Journal of Public Economics 95 (2011) 1082–1095

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

Social norms and energy conservation

Hunt Allcott

MIT, United States
New York University, United States

ARTICLE INFO

Article history:
Received 9 June 2010
Received in revised form 2 February 2011
Accepted 7 March 2011
Available online 21 March 2011

JEL classifications:
C44
D12
L94
Q41

Keywords:
Social norms
Energy demand
Randomized field experiments

ABSTRACT

This paper evaluates a series of programs run by a company called OPOWER to send Home Energy Report letters to residential utility customers comparing their electricity use to that of their neighbors. Using data from randomized natural field experiments at 600,000 treatment and control households across the United States, I estimate that the average program reduces energy consumption by 2.0%. The program provides additional evidence that non-price interventions can substantially and cost effectively change consumer behavior: the effect is equivalent to that of a short-run electricity price increase of 11 to 20%, and the cost effectiveness compares favorably to that of traditional energy conservation programs. Perhaps because the treatment included descriptive social norms, effects are heterogeneous: households in the highest decile of pre-treatment consumption decrease usage by 6.3%, while consumption by the lowest decile decreases by only 0.3%. A regression discontinuity design shows that different categories of "injunctive norms" played an insignificant role in encouraging relatively low users not to increase usage.

1. Introduction

Climate change has emerged as one of the most important economic policy issues of the early 21st century, and many view energy efficiency as an appealing approach to reducing greenhouse gas emissions. Traditionally, economists and policymakers have focused on relative prices as the primary force driving energy demand. As a result, carbon cap-and-trade programs are the centerpiece of proposed climate change policies, and subsidies for energy efficient durable goods draw the vast majority of public energy efficiency funding in the U.S. (Gillingham et al., 2006).

There are three problems with price-based approaches to energy conservation. First, it has not been politically feasible to implement Pigouvian carbon taxes or a carbon emissions trading program in the U.S., suggesting that average wholesale energy prices are below social cost. Second, measuring the effects of an energy efficiency subsidy on energy use requires knowledge of the elasticities of demand for energy efficient durable goods and for energy conditional on capital stock. Lacking context-specific values of these parameters, subsidy-based programs are typically evaluated using a controversial approach called "deemed savings"; randomized controlled impact evaluations are exceedingly rare. A third problem is that while subsidies are in theory innocuous because they are transfers, they are in practice a large drain on increasingly-limited public funds.

Spurred by these problems, interest has dramatically increased in non-price energy conservation programs that are informed by insights from behavioral science and evaluated via randomized trials. Non-price interventions are typically inexpensive relative to subsidies, and as demonstrated by Bertrand et al. (2010) in the context of consumer finance, carefully-crafted psychological cues can have effects on demand that are comparable to large changes in relative prices. A critical challenge, however, is to craft interventions that are powerful and cost-effective when implemented at large scale.

This paper examines one of the most notable non-price energy conservation programs, which is run by a company called OPOWER. OPOWER mails Home Energy Report letters (HERs) that compare a household's energy use to that of similar neighbors and provide energy conservation tips. The neighbor comparisons were directly influenced by academic work showing that providing social norm information induces people to conserve energy (Schultz et al., 2007; Nolan et al., 2008). More broadly, the program was motivated by similar evidence on the power of social norms in a variety of domains, including voting (Gerber and Rogers, 2009), retirement savings (Beshears et al., 2009), and charitable giving (Frey and Meier, 2004). As of the end of 2010, OPOWER had contracts to run programs at 47 utilities in 21 states, including six of the largest ten utilities in the U.S.

The first parts of this paper are an impact evaluation of all of the OPOWER programs begun before the end of 2009. With nearly 600,000 households in treatment and control groups, this is one of the largest randomized field experiments in history. I show that the point estimates of the Average Treatment Effects (ATEs) of OPOWER's first

E-mail address: allcott@mit.edu.

0047-2727/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jpubeco.2011.03.003

17 experiments range from 1.4 to 3.3%, with an unweighted mean of 2%.[1] While there is often concern over the durability of treatment effects in similar non-price interventions (Ferraro and Price, 2010), the Home Energy Reports appear to have constant or increasing effects as they are repeatedly delivered over the first two years of treatment.

These effect sizes have several different economic interpretations. First, different energy conservation programs are typically compared on a basis of program implementation cost per kilowatt-hour of electricity saved. OPOWER's initial set of programs have cost effectiveness ranging from 1.3 to 5.4 cents per kilowatt-hour, with an unweighted mean of 3.3. These results compare favorably to estimates for traditional energy efficiency programs, and because they are estimated using randomized trials, they are much more certain. The welfare effects, however, are ambiguous: the costs that households incur to reduce energy use are unobserved, as is the change in welfare from learning that one compares favorably or poorly to neighbors.

A second way of interpreting effect sizes is to calculate the energy price changes that would induce the same changes in demand. Calibrating with estimated price elasticities, I show that the effects of sending Home Energy Reports are equivalent to a 11 to 20% short-run price increase or a 5% long run price increase. Taken as a whole, these effects are remarkable: simply sending letters can significantly and cost-effectively affect energy use behaviors.

The remainder of the paper builds on theoretical predictions of heterogeneous treatment effects. In particular, many models predict that the "descriptive norm" element of the Home Energy Report treatment, in which a household's energy use is compared to that of its neighbors, would cause households that previously used more than the norm to decrease usage, but would cause households that used less than the norm to use more. Social psychologists call these unintended consequences "boomerang effects" (Clee and Wicklund, 1980), and they are certainly undesirable if the objective is to induce energy conservation. Combining data across all of OPOWER's experiments, I show that Conditional Average Treatment Effects are larger than 6% in the highest decile of pre-treatment usage and close to zero in the lowest decile, but even these households that compare most favorably to their neighbors do not increase energy use. In this sense, the OPOWER intervention does not cause a "descriptive norm boomerang effect."

The Schultz et al. (2007) experiment that motivated OPOWER's work had found a boomerang effect for relatively low users. To combat this, they employed what social psychologists call "injunctive norms," which convey that energy conservation is pro-social (Cialdini et al., 1990). Specifically, they added a treatment condition that included hand-drawn "smiley faces" on the descriptive norm feedback reports given to these relatively low users. Although the group of low users that received this injunctive norm did not use statistically significantly less energy than the group that that did not, this group's increase in energy use was also not statistically distinguishable from zero. Based on this result, it was believed that injunctive norms could eliminate the boomerang effect.

OPOWER's Home Energy Reports therefore include injunctive norms, which are defined based on sharp cutoffs. Households are labeled as "Great" if they use less than the 20th percentile of their neighbor comparison group, "Below Average" if they use more than the mean, and "Good" if they are in between. The "Great" group receives two "smiley face" emoticons, the "Good" group receives one, and the "Below Average" group initially received "frownie faces" until customer complaints ended this practice. The treatment effects are substantially different across the three groups, although this could be caused either by the categorizations or by other factors correlated with baseline energy use that could affect how households respond to the treatment: for example, high users may have lower-cost opportunities to conserve. Notice, however, that households that had used just more energy than the 20th percentile of their comparison group are in the limit identical to households that use just less energy, but the former are labeled "Good," while the latter are labeled "Great." Similarly, households using just more than the mean of their comparison group were labeled "Below Average," while households using just less were labeled "Good."

I use a regression discontinuity (RD) design to test for whether these normative categorizations cause differential effects on energy use. I show that while the treatment effects differ substantially for households in the three different categories, the causal effects of the categorizations themselves are "tightly estimated zeros." Being labeled "Good" instead of "Great" has a differential treatment effect of less than 0.20 percentage points, or about one-tenth of the ATE. Being labeled "Below Average" instead of "Good" has a differential ATE of less than 0.16 percentage points. Therefore, the fact that we do not observe a descriptive norm boomerang effect is likely due not to the different categorizations. Instead, the potential effect is likely mitigated by the energy conservation tips or other aspects of the injuntive norms that affect all categories equally.

The paper proceeds by first giving more detail on the treatment and potential pathways of effects. The rest of Section 2 then provides background and descriptive statistics on OPOWER's experiments. Section 3 details the average treatment effects, from the econometric strategy to the parameter estimates and resulting cost effectiveness. In the spirit of Lalonde (1986), this section also documents the poor performance of non-experimental estimators. Section 4 discusses heterogeneous treatment effects and the RD design. Section 5 concludes.

2. Experiment overview

2.1. The treatment and mechanisms of effects

The Home Energy Reports are several-page letters with two key components. The first is the Social Comparison Module, which appears at the top of the letter's first page. As illustrated in Fig. 1, the graph on the left side of the Social Comparison presents the "descriptive norm" by comparing the household to the mean and 20th percentile of its comparison group. A household's comparison group comprises approximately 100 geographically-proximate houses with similar characteristics, including similar square footage and same heating type (gas vs. electric). The "Efficiency Standing" on the right side of the Social Comparison Module adds the injunctive norm by categorizing the household as "Great," "Good," or "Below Average."

The Report's second key component is the Action Steps Module. As illustrated in Fig. 2, these energy conservation tips include both changes to the household's stock of energy-using durable goods and to the use of that capital stock. These suggestions are targeted to different households based on historical energy use patterns and demographic characteristics. For example, households whose energy use was relatively high the previous summer were more likely to receive suggestions to purchase new energy efficient air conditioners.

To conceptualize the mechanisms through which the treatment acts, informally consider a model of energy demand in the style of Becker's (1965) household services model. The household derives

[1] There are also other analyses of OPOWER's projects. For regulatory reasons, each experiment is evaluated by industry program evaluators, so there are a growing number of consulting reports, such as Violette et al. (2009). A working paper by Ayres et al. (2009) evaluates OPOWER's programs in Sacramento and Puget Sound, and I therefore refrain from directly discussing those programs and refer readers to that paper for additional information. An earlier version of this paper (Allcott, 2009) focused specifically on OPOWER's program with Connexus Energy in Minnesota. A working paper by Costa and Kahn (2010) shows that OPOWER's CATEs at one West Cost site are stronger for liberal voters than for conservatives. More broadly, there is a long psychology literature on similar energy use information feedback programs, as reviewed in Abrahamse et al. (2005), Darby (2006), and Stern (1992).

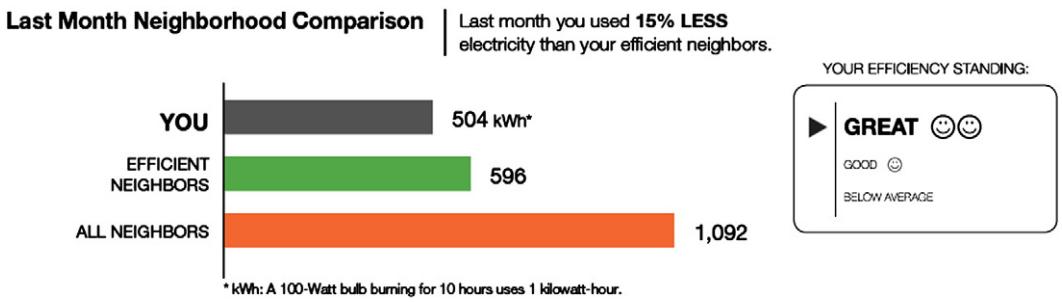

Fig. 1. Home energy reports: social comparison module.

utility from "energy services," such as warmth and television, and a composite good. As in Dubin and McFadden (1984) and Davis (2008), the household invests in "energy efficiency," or the rate of transformation of energy input into energy services, which can be increased at some cost. Conditional on energy efficiency, the household sets demand for energy and the composite good.

Consumers also receive "moral utility" (Levitt and List, 2007) from energy conservation, as this contributes to public goods such as reduced greenhouse gas emissions. This moral utility term depends on beliefs about the social norm. It seems likely that untreated households believe that they are closer to the social norm than they actually are, meaning that the treatment causes low (high) usage households to update beliefs about the social norm upward (downward).

Perhaps at the expense of other pathways, consider three primary mechanisms through which the treatment could act. First, the Action Steps tips provide information that allows the household to increase energy efficiency at lower cost. Second, if households are uncertain about some part of their production function, the social comparisons may facilitate social learning about their privately-optimal level of energy use, as documented in other contexts by Beshears et al. (2009), Cai et al. (2009), Conley and Udry (2010), Foster and Rosenzweig (1995), Mobius et al. (2005), Munshi (2004), and Munshi and Myaux (2006). Third, the treatment may directly affect the "moral cost" of energy use. This could happen because injunctive norms or other factors increase the moral cost of energy use for all recipients. Alternatively, the treatment could increase the moral cost for households using more than the norm and decrease it for those using less. This would generate "conditional cooperation," in which households increase (decrease) their contribution to a public good after being informed that others are contributing more (less) than previously believed, as in Alpizar et al. (2008), Fischbacher et al. (2001), Frey and Meier (2004), and Shang and Croson (2004).

2.2. Experimental design

This paper analyzes the Home Energy Report projects that OPOWER had begun by late 2009. There were projects with twelve utilities: six in California and Washington, six in the midwest, one in the urban Northeast, and one in a suburban area in a Mountain state. For business reasons, OPOWER has asked that the experimental results not be associated with the names of each partner, so the experiments will be referred to by numbers. The exception is Connexus Energy in Minnesota, Experiment 4, which will be used as an example in several instances later in the paper.

Regulated utilities typically have the incentive to increase instead of decrease their customers' energy use. Why are utilities working with OPOWER? The company's partners are typically either non-profit municipal utilities whose goals include energy conservation

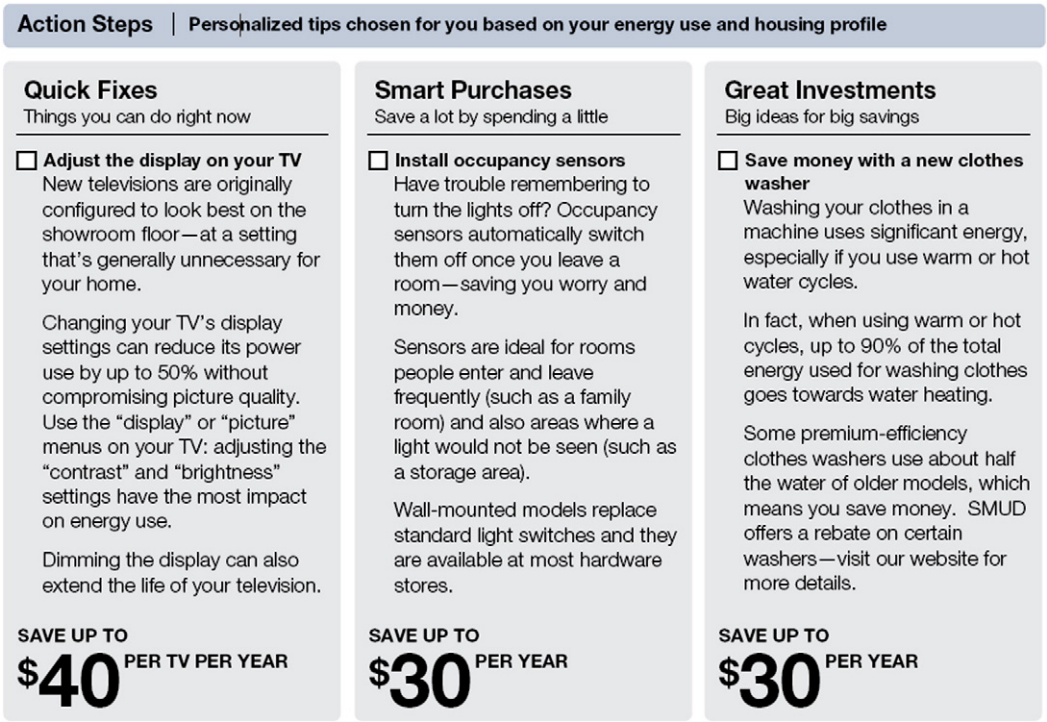

Fig. 2. Home energy reports: action steps module.

Table 1
Overview of OPOWER projects.

Experiment			N		
Number	Region	Start date	Households	Treatment	Observations
1	Rural Midwest	February, 2009	8175	8175	343,729
2	Urban Midwest	July, 2009	37,484	18,790	1,264,375
3	Urban Midwest	July, 2009	56,187	28,027	1,873,482
4	Rural Midwest	January, 2009	78,273	39,024	3,421,306
5	Suburban Mountain	October, 2009	11,612	7,254	394,525
6	Suburban Mountain	October, 2009	27,237	16,947	914,344
7	West Coast	October, 2009	24,940	23,906	570,386
8	Rural Midwest	April, 2009	17,889	9,861	794,457
9	Urban Northeast	September, 2009	49,671	24,808	1,712,530
10	Rural Midwest	February, 2009	8429	8,390	360,577
11	West Coast	October, 2008	79,229	34,893	3,121,879
12	West Coast	January, 2009	25,211	5,570	985,148
13	West Coast	January, 2009	17,849	3,852	672,629
14	West Coast	January, 2009	22,965	22,846	893,322
15	West Coast	September, 2009	39,336	19,663	671,990
16	West Coast	March, 2008	59,666	24,761	2,543,372
17	West Coast	April, 2008	24,293	9903	1,036,768
Combined		March, 2008	588,446	306,670	21,574,819

or regulated investor-owned utilities in one of 24 states where policymakers have enacted energy conservation mandates called Energy Efficiency Resource Standards (EERS). These regulations require that electricity and natural gas retailers run energy conservation programs that reduce the quantity of energy demanded in their service territory by some amount relative to counterfactual, typically a few percent over several years (ACEEE, 2010). For example, under Minnesota's New Generation Energy Act of 2007, utilities in that state are required to run conservation programs that reduce energy demand by 1.5% each year.

The eligible experimental populations at each utility included residential customers with sufficient electricity bill history to construct historical neighbor comparisons.[2] In some utilities, the entire customer base was included, while in others, only heavier users were eligible. OPOWER randomized the experimental population into a Treatment group, which would be mailed Home Energy Reports, and a Control group, which would not. At two utilities and a subpopulation of a third, the populations were not large enough to make control groups worthwhile. While data from these treatment-only experiments will be used for part of a regression discontinuity analysis in Section 5, they will not be used to estimate Average Treatment Effects of the Home Energy Reports.

Reports are sent to the Treatment group monthly, bimonthly, or quarterly, depending on the utility. In some of the more recent projects, letters are sent each month for the first several months of the program, with a lower frequency after that. In four experiments, the populations were divided into sub-populations with higher and lower baseline usage, with the Treatment groups in the high-usage subpopulation receiving more frequent Reports. In total, there were 17 separate experimental populations randomized into treatment and control across the 12 utilities. In Connexus and in Experiment 11, the population was randomized into monthly vs. quarterly frequency, and in Experiment 3, population was randomized between bimonthly and quarterly frequencies. Table 1 gives an overview of the start date and size of experiment for each project. In total, I observe 22 million utility bills from nearly 600 thousand households across the United States.

Most utilities send a worker to read each customer's electricity meter each month.[3] Within several days of the meter read, the results are sent electronically to OPOWER, where each household's social comparison is computed. The Home Energy Report is printed by an outside contractor and sent via U.S. Mail. Any meter reads more than 30 days after the day on which OPOWER generated the first report are considered "post-treatment." This is typically the second meter read after the one upon the first Report was based.

2.3. Data and baseline characteristics

Table 2 presents descriptive statistics for each experiment. Baseline electricity usage ranges from 19 to 60 kilowatt-hours (kWh) per day. For context, consider that a medium-sized (60 W) lightbulb used 5 hours each day consumes 0.3 kWh. A typical window air conditioner running at its highest setting for 5 hours uses 5 kWh. As illustrated in Fig. 3, heating and cooling are the primary uses of household electricity in the United States: over half of annual electricity consumption is for refrigerators, air conditioners, and space and water heating. In the most recent available data, computers, televisions, and lighting combined account for only 15% of electricity use (US Energy Information Administration, 2001).

As shown by the *p*-values in parenthesis of the fourth column of Table 2, baseline usage is balanced between Treatment and Control except in Experiments 2, 6, and 8. These imbalances are difficult to explain, and OPOWER has now begun to confirm covariate balance before finalizing the randomization. As the empirical specifications will use household fixed effects, any imbalance in the pre-treatment outcome does not mechanically bias results. Readers concerned about these imbalances, however, could discount the results from these three experiments.

Aside from the monthly electricity meter readings for each household, I also observe OPOWER's social comparison information for every report at every household, including whether they were rated as "Below Average," "Good," or "Great," and how far they were

[2] There were several other technical restrictions on the experimental population. Households had to have valid names and addresses, no negative electricity meter reads, at least one meter read in the last three months, no significant gaps in usage history, exactly one account per customer per location, and a sufficient number of neighbors. A handful of utility staff and "VIPs" were automatically enrolled in the reports and are therefore excluded from the analysis. Households on special medical rate plans were also excluded. These additional exclusions eliminate only a small portion of the potential population. None of the exclusions make the results less externally valid, as future programs will also be carried out in similarly restricted populations.

[3] The mechanics of the meter reading process actually vary somewhat by utility. The utility in Experiment 11 has automated metering infrastructure that records energy use daily. The utilities in Experiments 15 and 7 send workers to read each household's electricity meter once every two months, and 81% of, billing period lengths are between 57 and 65 days. All other program utilities read meters monthly, with 93% of billing period lengths between 28 and 34 days. A small percentage of bills are based on estimated meter reads, but the empirical analysis considers actual meter reads only.

Table 2
Descriptive statistics.

Experiment		Stats				
Number	Region	Y_0 (kWh/day)	$Y_0^T - Y_0^C$	% Moved	% Moved (T-C)	% Opt out
1	Rural Midwest	23 (12)	Non-Exper	14.3	Non-Exper	0.1
2	Urban Midwest	60 (31)	−1.18 (0.00)	5.5	0 (0.91)	0.4
3	Urban Midwest	31 (6)	0.04 (0.44)	7.9	−0.3 (0.22)	0.2
4	Rural Midwest	30 (17)	0.04 (0.74)	6.7	−0.1 (0.40)	1.7
5	Suburban Mountain	40 (12)	−0.06 (0.80)	13.7	−1.5 (0.03)	2.9
6	Suburban Mountain	19 (6)	0.22 (0.00)	20.2	−0.1 (0.81)	1.4
7	West Coast	18 (11)	0.09 (0.47)	11.3	−0.2 (0.64)	0.6
8	Rural Midwest	39 (27)	0.98 (0.02)	6.7	−0.4 (0.35)	2.0
9	Urban Northeast	30 (15)	−0.21 (0.12)	5.4	−0.2 (0.25)	0.5
10	Rural Midwest	24 (12)	Non-Exper	16.2	Non-Exper	0.5
11	West Coast	30 (14)	0.02 (0.86)	10.9	0 (0.98)	1.4
12	West Coast	39 (24)	0.12 (0.74)	13.2	−0.6 (0.22)	0.7
13	West Coast	20 (14)	0.25 (0.32)	17.9	0.1 (0.84)	0.4
14	West Coast	29 (21)	Non-Exper	12.6	Non-Exper	0.6
15	West Coast	37 (18)	0.01 (0.96)	5.9	0 (0.84)	0.7
16	West Coast	37 (14)	−0.54 (0.19)	15.5	0.5 (0.36)	3.3
17	West Coast	16 (4)	0.03 (0.76)	16.0	0.3 (0.70)	1.0

Y_0 is the average of electricity use for the 12 months preceding the beginning of treatment. $Y_0^T - Y_0^C$ and % Moved columns: p-values in parenthesis.

from the cutoffs to be in each of the other categories. I observe both the social comparisons that the Treatment group did receive and what the Control group would have received. Weather data from the National Climatic Data Center are used to construct the average Heating Degree-Days and Cooling Degree-Days over the days in each billing period, which are associated with the amount of electricity that should be required to keep a house at a comfortable temperature.[4]

Finally, I observe an extensive set of household-level covariates from utility surveys, public records such as property tax assessments, and private-sector marketing data providers. Depending on the utility, observed house characteristics may include year constructed, whether gas or electric heat, assessed value, square footage, whether single-family or multi-family dwelling type, whether rented or owner-occupied, whether it has a fireplace or pool, and the number bedrooms and bathrooms. Occupant characteristics may include number of residents, age of household head, and income.

2.4. Attrition

The programs experience two forms of attrition, moving and opting out. The fifth column of Table 2 lists the cumulative probability of moving for a household over the life of each experiment, which ranges from 5.4% in Experiment 9 to 20% in Experiment 6. Households that close accounts are different: they are younger, use less electricity, live in smaller, older homes, have lower incomes, and are more likely to rent and live in multifamily buildings. Naturally, the treatment does not cause households to move. As shown in the sixth column of Table 2, moving is not unbalanced with 90% confidence in 13 of the 14 randomized experiments. In all 14, F-tests show that there is no statistical difference between Treatment and Control in movers' observable characteristics.

Energy bills are not observed after the resident moves. Households that moved after the program began are included in the base specifications during the period when their usage is observed, giving an unbalanced panel. Excluding these households has no discernible influence on the results.

[4] More precisely, average Cooling Degree-Days is the mean, over all of the days in the billing period, of the maximum of zero and the difference between the day's average temperature and 65°. A day with average temperature 95 has 30 CDDs, while a day with average temperature 60 has zero CDDs. Average Heating Degree-Days is the mean, over all the days in the billing period, of the maximum of zero and the difference between 65° and the day's average temperature. A day with average temperature 95 has zero HDDs, while a day with average temperature 60 has five HDDs.

The second form of attrition is that some households asked to stop receiving the Reports. As shown in the rightmost column of Table 2, cumulative opt-out rates range from 0.1% to 3.3%. The most common reasons for opting out is the perception that the comparisons are unfair or inaccurate or that the reports are a "waste of resources." Across the different utilities, customers that opt out tend to have higher pre-treatment electricity usage and are also older and lower-income. Although they opted out of receiving Reports, their electricity bills are still observed.

3. Average treatment effects

3.1. Estimation

The initial estimand of interest is the Average Treatment Effect $\tau = E[Y_{it}(1) - Y_{it}(0)]$ in the population of experimental households, where $Y_{it}(1)$ and $Y_{it}(0)$ denote the "potential outcomes" for household i's electricity use at time t if the household were treated and were not treated, respectively (Rubin, 1974). As some households opted out, the "Treatment" here is defined as "being mailed the Home Energy

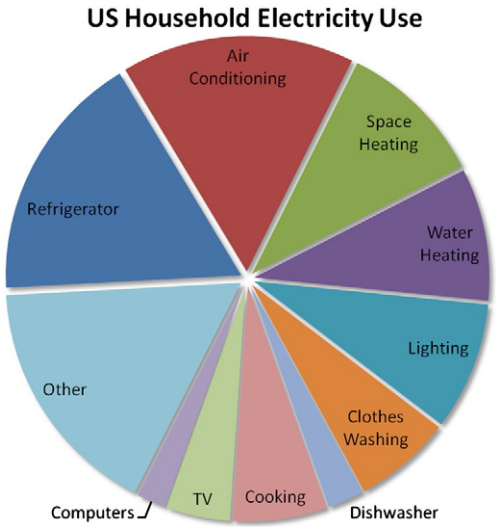

Fig. 3. US household electricity use.

Table 3
Connexus ATE specifications.

	I	II	III	IV	V
T × Monthly × Post	−2.65	−2.72	−2.72	−2.69	−2.74
	(0.27)	(0.18)	(0.18)	(0.16)	(0.18)
T × Quarterly × Post	−2.46	−2.26	−2.26	−2.23	−2.26
	(0.37)	(0.21)	(0.21)	(0.18)	(0.21)
Post	−3.70	−5.82	−2.41	−5.04	−0.63
	(0.12)	(0.11)	(0.46)	(0.36)	(0.46)
T	0.19				
	(0.40)				
Degree-day bins	No	No	No	No	Yes
Month × Year dummies	No	No	Yes	Yes	Yes
House fixed effects	No	Yes	Yes	No	Yes
House × Month fixed effects	No	No	No	Yes	No
Observations (thousands)	3421	3421	3421	3421	3421
R^2	0.0016	0.0016	0.0586	0.0000	0.0651
F statistic	874	2868	4643	3564	

Standard errors in parentheses. Dependent variable is the household's average daily electricity consumption (kilowatt-hours), normalized by average control group consumption in the Post period.

Table 4
ATEs for all experiments.

Experiment	ATEs (%)		
Number	Monthly	BiMonthly	Quarterly
1	Non-Exper	–	–
2	−1.83 (0.20)	–	–
3	–	−1.40 (0.19)	−1.37 (0.19)
4	−2.72 (0.18)	–	−2.26 (0.21)
5	–	−2.70 (0.44)	–
6	–	–	−1.64 (0.33)
7	–	−2.48 (0.25)	–
8	–	−3.32 (0.54)	–
9	–	−1.63 (0.15)	–
10	Non-Exper	–	–
11	−1.96 (0.14)	–	−1.49 (0.20)
12	−1.39 (0.34)	–	–
13	–	–	−1.44 (0.51)
14	–	Non-Exper	–
15	–	−1.89 (0.21)	–
16	−3.14 (0.37)	–	–
17	–	–	−1.84 (0.43)
Mean ATE	−2.03		

Reports or actively opting out."[5] The primary specification is a difference-in-differences estimator that models energy use conditional on Treatment group indicator T_i, post-treatment indicator P_{it}, month-by-year dummy variables μ_{my} and household fixed effects υ_i:

$$Y_{it} = \tau T_i P_{it} + \beta \cdot P_{it} + \mu_{my} + \upsilon_i + \varepsilon_{it} \quad (1)$$

This is estimated in OLS using the standard fixed effects estimator, using Huber–White ("robust") standard errors, clustered by household. As discussed by Bertrand et al. (2004), these standard errors are consistent in the presence of any correlation pattern in the errors ε_{it} within household over time.

3.2. Results

Because the sample sizes are large, estimated treatment effects tend to be robust to different configurations of fixed effects and controls. As an example, Table 3 presents estimated ATEs for the Connexus program. Column III is the primary specification detailed above, while the other four columns use different configurations of fixed effects, month-by-year dummies, and weather controls. The ATEs can be interpreted as percentage change, as electricity usage has been normalized by dividing by the average post-period control group consumption in each experiment and multiplying by 100. In this experiment, households were randomly assigned between a monthly Treatment group, a quarterly Treatment group, and a Control group. The point estimates of ATEs are clustered around negative 2.7% and 2.3% for monthly and quarterly treatment, respectively, and they are not statistically different across specifications.

Table 4 presents the ATEs for all experiments estimated using the primary specification detailed above. The ATEs range from 1.37% in the quarterly treatment arm of Experiment 3 to 3.32% in Experiment 8. The unweighted mean is 2.03%.

Some of the differences in ATEs are associated with treatment frequency: the unweighted mean ATEs for bimonthly and monthly treatments are 2.2%, while the average for quarterly experiments is 1.7%. Some of this association is causal: in Connexus and Experiment 11, the population was randomly assigned between monthly and quarterly, and the increased frequency causes a 0.5% larger ATE. In other experiments, populations with higher pre-treatment usage were assigned to more frequent treatment. Section 4 documents how households with higher pre-treatment usage have larger treatment effects conditional on frequency.

Even after conditioning on frequency and differences in observed household-level covariates, there is substantial unexplained variation in treatment effects across experiments. This presents a practical challenge for utilities that are considering adopting the program: it is more difficult to precisely forecast the program's potential performance in their location, and thus more difficult to determine whether to adopt. In a separate paper, Allcott and Mullainathan (2010b) explore the generalizability of this set of experimental results.

How persistent are these effects over time? Again using Connexus as an example, Fig. 4 presents treatment effects for each month of the experiment, for both the monthly and quarterly treatment groups. This figure is generated by interacting the full set of month-by-year dummies with both the monthly and quarterly treatment dummies, including the same set of month-by-year dummies as controls and using household fixed effects. The first reports were sent in January 2009, and this is the excluded month.

Fig. 4 shows that after treatment begins, the treatment effects take several months to ramp up to something approximating a steady state. Notice that the percentage treatment effects are higher in the winter and summer months, when heating and cooling loads increase underlying demand, than in the fall. After nearly two years of continuing treatment, there is no evidence of any decline in the treatment effects. In fact, the treatment effects are larger in the summer of 2010 than in the summer of 2009, although higher temperatures in summer 2010 could be responsible for this. In the three experiments with a full two years of post-treatment data, Experiments 11, 16, and 17, the ATEs are higher in the second year than the first year, and the weather in the first and second years is comparable.

Notice, however, that the Connexus quarterly group's treatment effects decay between March and May 2009. Due to a technical problem, the second round of quarterly reports, which would have been sent in April 2009 and thus would have effects observed beginning in May, were delayed by one month. Although this is only one data point, this is consistent with the idea that the effects would diminish over time in the absence of additional Reports.

[5] Alternatively, the Treatment could have instead been dened as "being mailed the Home Energy Reports," in which case my estimand would be interpreted as an Intent-to-Treat effect. The "Treatment" is also not "opening the Home Energy Report." Although it is quite likely that many Treatment group households do not open the Reports, and thus that the effect of the "Treatment" thus dened would be higher, it is di ¢ cult to measure letter open rates, and thus not possible to estimate this effect.
The appeal of the definitions used here is that they generate a useful estimand from a policy perspective. OPOWER, and the utilities that contract with them and policymakers that regulate them, want to know the aggregate electricity conservation possible from applying the program to an eligible population. For the eligible population from which the experimental households were drawn, this quantity of interest can be derived simply by multiplying my ATE by the population size.

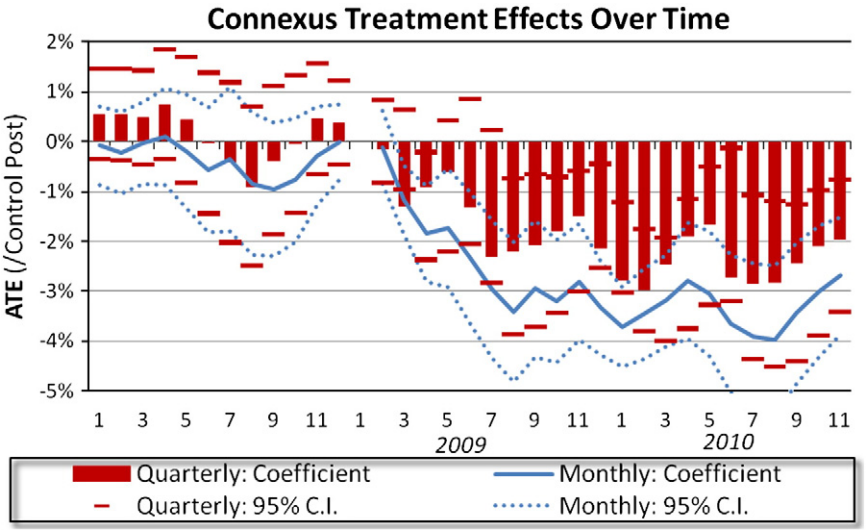

Fig. 4. Connexus treatment effects over time.

3.3. Interpreting effect sizes

3.3.1. Relative to changes in behavior

What actual activities underlie these changes in energy use? In one of their pilot programs, OPOWER has collected surveys in which Treatment group households were asked to self report what they had changed as a result of receiving the Home Energy Reports. Some of the reported effects were changes to household capital stock, including weather-stripping windows, improving insulation, or servicing the air conditioner. Many of the most frequently reported changes, however, were day-to-day usage behaviors: turning off lights, unplugging electronics, adjusting thermostats, and closing window blinds. Interestingly, these are behaviors that most consumers likely *already knew* could save them energy. This suggests that at least some of the letters' effects act through drawing attention or increasing the "moral cost" of energy use, instead of solely by providing new information or inducing changes in capital stock.

How do the percent ATEs translate into these real-world behaviors? The 2% mean ATE translates into 0.62 kWh per day. An air conditioner running at full power uses about 1 kW of power, so this treatment effect is equivalent to turning off an air conditioner that would have been on for 37 min each day. A standard incandescent lightbulb uses 60 W, so the treatment effect is also equivalent to 10.4 h of lightbulb use per day. Recall that many "treated" households likely do not open, understand, or act on the letters, meaning that the effects on those who do must be much larger. These effects seem remarkably large, given that the treatment is as simple as sending a letter.

3.3.2. Relative to prices

Another useful way to frame the effects of a non-price intervention is to calculate the price change that would produce similar effects. Given that treatment effects are visible very soon after the letters are received and the effects stabilize after several months, short run price changes are an appropriate comparison. Using the event study in Reiss and White (2008), the "60-day elasticity" of residential electricity demand in California with respect to a large, unanticipated price change is −0.10 to −0.18. With appropriate caution in generalizing this elasticity, it would imply that the effects of OPOWER's average existing program are equivalent to a short run electricity price increase of 11 to 20%. This again is remarkable: a simple non-price treatment changes consumer behavior as much as substantial price increases.

From a policy perspective, one might also be interested in comparing these effects to those of sustained increases in electricity prices that might result from climate change policies that would regulate carbon dioxide emissions. Reiss and White (2005) estimate that the long-run elasticity of residential electricity demand in California is −0.39, implying that OPOWER's effects would be equivalent to a sustained 5.2% increase in prices. For comparison, the US Energy Information Administration (2009) estimates that a recently-proposed carbon cap-and-trade program would increase electricity prices by 2.5% in 2020 and 20% in 2030. This underscores the appeal of non-price treatments like the Home Energy Reports: they appear to be somewhat less controversial than a carbon cap-and-trade or tax and are inexpensive for the policymaker to implement, but their effects are within a similar order of magnitude.

3.3.3. Cost effectiveness

OPOWER's cost effectiveness, defined as cents of cost to the program administrator per kilowatt-hour of electricity conserved, is a statistic of great practical interest. Just as international aid agencies and health programs have an array of possible projects that they can fund, energy conservation program administrators also have a set of available programs. In many settings, the administrator will have a regulatory energy conservation target such as an Energy Efficiency Resource Standard that it must achieve using a fixed budget. The administrator chooses to contract with OPOWER if the cost effectiveness compares favorably to other energy conservation opportunities.

OPOWER's cost effectiveness is the annualized cost of the Reports divided by kilowatt-hours saved per year. The numerator is estimated by multiplying the cost per report by the annualized number of Reports delivered during the program to date to the Treatment group. OPOWER has provided cost data on a confidential basis, but the cost of printing and mailing a short letter like the Home Energy Report is on the order of one dollar. The denominator is calculated by taking the average treatment effect $\hat{\tau}$, which is in percent, and multiplying it by the average daily kilowatt-hours of consumption Y_0 in the year preceding the program. The resulting formula is:

$$\text{Cost Effectiveness} = \frac{\text{Cost per Report} \cdot \text{Reports per Year}}{(\hat{\tau}/100) \cdot Y_0 \cdot 365} \tag{2}$$

The unweighted average cost effectiveness across experiments is 3.31 cents per kilowatt-hour saved. Fig. 5 displays the result for each experiment, with experiments organized from left to right in order of decreasing baseline usage. There is a clear general trend of improving

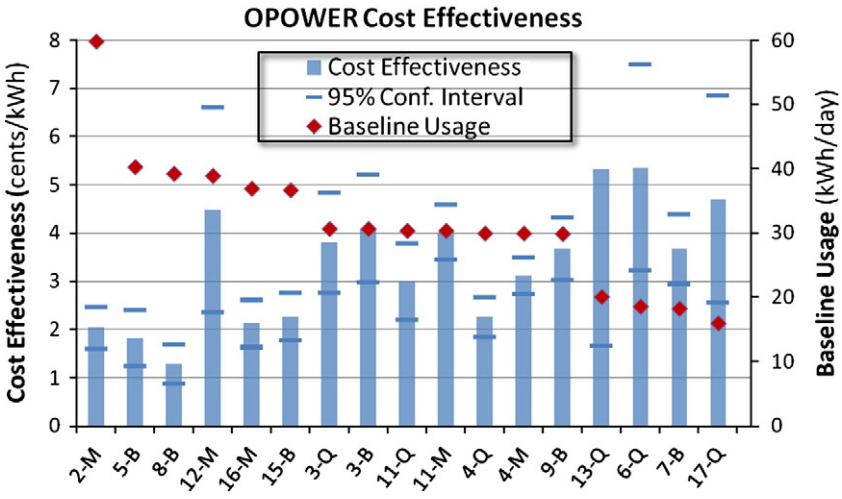

Fig. 5. Cost effectiveness.

cost effectiveness for populations of heavier users. This is because, as Section 4 details, treatment effects are larger for heavier users.

Treatment frequency also has an important impact on cost effectiveness. To see this, consider Experiments 3, 11, and 4 in the middle of Fig. 5. The population in Experiment 3 was randomized between quarterly ("Q") and bi-monthly ("B") frequencies, while the latter two populations were randomized between monthly and quarterly. As we have seen, more frequent treatment gives larger ATEs, but Fig. 4 shows that this does not offset the increased annualized cost. These results suggest that quarterly treatment is most cost effective.

Partially because there have been so few randomized impact evaluations, there is controversy in the applied literature over the cost effectiveness of typical energy conservation programs. Arimura et al. (2009) use utility-level data from a nationwide annual panel to correlate reported "demand-side management" program expenditures with changes in electricity use. They estimate cost effectiveness to be about 6 cents per kilowatt-hour, with confidence intervals ranging from 5.5 to 6.4. For their estimates to be unbiased, program expenditures must be uncorrelated with other policies or factors that could reduce energy demand. Friedrich et al. (2009), of an energy efficiency research and advocacy organization called the American Council for an Energy Efficient Economy, focus on utilities in 14 states with aggressive energy conservation programs. They divide reported administrative costs by engineering estimates of kilowatt-hours saved, giving state-level results ranging from 1.6 to 3.3 cents per kilowatt-hour. There has historically been concern that these "deemed savings" estimates are optimistic (Nadel and Keating, 1991).

One of the conceptual ways in which OPOWER has begun to change this space is by showing that randomized controlled trials both are feasible to implement and produce uncontroversial estimates of impacts and cost effectiveness. Randomized trials have also offered the company the opportunity to fine-tune their results, for example by designing more recent programs around more cost effective treatment frequencies. Taking the above two sets of benchmark cost effectiveness estimates at face value, many of OPOWER's existing programs compare favorably, which is one reason why increasing numbers of utilities have partnered with the company.

Energy conservation programs are in practice often evaluated using this "administrative cost effectiveness" metric, so this calculation is useful for comparison with existing work. This would be, however, a highly incomplete accounting of the welfare effects, both for energy conservation programs in general and for the OPOWER program in particular. There are two main channels of social welfare effects: social costs from energy production and private costs to consumer welfare.

Both channels are difficult to quantify. An important element of the social costs of energy use, the marginal damage of carbon dioxide emissions, is highly uncertain. Furthermore, although the magnitudes might be small, the treatment could affect other unobserved choices that generate carbon emissions. Consumers could, for example, also become motivated to drive their cars less, or could perhaps even drive more due to a "moral license" effect.

Each of the mechanisms through which the treatment could act has different consumer welfare implications. If the treatment affects energy use only by improving information or facilitating social learning, consumers have an unambiguous welfare gain. It is difficult to quantify this welfare gain, however, because the costs of changing capital stock or usage behaviors are unobserved. If the treatment acts only by affecting the moral cost of energy use, households that conserve energy unambiguously lose consumption utility due to the utility costs of substitution, but their change in moral utility is indeterminate. Moral utility could either increase because of the "warm glow" of contributing to the public good or decrease due to a feeling of social pressure to contribute (DellaVigna et al., 2010).

3.3.4. Prior research

How do these results compare to Nolan et al. (2008) and Schultz et al. (2007), the two papers that laid the conceptual foundations for OPOWER? It turns out that because of the small samples, 271 and 286 homes, respectively, those results provided little information on the likely treatment effects or cost effectiveness of OPOWER's programs. In Schultz et al. (2007) the ATE for a treatment most similar to OPOWER – treating both low and high users with both injunctive and descriptive norms – was 5.0% after one month. This was not statistically different than zero, and the 90% confidence interval also included 10%. In Nolan et al. (2008), the treatment most similar to OPOWER – the social comparison treatment relative to a control group that received only energy conservation information – had 10.1% and 7.3% effects after one and two months, respectively. After two months, this was not statistically different than zero, and the 90% confidence interval included 20%. These wide confidence intervals include values that could have made the programs either exceedingly cost effective or massively wasteful relative to alternative energy conservation programs.

Even if the pilots had been somewhat larger, treatment fidelity would have been a concern: the pilot treatment was hand-delivered on doorhangers, which are much more likely to be read than unrequested mail. The "smiley face" injunctive norms were hand-drawn by research assistants, an effort which could increase the

effects but would be difficult to scale. These issues underscore the value from measuring effects of the scaled programs.

3.4. Non-experimental estimators

Nearly all energy efficiency programs are still evaluated using non-experimental estimators or engineering accounting approaches. How important is the experimental control group to consistently-estimated ATEs? This issue is crucial for several of OPOWER's initial programs that were implemented without a control group but must estimate impacts to report to state regulators. While Lalonde (1986) documented that non-experimental estimators performed poorly in evaluating job training programs and similar arguments have been made in many other domains, weather-adjusted non-experimental estimators could in theory perform well in modeling energy demand. The importance of randomized controlled trials has not yet been clearly documented to analysts and policymakers in this context.

Without an experimental control group, there are two econometric approaches that could be used. The first is to use a difference estimator, comparing electricity use in the treated population before and after treatment. In implementing this, I control for weather differences non-parametrically, using bins with width one average degree day. This slightly outperforms the use of fourth degree polynomials in heating and cooling degree-days. This estimator is unbiased if and only if there are no other factors associated with energy demand that vary between the pre-treatment and post-treatment period.

A second non-experimental approach is to use a difference-in-differences estimator with nearby households as a control group. For each experiment, I form a control group using the average monthly energy use of households in other utilities in the same state, using data that regulated utilities report to the U.S. Department of Energy on Form EIA 826. The estimator includes utility-by-month fixed effects to capture different seasonal patterns – for example, there may be local variation in how many households use electric heat instead of natural gas or oil, which then affects winter electricity demand. This estimator is unbiased if and only if there are no unobserved factors that differentially affect average household energy demand in the OPOWER partner utility vs. the other utilities in the same state.

Fig. 6 presents the experimental ATEs for each experiment along with point estimates for the two types of non-experimental estimators. There is substantial variance in the non-experimental estimators: the average absolute errors for the difference and difference-in-differences estimators, respectively, are 2.1% and 3.0%. Across the 14 experiments, the estimators are also biased on average. In particular, the mean of the ATEs from the difference-in-differences estimator is −3.75%, which is nearly double the mean of the experimental ATEs. One potential explanation is a simple form of selection bias: the utilities that partner with OPOWER also devote more effort to other energy efficiency programs over the same period, causing an additional reduction in demand relative to non-partner utilities in the same state.

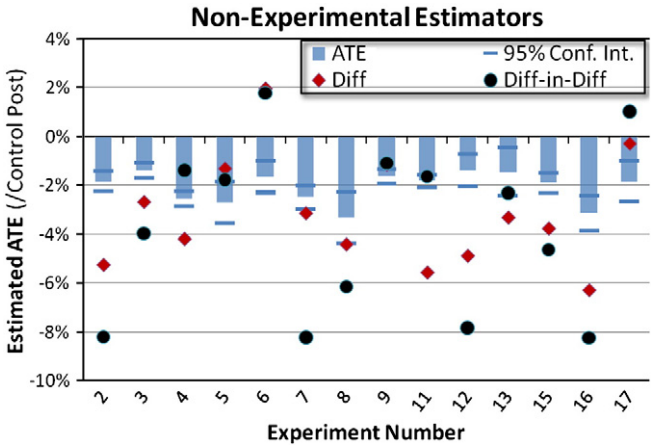

Fig. 6. Non-experimental specifications.

What's particularly insidious about the non-experimental estimates is that they would appear quite plausible if not compared to the experimental benchmark. Nearly all are within the confidence intervals of the small sample pilots by Schultz et al. (2007) and Nolan et al. (2008) that were discussed above. Evaluations of similar types of energy use information feedback programs have reported impacts of zero to 10% (Darby, 2006). Just as Lalonde (1986) motivated labor economists to focus on experimental and quasi-experimental estimators, results like these are crucial in documenting the importance of randomized impact evaluations of energy conservation programs.

4. Heterogeneous treatment effects

While the empirical focus so far has been on Average Treatment Effects, theory predicts that treatment effects could vary over time and across households. Although there is some heterogeneity on other observed characteristics, the primary observable source of heterogeneity is as a function of pre-treatment usage (Allcott, 2009). This could be high-usage households can reduce consumption at lower cost, or alternatively because social norm information has differential effects for households in different parts of the usage distribution.

I first examine Quantile Treatment Effects (QTEs), which are differences between corresponding quantiles of the distributions of household average post-treatment usage in the Treatment and Control groups. For example, the QTE at the 50th percentile is the difference between the median post-treatment usage in the Treatment group and the median post-treatment usage in the Control group. The conditional cooperation and social learning mechanisms predict a "descriptive norm boomerang effect": households with high pre-treatment usage should decrease usage, while those with low pre-treatment usage should increase usage. When examining Quantile Treatment Effects, this would manifest itself as a reduction in the dispersion of usage in the treatment group. If this were the only source of treatment effect heterogeneity, it would cause the QTEs to be positive for low quantiles, indicating an increase in usage, and negative for high quantiles.

In order to hold constant the effects of different treatment frequencies, I focus on the six experiments where frequency was randomly assigned or was the same for all households. For each of these experiments, I estimate the QTEs at the 49 even percentiles (2, 4, 6, 8,…,96, 98) using the procedure in Firpo (2007) and Froelich and Melly (2010). I then combine the sets of QTEs with a minimum distance estimator that minimizes the average of the variances of the 49 QTEs, using an approach akin to that in Eqs. (5) and (7) presented later in the paper.

Fig. 7 graphs the set of QTEs. The point estimates of the second and fourth percentiles are positive, although neither is statistically positive with 90% confidence. Of course, the Quantile Treatment Effects are not "quantiles of the treatment effect," so the fact that the QTES are negative does not necessarily imply that effects are negative for all subgroups or individual households.

A different approach which gives the same qualitative insight is to pool the data for the same six experiments and examine the Conditional Average Treatment Effects for households in different percentiles of their experiment's distribution of baseline usage. Fig. 8 illustrates the CATEs estimated by interacting dummy variables for decile of baseline usage with the treatment effect, controlling for interactions of these deciles with the post-treatment indicator, month-by-year effects, and household fixed effects. The more electricity a household used before the treatment, the more that it conserved post-treatment. The CATEs range from almost zero for the bottom decile of baseline usage to 6.3% in the top decile.

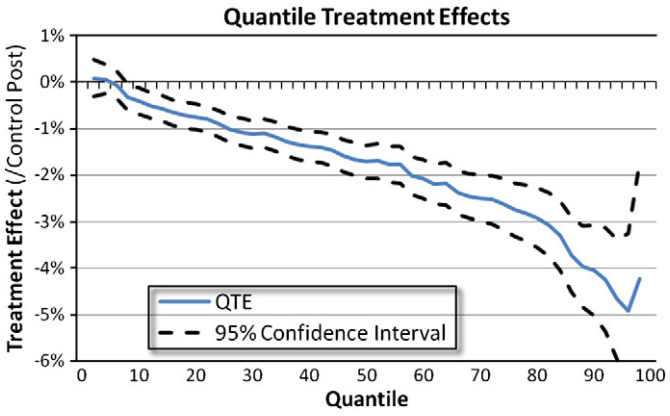

Fig. 7. Quantile treatment effects.

The QTE and decile analyses show that the descriptive norm boomerang effect is not strong enough to induce previously-low usage households to significantly increase usage. This could be because this effect is itself not strong in this setting, because the injunctive normative messages moderate the effect, or because it is outweighed by the impacts of information in the Action Steps Module.

Heterogeneity in treatment effects implies that "profiling," or targeting future treatment toward units with highest Conditional Average Treatment Effects, could raise the Average Treatment Effect on the Treated (ATT) and thus improve the program's cost effectiveness. Unlike some job training, health, or education programs where it may be impractical to enforce treatment assignment, treatments delivered through phone or mail can often be easily targeted. Combining these two insights implies that OPOWER is a natural and promising setting for profiling. In the (2009) working paper version of this analysis, I develop a statistical treatment rule under which a decision maker allocates treatment conditional on observed characteristics to maximize program cost effectiveness while treating a given share of the population. This builds on an existing profiling literature, including Berger et al. (2000), Dehejia (2005), Graham et al. (2009), Hirano and Porter (2006), Imai and Strauss (2009), and Manski (2004, 2009). Using Connexus as an example experiment, I show that if the OPOWER program were to be administered to half of the eligible population, profiling would increase the ATE by 74% relative to arbitrary assignment, thereby reducing the cost per kilowatt-hour conserved by 43%.

4.1. Effects of normative categorizations

Recall that households that used less than the 20th percentile of their neighbor comparison group are labeled as "Great," those who use more than the mean are labeled "Below Average," and those in between are labeled "Good." Could injunctive norms moderate the heterogeneity in effects across households with different pre-treatment usage? And are they responsible for mitigating the potential descriptive norms boomerang effect?

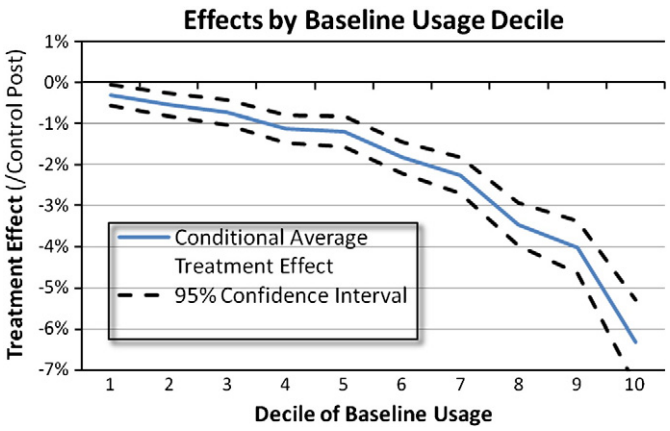

Fig. 8. Treatment effects by decile of baseline usage.

Indeed, one could imagine that these normative categorizations have different motivational effects. From a practical perspective, this would influence how OPOWER classifies households. For example, if being labeled "Great" appears to be more motivational than being labeled "Good," OPOWER might want to expand the "Great" category to include a broader part of the distribution. From a theoretical perspective, it is not obvious whether being labeled "Great" should induce people to conserve more, because they are motivated by positive feedback, or to conserve less, because they feel comfortable with their current performance. Similarly, being labeled "Below Average" could motivate people to improve or alternatively could discourage them from taking action.

4.1.1. Regression discontinuity design

Define as $t-$ the date that corresponds to the most recent Report generated in time to affect usage for meter read date t.[6] The "intuitive regression" that one might like to run would be to compare the treatment effect in bill t for households described as "Good" based on that recent Report $t-$ to the treatment effects for households that had been categorized as "Great" or "Below Average":

$$\begin{aligned} Y_{it} = {} & [\alpha_G \cdot 1(Great_{t-} + \alpha_B \cdot 1(Below\ Average_{t-}) + \tau_e] \cdot T_i P_{it} \\ & + [\beta_{Ge} \cdot 1(Great_{t-}) + \beta_{Be} \cdot 1(Below\ Average_{t-}) + \beta_e] \cdot P_{it} \\ & + \mu_{my} + \upsilon_i + \varepsilon_{it} \end{aligned} \tag{3}$$

The variables $1(Below\ Average_{t-})$ and $1(Great_{t-})$ are indicator variables for whether the household was categorized as "Below Average" or "Great" on Report $t-$. The coefficients $\hat{\alpha}_G$ and $\hat{\alpha}_B$ could naively be interpreted as the causal effects of these two categorizations relative to being categorized as "Good." Since other elements of the social comparisons and the knowledge effect differ across households in the different normative categorizations, however, $\hat{\alpha}_G$ and $\hat{\alpha}_B$ are not consistent estimators of these causal effects.

Instead of this specification, I therefore use a regression discontinuity estimator that exploits the fact that the injunctive categorizations were based on categories with sharp cutoffs. A group of households just below a cutoff are in the limit identical to a group just above, and thus their treatment effects from the descriptive norms and energy conservation tips are identical. Any differences in usage between households on either side of the cutoff are caused by their different injunctive categorizations.

To implement this, define c_{it-}^{20} and c_{it-}^{Mean} as the 20th percentile and mean cutoff points in the distribution of household i's neighbors' usage at time $t-$. Note that OPOWER has not varied these cutoff points across households or experiments. The variables D_{it-}^{20} and D_{it-}^{Mean} are the differences, in kilowatt-hours per month, between the household's energy use at time $t-$ and these 20th percentile and Mean cutoffs.

As suggested by Lee and Lemieux (2009) and Imbens and Wooldridge (2009), consider first a graphical analysis. Fig. 9 graphs electricity usage Y_{it} against forcing variable D_{it-} at the mean of households' comparison groups, which is the cutoff between being categorized as "Good" vs. "Below Average." The Usage variable on the y-axis, which as before is normalized as a percent of Control group

[6] More specifically, $t-$ is the date of the most recent Report generated more than 24 days and less than 55 days before meter read date t. If there is no Report in that period, the date of the most recent Report generated between 21 and 24 days before meter read date t is defined as t-. If there is no Report in that period, the most recent Report generated between 55 and 120 days before t is defined as t-. In the sporadic cases where there has been no Report generated for a household in the past 120 days, data point it is coded as not having an injunctive categorization.

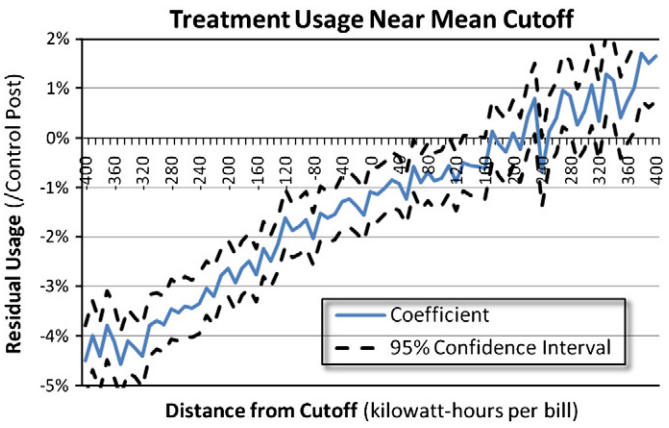

Fig. 9. Treatment group near mean comparison cutoff.

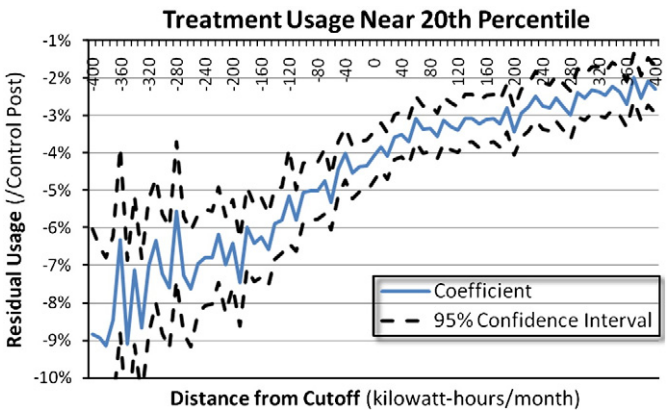

Fig. 10. Treatment group near 20th percentile cutoff.

post-treatment usage, is residual of month-by-year controls and household fixed effects. The line is upward sloping, as residual usage is serially autocorrelated: households that consume less compared to their peers on a given bill also tend to have lower residual usage on future bills. Fig. 10 is the analogous illustration near the 20th percentile of the household-specific comparison group, which is the cutoff between being categorized as "Great" vs. "Good."

The figures illustrate two key issues. First, the outcome variable Y_{it} is very close to linear in the forcing variable D_{it-}. This means that local linear regression, as suggested by Lee and Lemieux (2009) and Imbens and Wooldridge (2009), will be a natural specification. Second, there is no observable jump in the outcome at the cutoff. This suggests that we will estimate a zero effect, and the standard error – the "tightness" of that zero – will be the parameter of interest.

For the actual estimation, a separate regression is run for each cutoff using a rectangular kernel. This gives estimates of α_G, the effect of being "Good" instead of "Great," and α_B, the effect of being "Below Average" instead of "Good":

$$Y_{it} = 1(|D_{it-}|<h)\cdot 1(|D_{it-}|\neq 0)\cdot\left\{\begin{array}{c}\alpha\cdot 1(D_{it-}>0)+\beta_0\\ \beta_1 D_{it-}\cdot 1(D_{it-}<0)+\beta_2 D_{it-}\cdot 1(D_{it-}>0)\end{array}\right\} + \tau P_{it} + \mu_{my} + \upsilon_i + \varepsilon_{it}, \quad \forall i \text{ s.t.} T_i = 1 \tag{4}$$

The specifications again use household fixed effects, with standard errors clustered by household. While all observations are included in order to more precisely estimate month-by-year dummies and household fixed effects, each estimated α is identified only using observations with D_{it-} within bandwidth h of a cutoff.[7] Because of the large number of observations, regressions are run separately for each of the 17 experiments, and the set of estimated $\hat{\alpha}$ are combined using a minimum distance estimator weighted to minimize variance. Appendix A details the minimum distance estimator and cross-validation criterion function.

4.2. Results

Table 5 presents the regression discontinuity results around the Mean cutoff. The first row gives the results from the "intuitive regression" from Eq. (3). Conditional Average Treatment Effects are 1.08 percentage points larger for households recently categorized as "Good" instead of "Great," the two categories on either side of the 20th percentile of the neighbor comparison group distribution.

The lower portion of the table presents the RD estimates of α_B and values of the cross-validation criterion functions for a series of different bandwidths. The qualitative result of no statistically significant effect is robust even for very large bandwidths, although the standard errors drop somewhat as larger bandwidths admit more observations. Using the standard errors at bandwidth $h=300$, differential effects of larger than 0.16 percentage points in absolute

Table 5
RD results at mean cutoff.

Eq. (3)				$\hat{\alpha}$
				−1.08 (0.11)***
RD design				
h (kwh/month)	$\hat{\alpha}$	N	$CV_Y(h,5)$	$CV_Y(h,10)$
5	−0.54 (0.66)	29,436	3001.8	8067.1
10	−0.45 (0.47)	58,768	1708.2	1994.8
25	−0.33 (0.30)	145,744	1568.9	1636.7
50	0.18 (0.21)	289,771	1529.3	1610.5
100	0.12 (0.15)	573,089	1541.9	1603.2
200	0.08 (0.11)	1,092,428	1539.4	1605.8
300	0.09 (0.10)	1,521,405	1534.5	1607.1
400	0.1 (0.09)	1,847,840	1535.7	1608.8

Table 6
RD Results at 20th Percentile.

Eq. (3)				$\hat{\alpha}$
				−0.70 (0.14)
RD Design				
h (kwh/month)	$\hat{\alpha}$	N	$CV_Y(h,5)$	$CV_Y(h,10)$
5	−1.3 (0.90)	19,675	122,191	21,992
10	−0.45 (0.60)	40,272	49,237	5238
25	−0.06 (0.36)	102,736	3956	2191
50	0.2 (0.25)	206,479	2168.5	2089.8
100	0.09 (0.18)	406,637	2165.9	2068.1
200	0.01 (0.13)	778,548	2176.2	2067.7
300	−0.02 (0.12)	1,109,442	2173.8	2077.3
400	−0.12 (0.12)	1,400,046	2186.6	2040.1

Standard errors in parentheses. Dependent variable is the household's average daily electricity consumption (kilowatt-hours), normalized by average control group consumption in the Post period. All specifications include month-by-year dummies and household fixed effects. h refers to the bandwidth of the RD estimator; units are kilowatt-hours per month. The N column counts observations within the bandwidth. The $CV_Y(h,\pi)$ columns display the value of the cross-validation criterion functions at different π.

[7] Note also that the kernel excludes observations where $D_{it-}=0$. Some households are themselves the 20th percentile of their comparison group. OPOWER does not retain the distance between Y_{it-} and the next higher-consuming household's usage, meaning that the correct distance to their category cutoff cannot be calculated.

value can be ruled out with 90% confidence in a two-sided test. Dividing this by the $\hat{\alpha}_G$ from Eq. (3), no more than 15% of the larger effects for households labeled "Good" instead of "Great" are due to the different injunctive categorizations. These differences are instead likely driven by differential responses to descriptive norms or energy conservation information.

Table 6 presents the parallel results for the 20th percentile cutoff. Again using the Eq. (3), the treatment effects are 0.70 percentage points larger for households recently categorized as "Below Average" instead of "Good." At this 20th percentile cutoff, differential effects of more than 0.20 percentage points can be ruled out using the same test and bandwidth. No more than 28% of the differential effects for households labeled "Below Average" instead of "Good" are due to the different injunctive categorizations.

The results are highly robust. As Tables 5 and 6 show, the qualitative results are robust to any reasonable kernel bandwidth. The RD specifications above are restricted to Treatment group households only, but looking instead for discontinuities in the treatment effect around the cutoff increases the variance of the estimator but does not qualitatively change the results. Assuming the same slope above and below the cutoff, $\beta_1 = \beta_2$, has no statistically significant effect on the $\hat{\alpha}$, but this does not substantially reduce its standard error. As the graphical analysis suggests would be the case, however, omitting the linear terms $D_{it-} \cdot 1(D_{it-}<0)$ and $D_{it-} \cdot 1(D_{it-}>0)$ does affect the estimates at all but small bandwidths. Using a 4th-order polynomial in D_{it-} does not statistically affect $\hat{\alpha}$, largely because it substantially increases the standard errors. All of these additional specifications are omitted to conserve on space.

Put simply, the key result is that the different injunctive categorizations do not cause large differences in the treatment effects. While the injunctive norms may have some effect that is equally powerful across all households, being classified in one category or another compared to neighbors does not have large differential effects. Some combination of energy conservation tips plus injunctive norms that affect all three groups equally must be responsible for mitigating the descriptive norm boomerang effect in low-usage households.

5. Conclusion

This paper evaluates the effects of the OPOWER Home Energy Reports, which give households feedback on past energy consumption, compare them to their neighbors, and provide energy conservation tips. The program is a remarkable departure from traditional energy efficiency programs in that it is a non-price intervention designed with direct insight from behavioral science that is evaluated using randomized controlled trials. The perceived success or failure of these pilot programs will directly affect millions of dollars of future investment under new energy conservation regulations and climate policies and could more generally influence how future energy efficiency programs will be designed and evaluated.

I find that the Average Treatment Effects of OPOWER's programs range from 1.4 to 3.3% of baseline usage, with an unweighted mean ATE of 2.0%. I also show that treatment effects increase markedly as a function of pre-treatment usage, although not even the lowest-consumption households increase usage in response to the treatment. Thus, while the descriptive norm information by itself might have induced these households to increase usage, some other aspect of the treatment eliminated this potential "boomerang effect." A regression discontinuity analysis, however, shows that being assigned to a different injunctive norm category does not significantly change the treatment effect for households near the category cutoffs. Instead, the potential descriptive norm boomerang effect must be mitigated by energy conservation tips or by aspects of the injunctive norms that affect households in the different categories by similar amounts.

This analysis adds to growing appreciation of how non-price interventions can affect consumer behavior. Economists in general, and energy sector policymakers in particular, have historically focused on how prices and subsidies affect demand. This experiment shows that simply sending letters – a treatment that has no effect on relative prices and may have limited effects on information sets – can persistently affect usage by as much as a 11 to 20% short run price increase or a 5% long run increase. While the welfare effects are indeterminate and merit additional thought, OPOWER's programs are relatively cost effective from the program administrator's perspective. As climate change policies are implemented and utility regulation increasingly encourages energy conservation, such non-price treatments may receive greater attention, and insights from behavioral science may be increasingly taken to scale.

Acknowledgements

I thank, without implicating, Ian Ayres, Bob Cialdini, Tyler Curtis, Rajeev Dehejia, Kenneth Gillingham, Larry Goulder, Michael Greenstone, Matt Harding, Kosuke Imai, Seema Jayachandran, Karthik Kalyanaraman, Alex Kaufman, Ogi Kavazovic, Alex Laskey, Aprajit Mahajan, Justin Marion, Sendhil Mullainathan, Dave Rapson, Todd Rogers, Eldar Shafir, Joe Shapiro, Lan Shi, Marc Solomon, Dmitry Taubinsky, two anonymous referees, and seminar participants at the Congressional Budget Office, the Environmental Defense Fund, Harvard, the National Tax Association Annual Meetings, Stanford, and the University of Wisconsin for helpful conversations and feedback on this project.

Appendix A. Regression discontinuity details

This appendix provides detail on the minimum distance estimator used to combine RD estimates from different utilities and on the cross-validation criterion function. Because of the large number of observations, regressions are run separately for each experiment e, and the set of estimated $\hat{\alpha}_e$ are combined using a minimum distance estimator:

$$\hat{\alpha} = \sum_e w_e \hat{\alpha}_e \tag{5}$$

$$\hat{V}_\alpha = \sum_e w_e^2 \hat{V}_{\alpha e} \tag{6}$$

The weights w_e applied to the estimates from each experiment are chosen to minimize $\hat{V}_\alpha$:

$$w_e = \frac{1}{\hat{V}_{\alpha e}} \cdot \left(\sum_f \hat{V}_{-1}^{\alpha_f} \right)^{-1} \tag{7}$$

What remains is to choose the bandwidth h. Because of the very large sample size, it would be extremely computationally intensive to carry out the traditional "leave-one-out" cross validation approach introduced in Ludwig and Miller (2005) and detailed in Imbens and Lemieux (2007). Instead, I place each $it-$ observation within each experiment as a percentile in the experiment's distribution of D_{it-} and generate two placebo cutoffs that are π percentiles on either side of the one true cutoff c_{it-}. Eq. (4) is then run for each experiment, replacing $1(|D_{it-}|<h)$ with an indicator for whether D_{it-} is less than h below the lower placebo cutoff or h above the higher placebo cutoff. This regression generates predictions $\hat{Y}_{it}(h,\pi)$ for the set of observations $\mathcal{S}$ between the two placebo cutoffs. Note that the fitted values for observations above (below) the true cutoff are fitted only with the regression line above (below) the placebo cutoff. The cross-validation

criterion is the mean across observations in all experiments of the squared prediction errors for $\mathcal{S}$:

$$CV_Y(h,\pi) = \frac{\sum_{it} 1(it \in \mathcal{S})\left(Y_{it} - \hat{Y}_{it}(h,\pi)\right)^2}{\sum_{it} 1(it \in \mathcal{S})} \quad (8)$$

The rightmost three columns of Tables 5 and 6 show the number of observations within the bandwidth and the values of the cross-validation criterion function for $\pi = 5$ and $\pi = 10$. The shapes of these functions are comparable for other values of π. As Figs. 9 and 10 show, there is a linear relationship between the forcing and outcome variables over a wide interval around the cutoffs. As a result, once there is enough data to estimate the β_1 and β_2 controls, the cross-validation criterion function is very flat for a wide range of bandwidths. For example, $CV_Y(h,\pi = 5)$ drops from 3001.8 to 1529.3 as h is increased from 5 to 50 kWh per month, but it increases by less than 1% as h increases to 400. For values of h larger than that, slight non-linearities in the relationship between the forcing and outcome variables begin to bias the estimated α. To compute the "tightness" of the zero in Section 4, I choose $h = 300$, which gives relatively small standard errors but is still well within the range where the linearity assumption holds.

References

Abrahamse, Wokje, Linda, Steg, Charles, Vlek, Talib, Rothengatter, 2005. A Review of Intervention Studies Aimed at Household Energy Conservation. Journal of Environmental Psychology 25 (3), 273–291 (September).

ACEEE, 2010. State Energy Efficiency Resource Standards. http://www.aceee.org/files/pdf/State%20EERS%20Summary%20Aug%202010.pdf 2010.

Allcott, Hunt, 2009. Attention, Social Norms, and Energy Conservation. MIT Center for Energy and Environmental Policy Research Working Paper 09–014 (September).

Allcott, Hunt, Mullainathan, Sendhil, 2010b. External Validity and Partner Selection Bias. Working Paper, MIT Department of Economics (June).

Alpizar, Francisco, Carlsson, Fredrik, Johansson-Stenman, Olof, 2008. Anonymity, Reciprocity, and Conformity: Evidence from Voluntary Contributions to a National Park in Costa Rica. Journal of Public Economics 92, 1047–1060.

Arimura, Toshi, Newell, Richard, Palmer, Karen, 2009. Cost-Effectiveness of Electricity Energy Efficiency Programs. Resources for the Future Discussion Paper 09–48 (November).

Ayres, Ian, Raseman, Sophie, Shih, Alice, 2009. Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage. NBER Working Paper 15386 (September).

Becker, Gary, 1965. A Theory on the Allocation of Time. Economic Journal 75, 493–517.

Berger, Mark, Black, Dan, Smith, Jeffrey, 2000. Evaluating Profiling as a Means of Allocating Government Services. Working Paper, Syracuse University (September).

Bertrand, Marianne, Duflo, Esther, Mullainathan, Sendhil, 2004. How Much Should We Trust Difference-in-Differences Estimates? Quarterly Journal of Economics 119 (1), 249–275.

Bertrand, Marianne, Karlan, Dean, Mullainathan, Sendhil, Shafir, Eldar, and Zinman, Jonathan, 2010. What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment. Quarterly Journal of Economics 125 (1), 263–306.

Beshears, John, James, Choi, David, Laibson, Brigitte, Madrian, Katherine, Milkman, 2009. The Effect of Providing Peer Information on Retirement Savings Decisions. Working Paper, Harvard University (March).

Cai, Hongbin, Chen, Yuyu, Fang, Hanming, 2009. Observational Learning: Evidence from a Randomized Natural Field Experiment. American Economic Review 99 (3), 864–882 (June).

Cialdini, Robert, Reno, Raymond, Kallgren, Carl, 1990. A Focus Theory of Normative Conduct: Recycling the Concept of Norms to Reduce Littering in Public Places. Journal of Personality and Social Psychology 58, 1015–1026.

Clee, Mona, Wicklund, Robert, 1980. Consumer Behavior and Psychological Reactance. Journal of Consumer Research 6, 389–405.

Conley, Timothy and Udry, Christopher, 2010. Learning About a New Technology: Pineapple in Ghana. American Economic Review 100 (1), 35–69.

Costa, Dora, Kahn, Matthew, 2010. Energy Conservation Nudges and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment. NBER Working Paper No. 15939 (April).

Darby, Sarah, 2006. The Effectiveness of Feedback on Energy Consumption. Working Paper, Oxford Environmental Change Institute (April).

Davis, Lucas, 2008. Durable Goods and Residential Demand for Energy and Water: Evidence from a Field Trial. RAND Journal of Economics 39 (2), 530–546 (Summer).

Dehejia, Rajeev, 2005. Program Evaluation as a Decision Problem. Journal of Econometrics 125, 141–173.

DellaVigna, Stefano, List, John, Malmendier, Ulrike, 2010. Testing for Altruism and Social Pressure in Charitable Giving. Working Paper, University of Chicago (June).

Dubin, Jeffrey, McFadden, Daniel, 1984. An Econometric Analysis of Residential Electric Appliance Holdings and Consumption. Econometrica 52 (2), 345–362 (March).

Ferraro, Paul, Price, Michael, 2010. Using Non-Pecuniary Strategies to Influence Behavior: Evidence from a Large-Scale Field Experiment. Working Paper, Georgia State University (April).

Firpo, Sergio, 2007. Efficient Semiparametric Estimation of Quantile Treatment Effects. Econometrica 75 (1), 259–276 (January).

Fischbacher, Urs, Gachter, Simon, Fehr, Ernst, 2001. Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. Economic Letters 71, 397–404.

Foster, Andrew, Rosenzweig, Mark, 1995. Learning by Doing and Learning from Others: Human Capital and Technical Change in Agriculture. Journal of Political Economy 103 (6), 1176–1209 (December).

Frey, Bruno, Meier, Stephan, 2004. Social Comparisons and Pro-Social Behavior: Testing 'Conditional Cooperation' in a Field Experiment. American Economic Review 94 (5), 1717–1722 (December).

Friedrich, Katherine, Eldridge, Maggie, York, Dan, Witte, Patti, Kushler, Marty, 2009. Saving Energy Cost-Effectively: A National Review of the Cost of Energy Saved through Utility-Sector Energy Efficiency Programs. ACEEE Report No. U092 (September).

Froelich, Markus, and Melly, Blaise, 2010. Estimation of quantile treatment effects with STATA. Stata Journal 10 (3), 423–457.

Gerber, Alan, Rogers, Todd, 2009. Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You. Journal of Politics 71, 1–14.

Gillingham, Kenneth, Newell, Richard, Palmer, Karen, 2006. Energy Efficiency Policies: A Retrospective Examination. Annual Review of Environment and Resources 31, 161–192.

Graham, Bryan, Imbens, Guido, Ridder, Geert, 2009. Complementarity and Aggregate Implications of Assortative Matching: A Nonparametric Analysis. NBER Working Paper 14860 (April).

Hirano, Keisuke, Porter, Jack, 2006. Asymptotics for Statistical Treatment Rules. Working Paper, University of Wisconsin (August).

Imai, Kosuke, Aaron, Strauss, 2009. Planning the Optimal Get-out-the-vote Campaign Using Randomized Field Experiments. Working Paper, Princeton University (May).

Imbens, G.W., Lemieux, T., 2007. Regression discontinuity designs: A guide to practice, Journal of Econometrics. doi:10.1016/j.jeconom.2007.05.001.

Imbens, Guido, Wooldridge, Jeffrey, 2009. Recent Developments in the Econometrics of Program Evaluation. Journal of Economic Literature 47 (1), 5–86 (March).

Lalonde, Robert, 1986. Evaluating the Econometric Evaluations of Training Programs. American Economic Review 76 (4), 604–620 (September).

Lee, David, Lemieux, Thomas, 2009. Regression Discontinuity Designs in Economics. NBER Working Paper 14723 (February).

Levitt, Steven, List, John, 2007. What Do Field Experiments Measuring Social Preferences Reveal about the Real World? Journal of Economic Perspectives 21 (2), 153–174 (Spring).

Ludwig, Jens, Miller, Douglas, 2005. Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. NBER Working Paper No. 11702 (October).

Manski, Charles, 2004. Statistical Treatment Rules for Heterogeneous Populations. Econometrica 72 (4), 1221–1246 (July).

Manski, Charles, 2009. Diversified Treatment Under Ambiguity. Working Paper, Northwestern University.

Mobius, Markus, Niehaus, Paul, Rosenblat, Tanya, 2005. Social Learning and Consumer Demand. Working Paper, Harvard University (December).

Munshi, Kaivan, 2004. Social Learning in a Heterogeneous Population: Technology Diffusion in the Indian Green Revolution. Journal of Development Economics 73 (1), 185–215 (February).

Munshi, Kaivan, Myaux, Jacques, 2006. Social Norms and the Fertility Transition. Journal of Development Economics 80 (1), 1–38 (June).

Nadel, Steven, Keating, Kenneth, 1991. Engineering Estimates versus Impact Evaluation Results: How Do They Compare and Why? Energy Program Evaluation: Uses, Methods, and Results: Proceedings of the 1991 International Energy Program Evaluation Conference.

Nolan, Jessica, Schultz, Wesley, Cialdini, Robert, Goldstein, Noah, Griskevicius, Vladas, 2008. Normative Influence is Underdetected. Personality and Social Psychology Bulletin 34, 913–923.

Reiss, Peter, White, Matthew, 2005. Household Electricity Demand, Revisited. Review of Economic Studies 72 (3), 853–883 (July).

Reiss, Peter, White, Matthew, 2008. What Changes Energy Consumption? Prices and Public Pressure. RAND Journal of Economics 39 (3), 636–663 (Autumn).

Rubin, Donald, 1974. Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies. Journal of Educational Psychology 66 (5), 688–701.

Schultz, Wesley, Nolan, Jessica, Cialdini, Robert, Goldstein, Noah, Griskevicius, Vladas, 2007. The Constructive, Destructive, and Reconstructive Power of Social Norms. Psychological Science 18, 429–434.

Shang, Jen, Croson, Rachel, 2004. Field Experiments in Charitable Contribution: The Impact of Social Influence on the Voluntary Provision of Public Goods. Working Paper, University of Pennsylvania.

Stern, Paul, 1992. What Psychology Knows about Energy Conservation. American Psychologist 47 (10), 1224–1232.

US Energy Information Administration, 2001. Residential Energy Consumption Survey. http://www.eia.doe.gov/emeu/recs/recs2001 2001.

US Energy Information Administration, 2009. Energy Market and Economic Impacts of H.R. 2454, the American Clean Energy and Security Act of 2009. http://www.eia.doe.gov/oiaf/servicerpt/hr2454/pdf/sroiaf%282009%2905.pdf 2009.

Violette, Daniel, Bill, Provencher, Klos, Mary, 2009. Impact Evaluation of Positive Energy SMUD Pilot Study. Summit Blue Consulting, Boulder, CO.

FURTHER READING

Allcott, Hunt, Mullainathan, Sendhil, 2010a. Behavior and Energy Policy. Science 327 (5970) (March 5th).
Andreoni, James, 1989. Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence. Journal of Political Economy 97 (6), 1447–1458 (December).
Becker, Gary, 1974. A Theory of Social Interactions. Journal of Political Economy 82 (6), 1063–1093 (November).
Bernheim, Douglas, 1994. A Theory of Conformity. Journal of Political Economy 102 (5), 847–877 (October).
Cialdini, Robert, 2003. Crafting Normative Messages to Protect the Environment. Current Directions in Psychological Science 12, 105–109.
Duflo, Esther, Saez, Emmanuel, 2003. The Role of Information and Social Interactions in Retirement Plan Decisions: Evidence from a Randomized Experiment. Quarterly Journal of Economics 118 (3), 815–842 (August).
Imbens, Guido, Kalyanaraman, Karthik, 2009. Optimal Bandwidth Choice for the Regression Discontinuity Estimator. IZA Discussion paper No. 3995 (February).