +2. Robust standard errors in parentheses. *, **, ***: statistically significant with 90, 95, and 99 percent confidence, respectively.

D.B Program Implementation Cost

Home Energy Report programs have setup costs, per-household marginal costs, and annual fixed costs. In evaluating a program's second year, we ignore setup costs. Panel A of Table A13 presents the per-household annual marginal costs. Based on a high-volume price quote from PFL (www.PrintingForLess.com), we assume \$0.4926 per HER for printing and mailing. This uses the appropriate printing and paper quality, production speed, and shipping method for HERs. HER recipients occasionally call the utility to ask questions, complain, or opt out of HERs. Opower data show that HER recipients typically call with 0.5 percent probability per year and that these calls cost the utility \$5 per call to answer. We estimate \$0.01 per household for server space to store data, and \$0.05 to purchase household-level demographic data to enhance the HERs. Overall, we estimate that the per-household marginal cost for one year of a program involving four HERs is \$2.06.

Panel B of Table A13 presents the per-utility annual costs that are fixed with respect to the number of households. Opower reported an estimated 51 hours of program design and reporting time for a client like Central Hudson. In addition, Central Hudson and Opower have in-person meetings approximately every quarter, and short phone meetings most weeks. We assume that Opower staff cost \$85 per hour, on the basis of a \$118,097 nationwide median annual salary for "program managers" (see http://www1.salary.com/Program-Manager-Salary.html) multiplied by a 1.5 loading factor to account for health insurance, vacation, and other benefits and divided by 2080 hours per year. Central Hudson reported to us that their fully-loaded staff time for this project costs \$62.64 per hour. Total utility-level fixed costs are \$16,339.

Central Hudson has four HER programs — the natural gas program we study, plus three others — with a total of about 100,000 households in treatment. Some of the per-utility fixed costs such as program design and reporting likely would increase with the number of programs, whereas others such as travel time for quarterly meetings likely would not. If the fixed cost is allocated equally to each of Central Hudson's 100,000 recipient households, this gives \$0.16 per household. Given that the second year of the program we study includes 9948 households that were allocated to the treatment group and did not opt out, this would sum to \$1625. Alternatively, if the fixed cost is allocated equally to each program, this is \$4,085 per program. Allocating this \$4,085 equally across the 9948 recipient households gives \$0.41 per household.

Table A13: **Implementation Cost Estimates**

Item	Explanation	Cost ($)
Panel A: Per-Household Annual Marginal Costs		
Printing and mailing	$0.4926/HER × 4 HERS	1.97
Utility call center	0.5% call probability × $5/call	0.025
Server space	$0.01 per recipient household	0.01
Demographic data	$0.05 per recipient household	0.05
Total		*2.06*
Panel B: Per-Utility Annual Fixed Costs		
	Opower ($85/hour)	
Program design and reporting	51 hours	4,335
Quarterly meetings (time)	8 hours/quarter× 2 people	5,440
Quarterly meetings (travel)	$250/quarter × 2 people	2,000
Weekly phone meetings	20 minutes/week × 1 person	1,473
	Central Hudson ($62.64/hour)	
Quarterly meetings	2 hours/quarter× 4 people	2,004
Weekly phone meetings	20 minutes/week × 1 person	1,086
Total		*16,339*
Annual fixed cost per household	Central Hudson has ∼100,000 HER recipients	0.16
Annual fixed cost per program	Central Hudson has four HER programs	4,085

Notes: This table presents the implementation costs for an ongoing Opower Home Energy Report program. See text for details.

D.C Speculative Evaluation of a Typical Full Opower Program

In this appendix, we address two shortcomings of the welfare evaluation in Table 8. First, Table 8 evaluates only the second year of an Opower program. Second, it evaluates a particular Opower program, which may or may not be typical.

Table A14 evaluates the full course of a typical Home Energy Report program. We use the energy savings from "site 2" studied by Allcott and Rogers (2014), an electricity-focused program with savings approximately equal to the average savings of other Opower programs. Using Table 8 from Allcott and Rogers (2014), four years of Home Energy Reports are projected to save 1875 kilowatt-hours (kWh) in total, including significant savings after the program ends. At the 2014 national average electricity price of $0.125/kWh, this amounts to $234 dollars, as shown in Panel A.[30] We assume that the long-run marginal source of electricity is a combined cycle gas plant, with cost and heat rate characteristics from the U.S. Energy Information Administration's Annual

[30]See http://www.eia.gov/electricity/monthly/pdf/epm.pdf.

Energy Outlook.[31] This gives energy acquisition cost savings of \$176 and externality reduction of \$53, using the externality damage assumptions detailed in the body of this paper.

For implementation cost, we use the price that Opower charges utilities, which is about \$8 per household per year for six HERs. We assume that this covers costs to set up and operate the HER program as well as relevant overhead costs for sales, marketing, and research and development.[32]

Panel B shows the consumer welfare and social welfare effect of the program under two assumptions. In column 1, we ignore non-energy costs, assuming that $\Delta V = -\Delta\tilde{e} \cdot p_e$. In column 2, we adjust for non-energy costs using our estimate that $\Delta V \approx -\Delta\tilde{e} \cdot p_e \times 0.54$. By the "total resource cost" metric, which is what regulators use to evaluate energy efficiency programs, failing to adjust for non-energy costs overstates gains by a factor of 4.0. Similarly, failing to adjust for non-energy costs overstates social welfare gains by a factor of 2.2. We label this calculation as "speculative" because it hinges on the assumption that $\Delta V \approx -\Delta\tilde{e} \cdot p_e \times 0.54$.

Table A14: **Social Welfare Effects of a Full Home Energy Report Program**

	(1)	(2)
Panel A: Benefits and Costs Other than Consumer Welfare (\$/recipient)		
Implementation cost: C_n, P_n	32	
Retail energy savings: $-\Delta\tilde{e} \cdot p_e$	234	
Energy acquisition cost savings: $-\Delta\tilde{e} \cdot c_e$	176	
Utility net revenue loss: $-\Delta\tilde{e} \cdot (p_e - c_c)$	58	
Externality reduction: $-\Delta\tilde{e} \cdot \phi_e$	53	
Panel B: Consumer Welfare and Social Welfare Effect (\$/recipient)		
Assumption:	$\Delta V = -\Delta\tilde{e} \cdot p_e$	$\Delta V \approx -\Delta\tilde{e} \cdot p_e \times 0.54$
Consumer welfare effect: ΔV	234	126
Total resource cost: $\Delta V + \pi_e \Delta\tilde{e} - P_n$	144	36
Social welfare effect: $\Delta V - C_n + (\pi_e - \phi_e)\Delta\tilde{e}$	197	89

Notes: See text for details.

[31] http://www.eia.gov/forecasts/aeo/electricity_generation.cfm

[32] Opower (2014) reports that the company has a 65 percent gross margin in 2013, which would suggest that the price overstates program implementation cost. On the other hand, the company operated at a net loss through that year, suggesting that the gross margin is actually not sufficient to cover sales, marketing, R&D, and other relevant overhead.

E Machine Learning Algorithm

We use a simple forward stepwise regression approach, with separate regressions to predict energy savings and WTP. To predict energy savings, we modify Equation (8) to allow heterogeneous treatment effects:

$$Y_{it} = \tau_x S_{4t} T_i \tilde{\mathbf{X}}_i + \lambda_1 S_{4t} \tilde{\mathbf{X}}_i + \lambda_2 \tilde{\mathbf{X}}_i + \sum_s \tau_s S_{st} T_i + \nu_m \tilde{Y}_{it} + \omega_m + \varepsilon_{it}, \tag{11}$$

where $\tilde{\mathbf{X}}_i$ is a vector of covariates, and S_{4t} is an indicator for whether bill date t is in winter 2015-2016. To predict WTP, we simply regress WTP w_i on $\tilde{\mathbf{X}}_i$:

$$w_i = \beta_x \tilde{\mathbf{X}}_i + \epsilon_i \tag{12}$$

We divide the 19,929-household sample into $K = 5$ random partitions. We define δ^0 as random assignment — the decision rule based on an empty set of predictors x^0. Then, for covariates $j = 1, ..., J$, we:

1. For each X variable x
   (a) For $k = \{1, K\}$
      i. Set $\tilde{\mathbf{X}}_i = \{x, x^{j-1}\}$ and estimate τ_x and β_x in the $K-1$ training sets excluding k
      ii. Use those coefficients to predict $\hat{L}_i|\{x, x^{j-1}\}$ out of sample in partition k
   (b) Propose $\delta_i^{jx} = 1$ for 1/2 of observations with largest $\hat{L}_i|\{x, x^{j-1}\}$
   (c) Estimate $L(\delta^{jx})$ in full dataset using proposed δ^{jx}
2. Choose the x^* with highest $L(\delta^{jx})$. Call this δ^j and $L(\delta^j)$
3. If $L(\delta^j) > L(\delta^{j-1})$, increment to $j+1$ and set $x^j = \{x^*, x^{j-1}\}$
   (a) Else, stop: δ^{j-1} is optimal decision rule