American Economic Journal: Applied Economics 2019, 11(1): 236–276
https://doi.org/10.1257/app.20170328

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

By Hunt Allcott and Judd B. Kessler*

"Nudge"-style interventions are often deemed successful if they generate large behavior change at low cost, but they are rarely subjected to full social welfare evaluations. We combine a field experiment with a simple theoretical framework to evaluate the welfare effects of one especially policy-relevant intervention, home energy social comparison reports. In our sample, the reports increase social welfare, although traditional evaluation approaches overstate gains because they ignore significant costs incurred by nudge recipients. Overall, home energy report welfare gains might be overstated by $620 million. We develop a prediction algorithm for optimal targeting; this approach would double the welfare gains. (*JEL* C93, D91, L95, Q41, Q48)

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 restricting 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]

*Allcott: New York University, 19 West 4th Street, New York, NY 10012, NBER, E2e, JPAL, and ideas42 (email: hunt.allcott@nyu.edu); Kessler: Wharton, 1454 Steinberg Hall-Dietrich Hall, 3620 Locust Walk, Philadelphia, PA 19104, NBER, and JPAL (email: 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 the 2017 ASSA Annual Meeting, Berkeley, Chicago Booth, the Consumer Financial Protection Bureau, Cornell, Duke, Harvard, Microsoft Research, New York University, Stanford, Stanford Institute for Theoretical Economics, the University of Hawaii, 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 randomized control trial (RCT) was registered in the American Economic Association Registry for RCTs under trial number 713. Code to replicate the analysis is available from Allcott's website.

[†] Go to https://doi.org/10.1257/app.20170328 to visit the article page for additional materials and author disclosure statement(s) or to comment in the online discussion forum.

[1] In September 2015, President Obama signed an executive order that directed federal agencies to use behavioral insights when they "may yield substantial improvements in social welfare and program outcomes" (Obama 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 11,000 times.

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 may experience two types of costs and benefits 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, suggesting that the growing support for nudge interventions could be misguided and possibly even socially harmful.

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. Opower, the leading HER provider, works with about 100 utility companies in nine countries, sending HERs regularly to about 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 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?

[3] Of course, if the policymaker has correctly designated a "good" behavior to nudge people toward, behavior change may generate 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) reviews 156 published field trials studying social comparisons and other informational interventions to induce energy conservation.

[5] These include Summit Blue Consulting (2009), Ashby et al. (2012), Integral Analytics (2012), KEMA (2012), Opinion Dynamics (2013), and Perry and Woehleke (2013), among many others.

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.

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 elicit WTP by asking recipients to 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 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 57 percent of retail energy cost savings, meaning that the remaining 43 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 43 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 $0.77 per household. However, the standard approach of ignoring non-energy costs overstates this welfare gain by a factor of 3.7. We find the same qualitative results in a more speculative calculation where we generalize the 43 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 two. Since HERs have been delivered to millions of households worldwide, this adds up quickly: across all HER programs implemented as of January 2017, the social welfare gains could be overstated by $620 million.

The nudge's welfare effects are driven down by the fact that almost 60 percent of nudge recipients are not willing to pay the marginal social cost of the nudge, including many recipients who have negative WTP. However, one in five recipients is willing to pay at least $9, which is 4.8 times larger than marginal social cost. A natural response to such heterogeneous valuations would be to price the nudge at marginal social cost and let people opt in. In this context, however, inertia is extremely powerful—most people do not bother to opt into a HER program that would generate only a few dollars of benefits.[6] We show that even under generous assumptions, an opt-in program is unlikely to enroll enough people to be preferable to the current opt-out approach. Instead, we train a simple machine learning algorithm to set a "smart default"—that is, to target the program at the types of consumers for whom nudging generates the largest social benefits. The smart default approach can double the welfare gains, holding constant the number of nudge recipients.

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 et al. (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.[7] 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 actually quantifying 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.[8] Avoidance designs achieve the same conceptual goal as our MPL: both

[6] HERs involve much lower stakes than other contexts with powerful default effects, 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] See Weimer, Vining, and Thomas (2009); FDA (2011); Chaloupka et al. (2014); Ashley, Nardinelli, and Lavaty (2015); Chaloupka, Gruber, and Warner (2015); Cutler et al. (2015); Jin et al. (2015); and others.

[8] Herberich, List, and Price (2011) uses 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)

allow the researcher 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.

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.

A. *Consumers and Producers*

We model a population of P 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. Let e generate 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 $f(e;\alpha)$.

In the model, e is produced at constant marginal cost c_e and sold at uniform price p_e, giving markup $\pi_e = p_e - c_e$ per unit. Here, 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).

A 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

and Trachtman et al. (2015) use a different avoidance design to study motivations for charitable giving, although none of these three papers includes a social welfare analysis.

consumption relative to a social norm s: if $M = m_s - \mu(e-s)$, this equals $m - \mu e$ when we set $m = m_s + \mu s$. Note that 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 quasi-linear 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. As we discuss below, however, a price increase and 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 set $\tilde{e}$ based on $\hat{f}(e;\alpha,\gamma)$ instead of $f(e;\alpha)$. Second, there is a pricing distortion: 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

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

B. *Nudges*

The policymaker can implement a nudge for the population of P consumers at marginal cost c_n per consumer plus fixed cost F_n, giving total cost $C_n = Pc_n + F_n$. The policymaker maintains a balanced budget using lump-sum tax $T = C_n/P$. We formalize the nudge as a binary instrument $N \in \{0,1\}$ that changes consumers' γ, m, and μ. Specifically, each consumer has possibly different potential outcomes θ_N for $N = 0$ versus $N = 1$, in which γ, m, and μ could differ. We define $\Theta = \{\theta_0, \theta_1\}$ and let $F(\Theta)$ denote its distribution. In other 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).

C. *Private and Social Welfare Effects of Nudges*

We define "pretax 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 pretax consumer welfare is

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

This equation shows that an energy conservation nudge, for example, affects consumers by saving them money, changing the amount of household "energy services" they enjoy, and by changing how they feel about their energy use.

Social welfare is consumer welfare plus the pricing distortion times the behavior change:

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

The effect of the nudge on social welfare is

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

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 final term C_n then accounts for the cost of the nudge.

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*, which is shaded with diagonal lines on Figure 1.

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 $-\Delta\tilde{e} \cdot p_e$ (rectangle *acdg*), 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.,

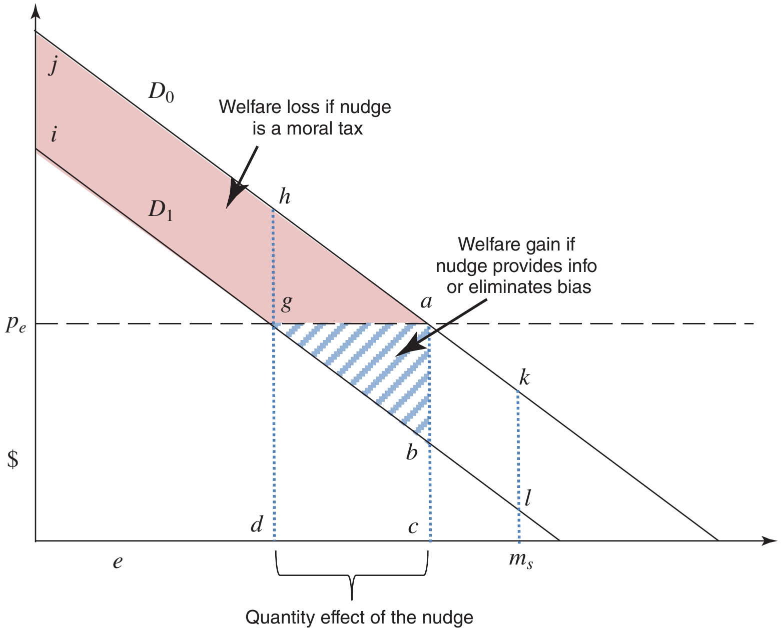

j
i
D_0
D_1
Welfare loss if nudge is a moral tax
h
Welfare gain if nudge provides info or eliminates bias
g
a
p_e
k
$
b
l
d
c
e
m_s
Quantity effect of the nudge

Figure 1. Illustrating the Effects of a Nudge on Consumer Welfare

Notes: This figure illustrates how two nudges with the same effect on quantity can have very different welfare effects. If a nudge is a "moral tax" on behavior, it reduces welfare by the solid shaded trapezoid *agij*. If a nudge provides information or eliminates bias, it increases welfare by the diagonally shaded triangle *abg*.

trapezoid *agij*, which has solid shading on Figure 1. 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. This generates the same financial savings $(acdg)$ and consumption utility loss $(acdh)$ as the moral tax case, but now the moral subsidy generates moral utility gain *ghkl*. In total, consumer welfare changes 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 generally, the nudge can have unbounded positive or negative effects on ΔV unless further restrictions are placed on m.

This discussion highlights how traditional cost-effectiveness 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.

D. *Estimation*

In the remainder of the paper, we estimate the social welfare effect ΔW of a specific nudge: Home Energy Reports. In this application, a retailer sells e with decreasing block pricing, and one could also imagine heterogeneous or nonlinear prices in other applications. To accommodate those cases, we can write the effect of the nudge on the retailer's net revenues as $\Delta\Pi$ instead of $\int \pi_e \Delta\tilde{e}\, dF(\Theta)$. This gives a more general version of equation (7):

$$(8) \qquad \Delta W = \int \Delta V - \phi_e \Delta\tilde{e}\ dF(\Theta) + \Delta\Pi - C_n.$$

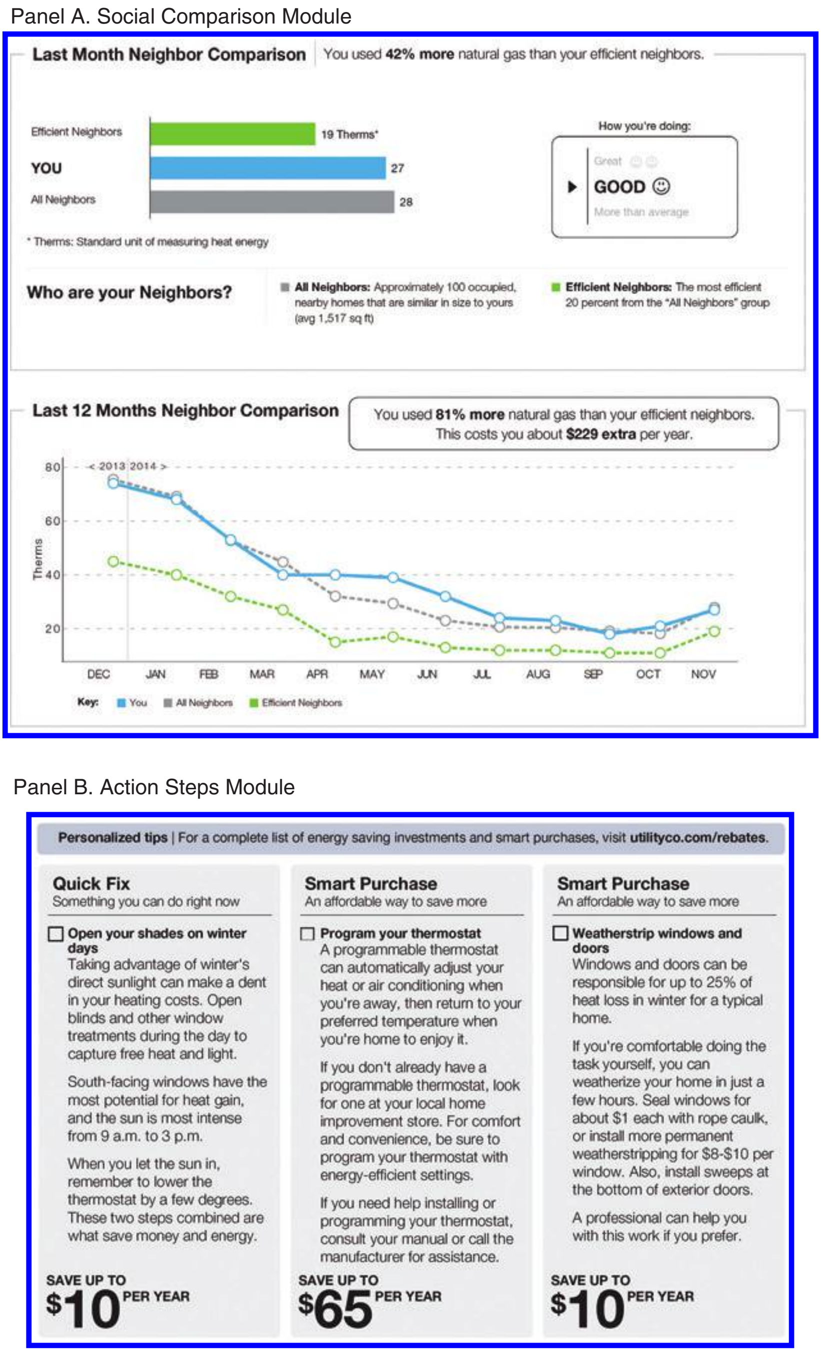
We estimate the change in energy use $\Delta\tilde{e}$ and retailer net revenues $\Delta\Pi$ by implementing HERs as a randomized control trial. We use outside estimates of energy use externalities ϕ_e, and we estimate nudge implementation cost C_n from accounting data. To estimate the change in consumer 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 twentieth 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."

We study one HER 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 that were not included in one of Central Hudson's several previous HER programs, Opower randomly assigned half to treatment and half to control. The treatment group received up to four HERs during the winter "heating season" from late October 2014 through late April 2015. Central Hudson employees read each household's natural gas meter every two months, and a HER was generated and mailed shortly after each meter read in order to provide timely and relevant



Panel A. Social Comparison Module

Last Month Neighbor Comparison | You used **42% more** natural gas than your efficient neighbors.

* Therms: Standard unit of measuring heat energy

Who are your Neighbors?

■ **All Neighbors:** Approximately 100 occupied, nearby homes that are similar in size to yours (avg 1,517 sq ft)

🟩 **Efficient Neighbors:** The most efficient 20 percent from the "All Neighbors" group

Last 12 Months Neighbor Comparison

You used **81% more** natural gas than your efficient neighbors.
This costs you about **$229 extra** per year.

Panel B. Action Steps 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.

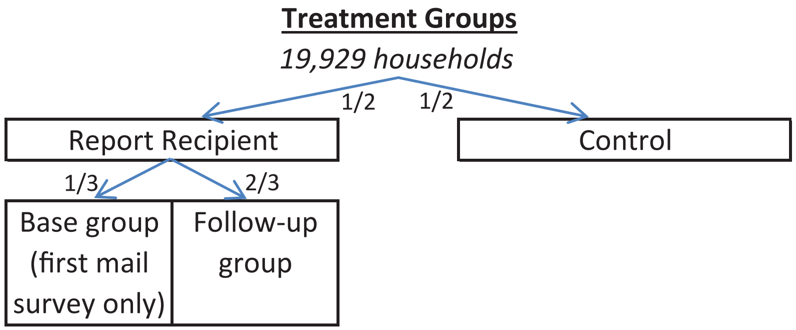
A professional can help you with this work if you prefer.

SAVE UP TO $10 PER YEAR

FIGURE 2. THE OPOWER HOME ENERGY REPORT

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.

Source: Opower



Process

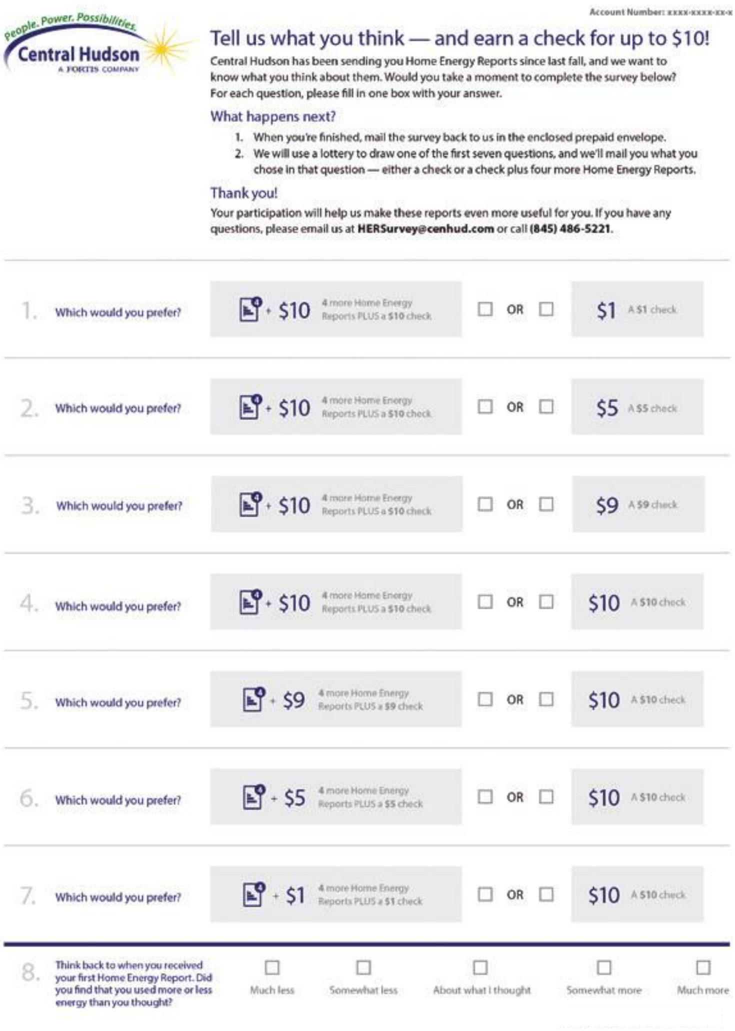
1. **Four reports** (October 2014-April 2015)
2. **First mail survey** (in final report)
3. **Follow-up mail survey** (own envelope, May 2015)
4. **Phone survey** (June-August 2015)
5. **Next four reports and/or check** (October 2015-April 2016)

Figure 3. Experimental Design

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.

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.

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



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

Figure 4. Mail Survey

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 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.

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.

In order to measure the direction and magnitude of nonresponse bias, a randomly-selected two-thirds of HER recipients were sent a follow-up mail survey on May 26, 2015. We call this group the "follow-up group," while the other one-third is the "base group." This follow-up survey was not part of a 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.

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.[9] 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. Online 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] 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.

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.

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.[10] 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 nonresponse 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.

Evaluating the Program's Second Year.—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 a 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) shows that continued HERs cause incremental conservation even after receiving 8 to 24 reports over 2 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

[10] Because Central Hudson needed to continue sending HERs to most households to satisfy regulatory requirements under the Energy Efficiency Portfolio Standard, we placed 98.8 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.

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.[11] 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 versus later HERs differently. This discussion highlights why it is both interesting and relevant to evaluate the program's second year.

III. Data

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

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. Central Hudson reads customers' natural gas meters on very regular bimonthly cycles: 94 percent of billing period durations are between 55 and 70 days.

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 twentieth percentile of the distribution of neighbors' use. We observe that mean and twentieth percentile for all HERs, including HERs that control group households would have received.

Table 1 presents customer 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 algorithms—which we introduce in the following sections—less effective.

The population is relatively wealthy, although these data may overestimate household income.[12] 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.

[11] 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.

[12] The mean household is in a census block group with median household income of $64,000 according to census data.

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	1,500
House value ($000s)	16,741	271	173	18	2,527
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 (ccf) 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.

Our household covariates, denoted **X** in the sections that follow, are these same variables, except that we take natural logs of income, net worth, house value, age, and house age.[13] Online 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.

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 1,690 households (17.9 percent) completed the phone survey; 2,312 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 nongovernment surveys, and indeed higher than we expected. We discuss our strategy for extrapolating to the population of nonrespondents in Section IV.

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

[13] 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 2—Survey Response Rates

	Response rate (percent)
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

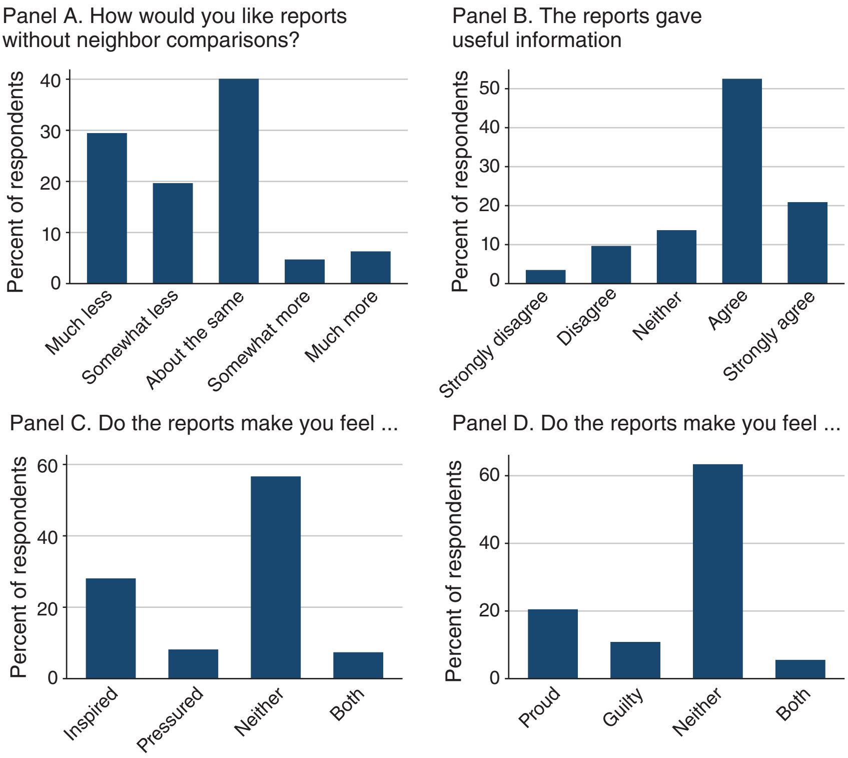


Figure 5. Qualitative Evaluations of Home Energy Reports

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

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) than negative (pressured or guilty). These qualitative results suggest that most people “like” HERs and want to receive them if they are free.

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 (WTP) 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.[14] 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 versus mail) and across different formats (i.e., mail versus phone). We therefore give them special attention in Section IIIB.

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. 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.[15]

[14] 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.

[15] We listened to about 25 of the 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.

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).[16] 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 the customer's expected savings, which suggests that consumers do not like guilt per se—they like guilt only because it helps them reduce expenditures. See online Appendix Table A4 for formal results.

As we detail below, 34 percent of respondents reported negative WTP. In online Appendix Table A5, we confirm that negative WTP is strongly associated with the same set of qualitative assessments in expected ways. Furthermore, all six households that both opted out of the HERs and responded to the survey had negative WTP. These strong correlations build confidence that both the MPL and the qualitative questions elicited meaningful responses.

Eighty-seven 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 online Appendix Table A6 for formal results.

Two-hundred-seventy-seven 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

[16] 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."

correlated, but not perfectly correlated. Online 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 four 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 6 percent and thought that checks would arrive six months before the HERs' benefits, their WTP would be about 3 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 our welfare calculations. 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 versus a one-time check due to a behavioral bias such as focusing bias (Köszegi and Szeidl 2013), it is again not clear what WTP to respect for welfare analysis.

Fourth, Beauchamp et al. (2016) demonstrates 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.[17]

[17] 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. (2016), 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 Estimate of Willingness-to-Pay

In this section, we calculate average WTP, which will be our measure of the consumer welfare effects in equation (8).

Before calculating WTP, it is important to clarify the target population and time period for which we want the estimate 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. Thus, $\mathcal{P}_n$ 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 online Appendix Table A10 to estimate $\Pr(H_i = 1 \mid \mathbf{X}_i; \mathcal{P})$ 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 $\left[\hat{\Pr}(H_i = 1 \mid \mathbf{X}_i)\right]^{-1}$. Of course, we are not able to correct for unobservable differences between sample and target populations, and we discuss this issue further below.

Figure 6 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, motivates the analysis of opt-in programs and targeting in Section VI.

Table 4 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 4 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 versus 2 suggests that survey responders may be slightly

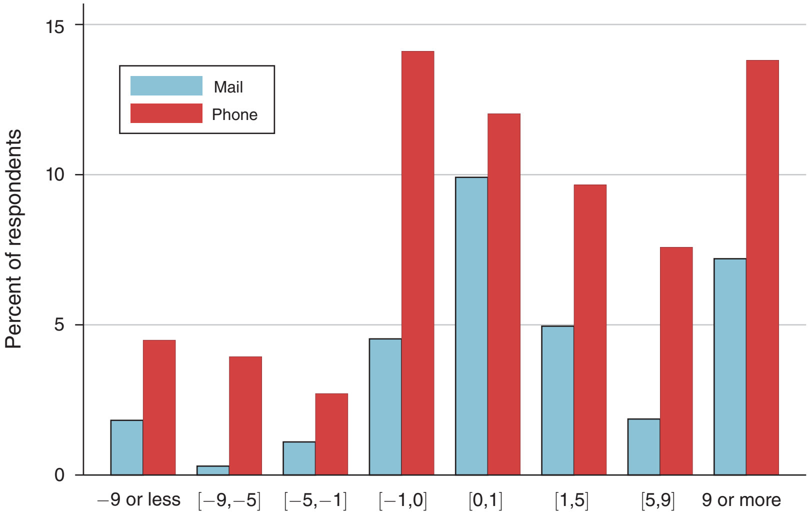


Figure 6. Willingness-to-Pay for Home Energy Reports

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

Table 4—Correlates of WTP and Their Correlation with Response

Dependent variable:	WTP (1)	Have WTP (2)
ln(Income)	0.0603 (0.244)	0.0295 (0.0225)
Retired	−1.588 (0.812)	0.182 (0.0751)
Married	0.683 (0.414)	−0.00765 (0.0368)
Rent	−0.780 (0.443)	−0.114 (0.0399)
Single family home	0.322 (0.424)	0.0629 (0.0382)
Buyer score	0.342 (0.219)	0.0500 (0.0199)
Observations	2,137	9,439

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 are in parentheses.

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 presents estimates of mean WTP, with standard errors in parentheses. Column 1 presents unweighted estimates, while column 2 uses row-specific IPWs to

Table 5—Estimates of Mean Willingness-to-Pay

	Unweighted (1)	Weighted (2)
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 versus follow-up mail	0.117	0.490
Returned first versus returned follow-up mail	0.001	0.024
Mail versus phone	0.059	0.081
Base group versus phone	0.026	0.160
Follow-up group versus phone	0.213	0.224
Returned first survey versus phone	0.000	0.004
Returned follow-up survey versus 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.

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 versus follow-up groups and households that actually returned the first survey versus 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 5 shows that respondents to the first mail survey are positively selected. Unweighted mean WTP is somewhat higher for the randomly-assigned base group versus follow-up group ($4.32 versus $3.22, $p \approx 0.117$), and mean WTP is much higher for households in either group that returned the first mail survey versus those that returned only the follow-up survey ($4.36 versus $2.58, $p \approx 0.001$). This positive selection is almost mechanical: the first mail survey was sent in the same envelope with a HER, and people who open and read HERs likely have higher WTP than people who do not.

By contrast, the phone survey and follow-up mail survey, which was sent in an envelope from a different outbound address and was not part of a HER, are not subject to this form of positive selection. Indeed, unweighted mean WTP is statistically and economically very similar for phone survey versus follow-up mail survey respondents (\$2.79 versus \$2.58, $p \approx 0.606$), and the weighted means are almost identical (\$2.67 versus \$2.63, $p \approx 0.925$). This similarity implies that these two samples are either not selected from nonrespondents or that they have the same sample selection bias despite coming from two different forms of contact (mail versus phone). Online Appendix Table A12 presents suggestive evidence in favor of the former explanation, showing that WTP does not vary statistically or economically for households that responded on earlier versus later phone survey attempts. Extending this logic suggests that phone survey nonresponders, 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.[18] 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 remain concerned about either positive or negative selection can also focus on the welfare evaluation for the subsample of survey respondents $\mathcal{P}_s$.

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. Note that this section is not necessary for our welfare calculations, and it requires additional speculative assumptions. While we believe exploring moral utility is valuable to better understand how and why nudges like HERs are effective, readers may choose to skip directly to Section V.

[18] 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 $\hat{\Pr}(H_i \,|\, \mathbf{X}_i; \mathcal{P}_n \backslash \mathcal{S}_1)$ 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 online Appendix Table A10. If w_i is WTP for household i and $N_n = 9{,}948$ is the number of households in target population $\mathcal{P}_n$, the predicted target mean WTP is $\left(\sum_{i \in \mathcal{S}_1} w_i + \sum_{i \in \{\mathcal{P}_n \backslash \mathcal{S}_1\}} \frac{w_i}{\hat{\Pr}(H_i \,|\, \mathbf{X}_i; \mathcal{P}_n \backslash \mathcal{S}_1)}\right) / N_n$.

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. Note that 39 percent of respondents to question 3 on the phone survey 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 nonzero savings. Moral utility, or some other unmodeled factor unrelated to financial gain or consumption utility, is needed to explain this nonzero 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 (or for *relatively* heavy users, in a formulation where moral utility depends on consumption relative to the social norm). 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.[19] 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 expected 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. See online Appendix Table A13 for formal results. Estimates show that expected 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. Using these estimates, we fit $\widehat{\Delta M_i} = \hat{\beta}_A \mathbf{A}_i$.

The first row of Table 6 presents results of univariate regressions of seven different variables on mean usage in winter 2014–2015, measured in ccf/day. In the

[19] 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 $d\Delta M/d\tilde{e}(\theta_1) = -\Delta\mu - \mu_0(d\Delta\tilde{e}/d\tilde{e}(\theta_1))$. We think of the moral price μ as being weakly positive. If $\mu_0 > 0$, then $d\Delta M/d\tilde{e}(\theta_1) < 0$ if $d\Delta\tilde{e}/d\tilde{e}(\theta_1) > \Delta\mu/\mu_0$, i.e., if behavior change does not increase too much in $\tilde{e}(\theta_1)$. If $\mu_0 = 0$, $d\Delta M/d\tilde{e}(\theta_1) < 0$ holds unambiguously.

Table 6—Measuring Moral Utility

	WTP (1)	Expected savings (2)	Inspired (3)	Pressured (4)	Proud (5)	Guilty (6)	$\widehat{\Delta M}$ (7)
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 1,000s 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 are in parentheses.

context of the model, this usage variable is $\tilde{e}(\theta_1)$. Column 1 reports that a one ccf/day increase in mean usage is unconditionally associated with a \$0.188 increase in WTP. Heavier users expect higher 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 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 6 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.

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 online Appendix Table A14 for formal results.

A fourth prediction is that if $\Delta\mu > 0$ but $\Delta m = 0$, then $\Delta M < 0$. In other 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 heavier users are less likely to say that the HERs make them feel good, and more likely to say that the HERs make them feel bad. In the context of our model, this means that HERs do seem to increase the "moral price" of energy use. This would seem to be consistent with Glaeser's (2006) concern. However, more people report positive affect than negative affect, and overall average WTP is positive. In the context of our model, this means that HERs are not only a moral tax. Instead, they also provide information and a positive affect windfall $\Delta m > 0$.

V. Welfare

In this section, we use the empirical estimate of WTP to calibrate the welfare formula from equation (8). Before doing this, we make explicit our key revealed preference assumption and quantify the other parameters necessary for the welfare calculation.

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. However, 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.[20]

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

[20] Online Appendix E.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 relative to the true ΔV_i, thus causing our calculation to overstate welfare gains. There is also suggestive evidence that consumers are overconfident, 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.

B. *Energy Use*, *Retailer Net Revenue*, *Externality*, *and Implementation Cost Parameters*

To complete the welfare analysis prescribed by equation (8), we need to estimate treatment effects of the second year of HERs on energy use $\Delta\tilde{e}$ and on Central Hudson's net revenues $\Delta\Pi$, calculate the externality ϕ_e, and measure implementation cost C_n.

Conceptually, the ideal way to estimate treatment effects of the second year of HERs on energy use and retailer net revenues would be to compare 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 addition, such estimates would not have been very precise at this sample size. Instead, we maintain a "no persistence" assumption, i.e.,—we assume that HERs sent during a given winter only affect energy use in that same winter—and estimate the causal effect of the program during its second winter.[21] As we will explain below, alternative assumptions would have very little effect on our welfare analysis because natural gas prices are so close to social marginal cost.

Because estimating these treatment effects is straightforward, we relegate that exercise to online Appendix C. We estimate that receiving the second year of Home Energy Reports changes natural gas use by $\Delta\tilde{e} \approx -6.59$ ccf for $\mathcal{P}_n$, the target population of treatment group households that did not opt out before the program's second year, and by $\Delta\tilde{e} \approx -7.65$ ccf for $\mathcal{P}_s$, the target population of MPL respondents with valid WTP that did not opt out. The slightly larger energy savings estimate for $\mathcal{P}_s$ suggests that survey respondents may have been slightly more engaged with the HERs compared to nonrespondents.

Central Hudson uses a decreasing block price schedule, with prices above marginal cost in order to help recover fixed costs. Because prices are nonlinear, we cannot simply multiply the treatment effects on natural gas by some constant markup to get the treatment effects on retailer net revenues $\Delta\Pi$. Instead, we estimate additional treatment effect regressions where the dependent variable is Central Hudson's net revenues accrued on each bimonthly natural gas bill. In online Appendix C, we find that the second year of HERs reduced Central Hudson's net revenues $\Delta\Pi$ by \$2.53 and \$2.84 for the average household in populations $\mathcal{P}_n$ and $\mathcal{P}_s$, respectively. Using similar regressions, we also estimate that HERs reduce consumers' average retail natural gas expenditures by \$4.91 and \$5.61 for $\mathcal{P}_n$ and $\mathcal{P}_s$, respectively.

The average marginal markup—that is, the average markup on the natural gas that households conserved as a result of receiving HERs—equals the retailer net

[21] To our knowledge, there is no published evidence on whether HERs have persistent effects on natural gas use. Allcott and Rogers (2014) finds that the effects of the first four HERs on electricity use quickly decay away, which is consistent with the "no persistence" assumption. Allcott and Rogers (2014) also finds persistent effects of HERs on electricity use after discontinuing treatment, but this is in samples where most households had received 24 HERs over two years, which is much more extensive than the first year or two of the Central Hudson program. More recent work by Brandon et al. (2017, table 4) suggests that about one-quarter of the effects of the first year of HER programs on electricity usage are due to persistent capital stock changes, as are about one-third of the effects of the first two years. If this is also true for natural gas, it suggests a small violation of the no persistence assumption.

revenue loss per unit of energy conserved. For $\mathcal{P}_n$, this is $\hat{\pi}_e \equiv \Delta\Pi/\Delta\tilde{e} \approx -\$2.53/-6.59 \approx \$0.38$ per ccf, and the ratio is similar for $\mathcal{P}_s$.

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 US 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 3 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 (2014), giving an additional \$0.10/ccf externality. Thus, the total environmental externality ϕ_e is $\$0.045 + \$0.264 + \$0.10 \approx \$0.41$ per ccf.

As the welfare formulas from Section I show, the social welfare effects of energy conservation that the consumer does not internalize are the retailer net revenue and externality effects: $\Delta\Pi - \phi_e\Delta\tilde{e}$. Using that $\hat{\pi}_e \equiv \Delta\Pi/\Delta\tilde{e}$, this can also be written as the average pricing distortion (which is the average marginal markup minus the environmental externality) times the amount of conservation: $(\hat{\pi}_e - \phi_e)\Delta\tilde{e}$. Intuitively, we can think of energy conservation as reducing environmental externalities by amount $\phi_e\Delta\tilde{e}$, but also imposing a pecuniary externality $\hat{\pi}_e\Delta\tilde{e}$ on other consumers, who now must contribute more to the retailer's fixed cost recovery.[22] This insight is not new (see Davis and Muehlegger 2010), and Central Hudson's pricing structure is not unusual: Central Hudson's retail markup is closely comparable to the 40 percent average markup for residential and commercial natural gas consumers nationwide, as calculated by Davis and Muehlegger (2010).[23]

For many readers, the intuitive case is that $\hat{\pi}_e - \phi_e < 0$, so the average marginal energy price is below social marginal cost, and energy conservation benefits others in society. However, if $\hat{\pi}_e - \phi_e > 0$, then the energy price is actually *above* social

[22] 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. In our model and welfare estimates, we directly count retailer net revenues. Given that profit is in fact held constant by varying future retail prices, and this price variation in turn generates deadweight loss, the true welfare effects of HERs are lower (higher) than our estimates if future marginal retail prices are higher (lower) than marginal social cost.

[23] 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 and set constant marginal prices. There are various justifications for amortizing fixed costs into marginal prices using either linear pricing or decreasing block pricing, 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, retailer net revenue effects $\Delta\Pi$ still enter the welfare calculation.

marginal cost, and energy conservation imposes a net burden on others in society. If $\hat{\pi}_e - \phi_e = 0$, then the energy price is not distorted on average. In the latter case, equation (8) shows that the nudge's welfare effect is just the private benefits net of the nudge implementation cost.

Using the estimates above, $\hat{\pi}_e - \phi_e \approx \$0.38 - \$0.41 \approx -\$0.03/\text{ccf}$, so the average marginal retail gas price happens to be very close to social marginal cost. As Davis and Muehlegger (2010) points out, although this is sensitive to the social cost of carbon and other externality damage parameters, it significantly diminishes 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 primarily to the extent that they address market failures such as imperfect information or otherwise increase consumer welfare. Welfare gains from the nudge will need to be driven primarily by private gains to nudge recipients rather than by uninternalized social benefits.

Because natural gas is priced so close to social marginal cost, the welfare estimates will not be very sensitive to the energy savings estimates, and thus not very sensitive to violations of the "no persistence" assumption. As long as the average marginal markup $\hat{\pi}_e$ approximately equals the externality ϕ_e, any changes in $\Delta\tilde{e}$ will affect the externality reduction $\phi_e \Delta\tilde{e}$ and retailer net revenues $\hat{\pi}_e \Delta\tilde{e}$ by approximately offsetting amounts.

Finally, in online Appendix E.B, we use an accounting analysis to calculate nudge implementation cost C_n. 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.[24] 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 $F_n/P \approx \$0.16$/household. We also present an alternative calculation in which these costs are allocated equally to each of the four programs, giving $F_n \approx \$4{,}085$ per program. The number of households to be nudged in our program's second year (i.e., $\mathcal{P}_n$, the set of recipient households that did not opt out) is $P = 9{,}948$. This would give $F_n/P \approx \$0.41$ per household in the program we study.

C. *Results*

Table 7 presents the welfare analysis of the program's second year. Columns 1 and 2 present results after re-weighting 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 online Appendix Table A7 and WTP from column 2 of Table 5. Columns 3 and 4 present results for

[24] Different programs are well-defined in the sense that they have different specific customer subpopulations 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.

Table 7—Social Welfare Effects of a Second Year of Home Energy Reports

Target population:	All HER recipients		MPL respondents	
Panel A. Benefits and costs other than consumer welfare ($/*recipient*)				
Externality reduction	2.71		3.15	
(−) Retailer net revenue loss	2.53		2.84	
(−) Implementation cost	2.22		2.22	
(=) ΔWelfare, excluding consumer welfare	−2.04		−1.92	
Panel B. Mean WTP and social welfare effect ($/*recipient*)				
Assumption	Mean WTP (1)	ΔWelfare (2)	Mean WTP (3)	ΔWelfare (4)
Base case	2.81	0.77	2.97	1.06
Uniform WTP at MPL endpoints	2.59	0.55	2.74	0.82
WTP = {−12, 12} at MPL endpoints	2.34	0.30	2.48	0.56
WTP = {−15, 15} at MPL endpoints	2.76	0.72	2.92	1.01
WTP bounds closest to zero	1.61	−0.43	1.70	−0.21
Nonrespondents have WTP = 0	0.60	−1.44		
Weight = 1 for first mail respondents	2.68	0.64		
Fixed costs equally allocated	2.81	0.53	2.97	0.81
Gas savings 100% larger	2.81	0.95	2.97	1.36

Notes: Columns 1 and 2 present results after re-weighting 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 online Appendix Table A7, retailer net revenue loss from column 3 of online Appendix Table A9, and WTP from column 2 of Table 5. 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 online Appendix Table A7, retailer net revenue loss from column 4 of Table A9, and WTP from column 1 of Table 5. Using equation (8), ΔWelfare in columns 2 and 4 is consumer welfare gain + externality reduction − retailer net revenue loss − implementation cost, where Mean WTP in columns 1 and 3 is our measure of the consumer welfare gain.

$\mathcal{P}_s$, the target population of MPL respondents with valid WTP that did not opt out. These columns use energy savings from column 4 of online Appendix Table A7 and WTP from column 1 of Table 5. 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.

Panel A of Table 7 presents benefits and costs other than consumer welfare. For target population $\mathcal{P}_n$, the conservation induced by HERs reduces environmental externalities by \$2.71 per household and decreases Central Hudson's net revenues by \$2.53 per household. The social welfare effect excluding consumer welfare is externality reduction minus retailer net revenue loss minus implementation cost, or $2.71 - 2.53 - 2.22 \approx -\$2.04$ for target population $\mathcal{P}_n$.

Panel B of Table 7 completes the social welfare estimates by adding in WTP as the measure of consumer welfare gain. Columns 1 and 3 present WTP, while columns 2 and 4 present the resulting social welfare estimate using equation (8). The first row presents 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 5. The social welfare effects are \$0.77 and \$1.06 per household for the full population and for MPL respondents, respectively.

The multiple price list survey allows us to bound each respondent's WTP, but we made particular assumptions to go from bounds to point estimates. The next four rows of panel B consider sensitivity to alternative assumptions. The second, third, and fourth 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 second row 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 fifth row of panel B uses the bounds of each interval closest to zero—for example, all consumers with WTP between $-\$5$ and $-\$1$ are assigned WTP $= -\$1$.

The next two rows of panel B consider alternative adjustments for survey nonresponse when extrapolating WTP to the full HER recipient population. In Section IV, we speculated that if there is a nonresponse bias, it is likely positive. Under the extreme assumption that nonrespondents have zero WTP, mean WTP would be \$0.60, and welfare effects would be $-\$1.44$ per HER recipient. We view this as an unrealistically conservative assumption.[25] 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 \$0.64 per recipient.

The next 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 a HER program would not be large enough to generate enough social surplus to outweigh the program-level fixed costs. If implementation costs were 35 percent higher, externality damages were 29 percent lower, or WTP were 28 percent lower, the base case social welfare point estimate would be negative.

Because $(\hat{\pi}_e - \phi_e) \approx 0$, i.e., average 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, sampling error in our energy savings estimates and the "no persistence" assumption described earlier make little difference in our welfare estimates. The final row in panel B of Table 7 presents an alternative scenario where we double both the externality reduction and retailer net revenue loss assumptions. This affects the welfare estimates for the two target populations by \$0.18 and \$0.30.

We can also be more precise about the impact of sampling error. Applying the Delta method to the energy savings estimates in online Appendix Table A7, the 95 percent confidence interval on welfare effects for target $\mathcal{P}_n$ extends $1.96 \cdot \hat{SE}(\tau) \cdot 243 \cdot (\phi_e - \hat{\pi}_e) \approx \$0.16$ in either direction, where τ refers to the winter 2015–2016 treatment effect (in ccf/day) and 243 is the number of days in winter 2015–2016. WTP estimates in Table 5 are relatively precisely estimated, with a 95 percent confidence interval that extends \$0.31 in either direction for both the weighted and unweighted estimates.

[25] 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 5 percent of phone survey respondents reported not remembering HERs.

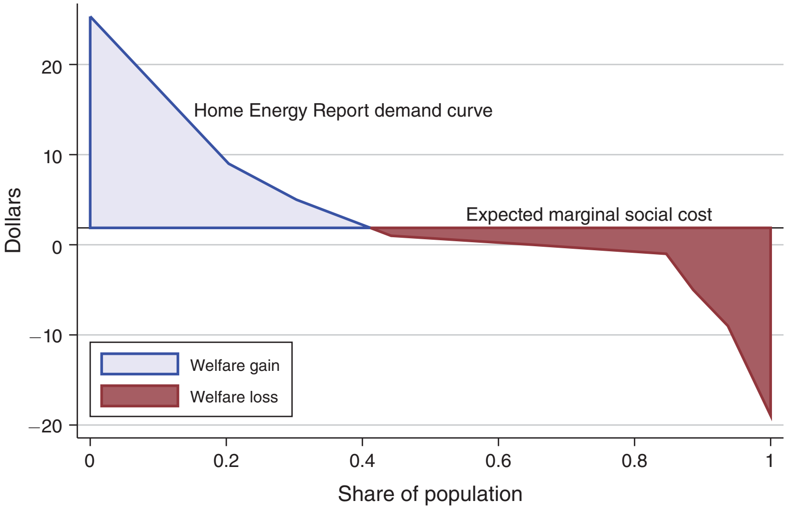


Figure 7. Social Welfare Analysis: Graphical

Note: 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 7.

Figure 7 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" of the nudge, i.e., the marginal implementation cost net of the average uninternalized social benefit $c_n - 243\hat{\tau}(\hat{\pi}_e - \phi_e)$, is approximately \$1.88 per household. The net social welfare effect is the area between the demand curve and expected marginal social cost, i.e., the lightly shaded area minus the darkly shaded area, minus fixed cost F_n. Leaving aside the variation in $\Delta\tilde{e}$ across households, which is less relevant because $(\hat{\pi}_e - \phi_e) \approx 0$, the social welfare effect trades off the gains to the 41 percent of consumers willing to pay more than \$1.88 with the losses to the 59 percent of consumers that are not. The consumer surplus in the lightly shaded area is large, and 20 percent of people are willing to pay at least \$9, which is 4.8 times larger than expected marginal social 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.

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 expenditure savings (\$4.91) implies that consumers in population $\mathcal{P}_n$ incur an average of \$2.10 in net utility costs, which we call "non-energy costs" for shorthand.[26] This benefit/cost ratio of \$4.91/\$2.10 ≈ 2.34 implies that

[26] 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

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 of the program on consumer welfare is to save people money on buying cigarettes.

How does ignoring non-energy costs affect the social welfare calculation in Table 7? If we set $\Delta V \approx \$4.91$, the welfare gain is \$2.87 per recipient—3.7 times larger than our base case estimate of \$0.77 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 online Appendix E.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 $\$2.10/\$4.91 \approx 43$ percent 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 two.

Given that Home Energy Reports have been sent to millions of households around the world, these program evaluation discrepancies add up very quickly. Online Appendix E.C extends our speculative calculation to show that when we aggregate over all the energy saved through HERs as of January 2017, the standard program evaluation approach (ignoring non-energy costs) suggests social welfare gains of \$1.22 billion. Accounting for our estimate of non-energy costs decreases that estimate to \$600 million. Thus, we estimate that failing to account for non-energy costs causes the social value of these nudges to be overstated by \$620 million.

VI. Allocating Nudges: Opt In versus Smart Defaults

Figure 6 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? We consider two approaches: an opt-in program and a machine learning algorithm that targets the nudge to maximize social welfare.

of future retail energy cost savings from the phone survey. Mean expected savings is larger than the observed \$4.91, so this would imply larger non-energy costs, which further reinforces our arguments in this section.

Table 8—Opt In and Smart Defaults: Results

Row	Policy	Percent of population receiving HERs (1)	Mean gas use change (ccf/ recipient-day) (2)	Mean WTP ($/ recipient) (3)	Welfare effect ($/ recipient) (4)	Total welfare effect ($000s) (5)
1	Existing opt-out program	50	−0.027	2.81	0.77	7.7
2	Opt in; zero switching cost	41	−0.027	9.78	7.42	60.7
3	Opt in; 1.5% opt-in rate	1.5	−0.027	24.5	16.93	5.1
4	Targeted on energy savings	50	−0.050	3.08	1.19	11.9
5	Targeted on WTP	50	−0.040	3.30	1.34	13.4
6	Targeted on welfare	50	−0.048	3.42	1.51	15.1

Note: Column 5 presents the aggregate social welfare effect across all 19,929 households, which is (column 1)/100 × (column 4) × (19,929/1,000).

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 second-year average treatment effect $\hat{\tau}$ from online Appendix Table A7. We then set the price at expected social marginal cost, i.e., $c_n - 243\hat{\tau}\left(\hat{\pi}_e - \phi_e\right) \approx \$1.88$.

Table 8 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 × (column 4) × (19,929/1,000). Row 1 presents the existing opt-out program as a benchmark.

Row 2 presents the welfare effects of an opt-in program assuming zero inertia—that is, we assume that all consumers opt into the second year of HERs if and only if they are willing to pay more than the $1.88 price. Under this assumption, 41 percent of consumers opt in, and they have mean WTP of $9.78. The total social welfare gain in column 5 is eight 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 United States, at a large utility in Ohio called American Electric Power. 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, even though there are many people who value the nudge at more than its price, switching costs or other forms of inertia prevent most of them 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 decisions such as choosing health insurance or retirement savings plans, 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 inertia assumption in row 2 is unrealistic.

We explore the importance of inertia under three assumptions. First, we assume only 1.5 percent of consumers opt in, as in Ohio. Second, we assume consumers opt in if and only if their WTP is larger than a switching cost, so the 1.5 percent that opt in will be drawn from the right tail of the WTP distribution. Third, we assume 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—one-third less 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.

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 $\delta : \mathcal{X} \rightarrow \{0, 1\}$ that maps household covariates from space $\mathcal{X}$ to treatment assignment $\{0, 1\}$ in order to maximize objective $L(\delta)$. 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 $L_\tau(\delta_\tau) = \sum_i -\tau_i \delta_\tau(\mathbf{X}_i)$, consumer welfare (i.e., WTP), where $L_{CW}(\delta_{CW}) = \sum_i w_i \delta_{CW}(\mathbf{X}_i)$, and social welfare, where $L_W(\delta_W) = -F_n + \sum_i \big(w_i + (\tilde{\pi}_e - \phi_e)\tau_i - c_n\big)\,\delta_W(\mathbf{X}_i)$.

This is a standard prediction problem, in which additional flexibility in the functional form of $\delta(\mathbf{X}_i)$ allows more precise in-sample fit but worsens out-of-sample performance. Intuitively, if we predict WTP w using sufficiently flexible functions of $\mathbf{X}$, we can appear to perfectly predict WTP, allowing a targeting algorithm that appears to perfectly allocate the nudge to high-WTP consumers in one sample but would perform poorly in another sample. To avoid such overfitting, we use cross-validation. Specifically, we take the following steps.

First, we randomly partition the sample of 19,929 households into five subsamples of equal size. For each subsample (the test set), we fit machine learning algorithms that predict $\hat{w}\,|\,\mathbf{X}_i$ and $\hat{\tau}\,|\,\mathbf{X}_i$ using the other four subsamples (the training set), and project $\hat{w}\,|\,\mathbf{X}_i$ and $\hat{\tau}\,|\,\mathbf{X}_i$ back into the test set. This gives out-of-sample predictions of $\hat{w}\,|\,\mathbf{X}_i$ and $\hat{\tau}\,|\,\mathbf{X}_i$ for all observations.

In that first step, we use several different algorithms to predict $\hat{w}\,|\,\mathbf{X}_i$ and $\hat{\tau}\,|\,\mathbf{X}_i$, choosing the ones that deliver the highest values of objectives L_{CW} and L_τ, respectively, as evaluated in the third step described below. To predict $\hat{w}\,|\,\mathbf{X}_i$, we use elastic nets (Zou and Hastie 2005), as this delivers slightly higher L_{CW} than ridge regression and random forests. To predict $\hat{\tau}\,|\,\mathbf{X}_i$, we use gradient forests (Athey, Tibshirani, and Wager 2017), as this delivers significantly higher L_τ than the approach of Imai and Strauss (2013). We tuned the algorithms separately within each training set. Online

Appendix Table A18 presents performance statistics and optimal tuning parameters for the machine learning algorithms that we tested.

Second, we set δ_{τ}, δ_{CW}, and δ_W equal to 1 for the halves of the population with above-median values of predicted savings $-\hat{\tau}|\mathbf{X}_i$, WTP $\hat{w}|\mathbf{X}_i$, and welfare gain $\hat{w}|\mathbf{X}_i + (\hat{\pi}_e - \phi_e)\hat{\tau}|\mathbf{X}_i$, respectively. These three sets of households maximize the predicted values of the three objectives L_{τ}, L_{CW}, and L_W, respectively, subject to the constraint that half of households are treated.

Third, we evaluate the performance of the targeting algorithms by estimating average WTP and the winter 2015–2016 treatment effect $\hat{\tau}$ using only the subsamples with δ_T, δ_{CW}, and δ_W equal to 1.

Rows 4–6 of Table 8 present results when maximizing energy savings, WTP, and welfare, respectively. Comparing rows 1 versus 6 of column 4 shows that targeting on welfare can approximately double the program's total welfare gains, moving from \$0.77 to \$1.51 per recipient. Interestingly, there is limited trade-off when targeting on the different objectives: maximizing energy savings also increases average WTP, and vice versa.

To help interpret the results of the prediction algorithms, Figure 8 presents differences in mean **X** variables (normalized into standard deviations) between targeted and nontargeted 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. This explains why there is limited trade-off 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.

A 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 IVA, 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—policies would ideally be written to explicitly encourage targeting to maximize welfare.

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

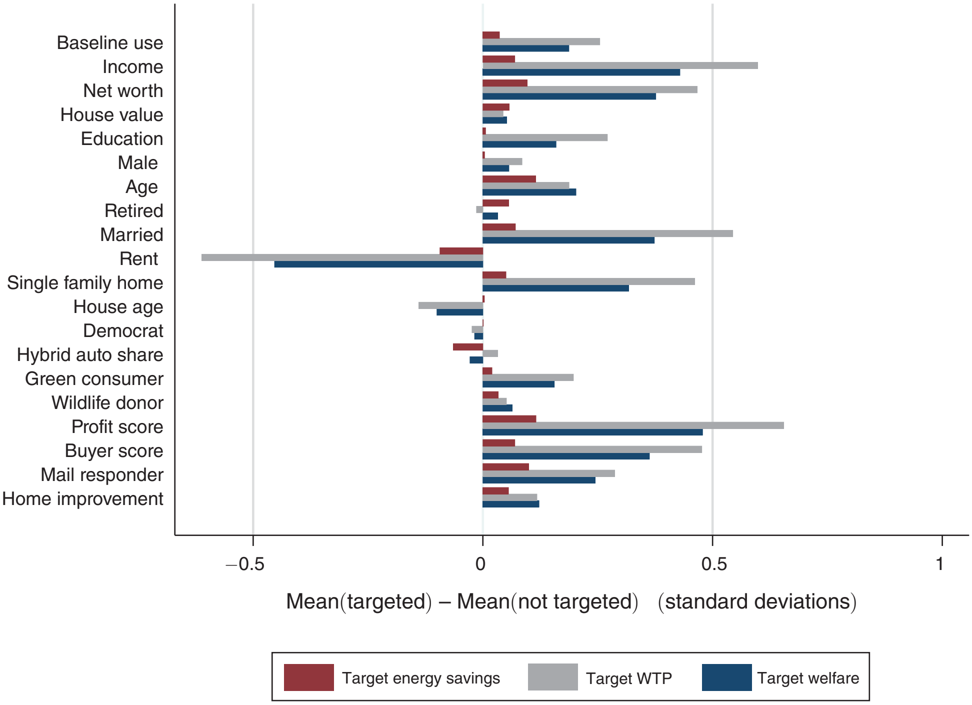


Figure 8. Demographic Differences between Targeted and Nontargeted Households

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 nontargeted households for each of these three maximands, in standard deviation units.

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 3.7. A speculative extrapolation of our results suggests that the overall social value of HERs may be overstated by $620 million. These results highlight the importance of measuring the full welfare effects of nudges.

REFERENCES

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 (5): 3237–45.

Allcott, Hunt. 2011. “Social norms and energy conservation.” *Journal of Public Economics* 95 (9–10): 1082–95.

Allcott, Hunt. 2015. “Site Selection Bias in Program Evaluation.” *Quarterly Journal of Economics* 130 (3): 1117–65.

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.

Ambuehl, Sandro, B. Douglas Bernheim, and Annamaria Lusardi. 2014. “A Method for Evaluating the Quality of Financial Decision Making, with an Application to Financial Education.” National Bureau of Economic Research (NBER) Working Paper 20618.

American Council for an Energy-Efficient Economy (ACEEE). 2015. *State Energy Efficiency Resource Standards* (*EERS*)*: April 2015*. American Council for an Energy-Efficient Economy. Washington, DC, April.

Andreoni, James, Justin M. Rao, and Hannah Trachtman. 2011. “Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving.” National Bureau of Economic Research (NBER) Working Paper 17648.

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://aceee.org/files/proceedings/2012/data/papers/0193-000218.pdf.

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): 617–24.

Athey, Susan, Julie Tibshirani, and Stefan Wager. 2017. “Solving Heterogeneous Estimating Equations with Gradient Forests.” https://arxiv.org/pdf/1610.01271v2.pdf.

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.

Beauchamp, Jonathan P., Daniel J. Benjamin, Christopher F. Chabris, and David I. Laibson. 2016. “Controlling for Compromise Effects Debiases Estimates of Preference Parameters.” https://scholar.harvard.edu/files/jonathanpbeauchamp/files/controlling_for_compromise_effects_paper.pdf.

Belloni, Alexandre, and Victor Chernozhukov. 2013. “Least squares after model selection in high-dimensional sparse models.” *Bernoulli* 19 (2): 521–47.

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.

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.

Bronnenberg, Bart J., Jean-Pierre Dubé, Matthew Gentzkow, and Jesse M. Shapiro. 2015. “Do Pharmacists Buy Bayer? Sophisticated Shoppers and the Brand Premium.” *Quarterly Journal of Economics* 130 (4): 1669–1726.

Caplin, Andrew. 2003. “Fear as a Policy Instrument.” In *Time and Decision: Economic and Psychological Perspectives of Intertemporal Choice,* edited by George Loewenstein, Daniel Read, and Roy Baumeister, 441–58. New York: Russell Sage Foundation.

Carroll, Gabriel D., James J. Choi, David Laibson, Brigitte Madrian, and Andrew Metrick. 2009. “Optimal Defaults and Active Decisions.” *Quarterly Journal of Economics* 124 (4): 1639–74.

Chaloupka, Frank J., Jonathan Gruber, and Kenneth E. Warner. 2015. “Accounting for ‘Lost Pleasure’ in a Cost–Benefit Analysis of Government Regulation: The Case of the Food and Drug Administration’s Proposed Cigarette Labeling Regulation.” *Annals of Internal Medicine* 162 (1): 64–65.

Chaloupka, Frank J., Kenneth E. Warner, Daron Acemoğlu, Jonathan Gruber, Fritz Laux, Wendy Max, Joseph Newhouse, et al. 2014. “An evaluation of the FDA’s analysis of the costs and benefits of the graphic warning label regulation.” *Tobacco Control*: 1–8.

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.

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): 247–80.

Davis, Lucas W., 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.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. 2012. “Testing for Altruism and Social Pressure in Charitable Giving.” *Quarterly Journal of Economics* 127 (1): 1–56.

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: 729–39.

DiNardo, John, Justin McCrary, and Lisa Sanbonmatsu. 2006. “Constructive Proposals for Dealing with Attrition: An Empirical Example.” https://pdfs.semanticscholar.org/b431/c3e7112e443eeab310b90016030ed53830f6.pdf.

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 (CEP) Discussion Paper 1222.

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.

Farhi, Emmanuel, and Xavier Gabaix. 2015. “Optimal Taxation with Behavioral Agents.” http://faculty.chicagobooth.edu/workshops/micro/pdf/Gabaix.pdf.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. 2015. “Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program.” National Bureau of Economic Research (NBER) Working Paper 21331.

Glaeser, Edward L. 2006. “Paternalism and Psychology.” *University of Chicago Law Review* 73 (1): 133–56.

Glaeser, Edward L. 2014. “The Supply of Environmentalism: Psychological Interventions and Economics.” *Review of Environmental Economics and Policy* 8 (2): 208–29.

Handel, Benjamin R. 2013. “Adverse Selection and Inertia in Health Insurance Markets: When Nudging Hurts.” *American Economic Review* 103 (7): 2643–82.

Herberich, David H., John A. List, and Michael K. Price. 2011. “How Many Economists Does It Take to Change a Light Bulb? A Natural Field Experiment on Technology Adoption.” https://www.researchgate.net/publication/228431296_How_Many_Economists_does_it_take_to_Change_a_Light_Bulb_A_Natural_Field_Experiment_on_Technology_Adoption.

Holland, Stephen P., Erin T. Mansur, Nicholas Z. Muller, and Andrew J. Yates. 2015. “Environmental Benefits from Driving Electric Vehicles?” National Bureau of Economic Research (NBER) Working Paper 21291.

Howarth, Robert, Drew Shindell, Renee Santoro, Anthony Ingraffea, Nathan Phillips, and Amy Townsend-Small. 2012. *Methane Emissions from Natural Gas Systems.* National Climate Assessment, U.S. Global Change Research Program. Washington, DC, February.

Integral Analytics. 2012. *Sacramento Municipal Utility District Home Energy Report Program: Program Years 2008–2011*. Integral Analytics. Cincinnati, November.

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*. Interagency Working Group on the Social Cost of Carbon, United States Government. Washington, DC, May.

Intergovernmental Panel on Climate Change. 2014. “Anthropogenic and Natural Radiative Forcing.” In *Climate Change 2013—The Physical Science Bias*, 659–740. Cambridge, UK: Cambridge University Press.

Ito, Koichiro. 2014. “Do Consumers Respond to Marginal or Average Price? Evidence from Nonlinear Electricity Pricing.” *American Economic Review* 104 (2): 537–63.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka. 2015. “The Persistence of Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand.” National Bureau of Economic Research (NBER) Working Paper 20910.

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): 154–86.

Kamenica, Emir. 2008. “Contextual Inference in Markets: On the Informational Content of Product Lines.” *American Economic Review* 98 (5): 2127–49.

Kantola, S. J., G. J. Syme, and N. A. Campbell. 1984. “Cognitive Dissonance and Energy Conservation.” *Journal of Applied Psychology* 69 (3): 416–21.

Keuring van Elektrotechnische Materialen te Arnhem (KEMA). 2012. *Puget Sound Energy’s Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation.* DNV KEMA Energy and Sustainability. Madison, WI, April.

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.

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

Larrick, Richard P., and Jack B. Soll. 2008. "The MPG Illusion." *Science* 320 (5883): 1593–94.

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): 1149–87.

Nolan, Jessica M., P. Wesley Schultz, Robert B. Cialdini, Noah J. Goldstein, and Vladas Griskevicius. 2008. "Normative Social Influence is Underdetected." *Personality and Social Psychology Bulletin* 34 (7): 913–23.

Obama, Barack. 2015. "Executive Order—Using Behavioral Science Insights to Better Serve the American People." Media Release, White House, Washington, DC, September 15, 2015.

Opinion Dynamics. 2013. *Massachusetts Cross-Cutting Behavioral Program Evaluation Integrated Report*. Opinion Dynamics. Waltham, June.

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

Pope, C. Arden, III, 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): 1132–41.

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): 429–34.

Sudarshan, Anant. 2014. "Nudges in the marketplace: Using peer comparisons and incentives to reduce household electricity consumption." http://www.anantsudarshan.com/uploads/1/0/2/6/10267789/nudges_sudarshan_2014.pdf.

Thaler, Richard H., and Cass R. Sunstein. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*. New Haven: Yale University Press.

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.

Summit Blue Consulting. 2009. *Impact Evaluation of Positive Energy SMUD Pilot Study*. Summit Blue Consulting. Boulder, May.

US Environmental Protection Agency (EPA). 1995. "External Combustion Sources." In *AP–42: Compilation of Air Emissions Factors*, Vol. 1, 5th ed. Washington, DC: US Environmental Protection Agency (EPA).

US Food and Drug Administration (FDA). 2011. "Required Warnings for Cigarette Packages and Advertisements, Final Rule." *Federal Register* 76 (120): 36628–36777.

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 (6): 1063–70.

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.

Whitehead, Mark, Rhys Jones, Rachel Howell, Rachel Lilley, and Jessica Pykett. 2014. *Assessing the Global Impact of the Behavioral Sciences on Public Policy: Nudging All over the World.* Economic and Social Research Council. Swindon, September.

Zou, Hui, and Trevor Hastie. 2005. "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B* (*Statistical Methodology*) 67 (2): 301–20.