Social Pressure and Self-Selection in Experimental Policy Evaluation

Alec Brandon*
University of Chicago

November 24, 2019
Click here for latest version

Abstract

A leading technique uses randomized experimentation to evaluate the impact of policy interventions. Applications of this technique often measure the impact of an intervention on a self-selected sample. In this paper, I propose that social pressure from the experimenter is a determinant of these selection decisions. I test this hypothesis in the context of a study on a free LED lighting program. Recruitment takes place door-to-door in the suburbs of Chicago. A day before their recruitment visit, each household is informed about the study with a flyer on their doorknob. I isolate the role of social pressure by varying whether households can select out of the study by checking an opt-out box on their flyer. My main finding is that the LED lighting program causes a 15 percent reduction in evening energy use in the sample recruited with an opt-out flyer, whereas no energy savings are observed amongst households recruited with a baseline flyer. Social pressure appears to be the driver of these disparate effects, as the opt-out flyer causes a 13 percent reduction in households answering their door and, conditional on answering, causes a 52 percent increase in selection into the study. These results have important implications for the way evaluation experiments with self-selected samples are conducted, reported, and modeled.

*I thank Leonardo Bursztyn, Steven Levitt, and, John List for their support with this project. This paper also benefited from conversations with Dan Alexander, Mirthe Boomsma, Luigi Butera, Chris Clapp, Jonathan Davis, Michael Dinerstein, Justin Holz, Jeff Livingston, Fatemeh Momeni, David Novgorodsky, Gautam Rao, Matthias Rodemeier, Sally Sadoff, Joe Seidel, Karen Ye, and many others. I thank Diego Ruiz for providing excellent research assistance. Rob Zakim and the UtilityAPI team provided invaluable support throughout. The design and analysis of this experiment was pre-registered as trial 4476 on the American Economic Association RCT Registry, available at: https://doi.org/10.1257/rct.4476-2.0.

1. Introduction

Over the past half-century, the experimental approach to policy evaluation has gained considerable prominence in economics (Orcutt and Orcutt, 1968; Ashenfelter, 1987; Duflo et al., 2007). The defining feature of this approach is the random assignment of observational units to competing policy regimes (Burtless and Orr, 1986). In the simplest case, an experimental sample is randomly assigned to a treatment or control group. The treatment group is granted access to the policy intervention being evaluated, while access is withheld from the control group. Once a sufficient amount of time has passed, outcomes are observed and the impact of the intervention is evaluated by comparing outcomes across the treatment and control groups.

A longstanding concern with this research design is that the composition of experimental samples might be determined, in part, by the procedures used to conduct the experiments.[1] When the response to an intervention is heterogeneous, this concern is especially important because the measured impact could be confounded by the very procedures used to obtain a sample and make that measurement. While this concern has received a great deal of attention (see, e.g., Heckman, 1992; Heckman and Smith, 1995; Harrison and List, 2004; Banerjee and Duflo, 2009; Allcott, 2015; Deaton and Cartwright, 2018), the literature has yet to settle on specific conventions or best practices, with evaluation experiments employing a great diversity of procedures in the service of obtaining an experimental sample.

To substantiate this claim, I survey the experimental evaluation literature appearing in the top five economics journals between 2005 and 2019.[2] Across the 113 experimental evaluation papers I identify, there is a relative split in how samples are obtained, with 38 percent recruiting self-selected volunteers and the other 62 percent obtaining their sample at the behest of a site, such as an energy provider or school district.[3] Focusing further on the papers that use self-selected samples, Figure 1 highlights three salient trends. First, papers are nearly unanimous in their abstention from reporting recruitment documents, such as scripts. Second, the majority of papers do not discuss how their sample was recruited and when this is discussed, strategies are quite varied.[4] Third, of the papers reporting how recruitment is

[1]This concern is distinct from others that focus on the role of experimental procedures once an experimental sample has been obtained. For example, sample attrition (Hausman and Wise, 1979) and compliance in the treatment group (Bloom, 1984) have to do with experimental samples that are already in place. Similarly, Hawthorne and John Henry effects refer to the reaction a sample might have to having their outcomes observed or being assigned to the control group, respectively (Peters et al., 2016).

[2]These journals are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

[3]Occasionally an experimental sample is both site and self-selected. For example, permission might be required from a site to then recruit a self-selected sample. I classify these papers as self-selected.

[4]In cases where a paper reported both in person and remote recruitment, I code the paper as utilizing in person

framed, there is an even split between research study and program lottery framings.

The goal of this paper is to examine more closely the ways in which these disparate recruitment procedures might influence the composition of experimental samples and the measured impact of policies. Focusing on the case where selection is voluntary, I advance the following hypothesis. Recruitment procedures that scrutinize the decision to select out of an experimental sample induce a subsample that is otherwise uninterested in the experiment to participate. Under this hypothesis, high pressure recruitment strategies, such as recruiting in person, amplify the social cost of declining a request to participate (Levitt and List, 2007). When interest in an experiment is also heightened by the anticipated response to the intervention (Roy, 1951), this hypothesis has a clear prediction for the confounding of experimentally measured policy impacts: Recruitment procedures that socially pressure self-selection attenuate the measured impact of policy interventions.

I test this hypothesis in the context of an experimental evaluation of a residential LED lighting program.[5] Recruitment for the experiment takes place door-to-door amongst 4,888 households in the suburbs of Chicago. Households selecting into the experimental sample are randomly granted or denied the program, which is a pack of eight free LED light bulbs.[6] A day before their recruitment visit, households are informed about the program and the procedures pertaining to its receipt with a flyer placed on their doorknob. I isolate the effect of social pressure by varying whether households can select out of the experimental sample by checking an opt-out box on their flyer (DellaVigna et al., 2012).[7] The presence of this opt-out box allows households uninterested in the program to ensure selection out of the experiment without the social scrutiny of research staff influencing their decision.[8]

This design allows for a simple test of social pressure in experimental evaluation. If social pressure is a driver of self-selection, then, relative to the baseline flyer, the opt-out flyer should reduce the frequency of households answering their door. As a result, the rate at which households select into the experimental sample should also be diminished

recruitment because research from other settings finds the highest response rates for in person recruitment (Rolnick et al., 1989; Mannesto and Loomis, 1991; Maguire, 2009).

[5] This program is based a real policy intervention recently launched by the Los Angeles Department of Water and Power (LADWP, 2019).

[6] In the energy efficiency literature, this research design is referred to as "recruit and deny" (Gandhi et al., 2016). Recent energy efficiency experiments employing "recruit and deny" designs include Faruqui et al. (2014); Jessoe and Rapson (2014); Harding and Lamarche (2016); Bager and Mundaca (2017); Fowlie et al. (2017); Ito et al. (2018); Brandon et al. (2019).

[7] DellaVigna et al. (2012, 2017); Giaccherini et al. (2019) employ this strategy to experiment with social pressure in the context of charitable giving, voter turnout, and energy efficient technology adoption, respectively. A related strand of research finds that inducing similar types of variation in selection to laboratory experiments (Eckel and Grossman, 2000; Dana et al., 2006; Broberg et al., 2007; Lazear et al., 2012) and charitable solicitations (Andreoni et al., 2017) can also have a profound influence on outcomes.

[8] Additionally, a \$10 financial incentive is varied for participation in the experiment, which was announced on the flyers.

Figure 1: Recruitment in Evaluation Experiments with Self-Selected Samples—2005-2019

Note: This figure presents the relative frequency of recruitment reporting conventions and methods for experimental evaluation papers published in the top five economics journals between 2005 and 2019. These journals are the *American Economic Review, Econometrica,* the *Journal of Political Economy,* the *Quarterly Journal of Economics,* and the *Review of Economic Studies*.

by the opt-out box. However, the opt-out box should draw a more motivated sample to their door, causing an increase in the participation rate amongst households who answer. Finally, if, in addition to social pressure driving self-selection, households also select into the experimental sample on the basis of their anticipated response to the intervention, the impact of the LED lighting program on household energy savings should amplified by recruitment with the opt-out flyer.

I report four main findings. First, the presence of the opt-out box reduces the frequency of households opening their door. The opt-out flyer reduces the proportion of households opening their door by 13 percent, relative to a rate of 26 percentage points for the baseline flyer.[9] Second, the opt-out flyer does not reduce likelihood of a household selecting into the experimental sample. Third, conditional on answering their door, the opt-out box increases the frequency of selection into the sample. Relative to a conditional participation rate of 6 percentage points for the baseline flyer, the opt-out flyer increases conditional participation by 51 percent. Fourth, the presence of the opt-out box causes a dramatic change in the energy savings caused by the LED lighting program. Amongst households recruited with an opt-out flyer, the lighting program causes a 15 percent reduction in energy use during (pre-registered) evening hours. No such energy savings are observed amongst households recruited with the baseline flyer.

The remainder of this paper is organized as follows. In Section 2 I present a simple theoretical framework to motivate my analysis. Then, in Section 3 I describe the design of the study in greater detail. Section 4 presents the empirical analysis and Section 5 briefly concludes.

2. Theory

In this section, I start by reviewing the research design of an evaluation experiment that relies on a self-selected sample. I then propose a simple model of social pressure in these self-selection decisions. Finally, I use this model is to derive the testable predictions assessed later in this paper.

2.1 Experimental Evaluation Research Design with Self-Selected Samples

Consider an experimenter who wishes to evaluate the impact of a policy intervention on an outcome of interest. To conduct their study, the experimenter executes the following research

[9]This calculation includes the households that check the opt-out box, which occurs at a rate of 13 percentage points amongst households with the option.

design. First, they recruit households from a relevant population. Let $D_i = 1$ indicate that household i is selecting into the experimenter's study and $D_i = 0$ denote selecting out. Second, the experimenter randomly assigns a treatment across the households in the study. This treatment is designed to mimic (or even exaggerate) the goods and services provided by the policy intervention being studied. If household i selects into the study then $Z_i = 1$ denotes they have been randomly assigned to receive the treatment and $Z_i = 0$ indicates nonreceipt. Third, the experimenter collects information on the outcome of interest for the households who selected into the study. Let Y_i reflect the realization of this outcome for household i.

This research design allows the experimenter to estimate the average effect of the treatment on households selecting into the study. To see this, let Y_{1i} denote the potential treated outcome for household i and Y_{0i} their potential untreated outcome.[10] Then the effect of the treatment on household i, R_i, is the difference between their potential treated and untreated outcomes, $Y_{1i} - Y_{0i}$. I denote this effect with R_i because from the perspective of the household it is a return provided by the treatment. The average of this return across the households selecting into the study is,

$$SATE = E[R_i|D_i = 1] = E[Y_i|D_i = 1, Z_i = 1] - E[Y_i|D_i = 1, Z_i = 0], \tag{1}$$

where the first equality defines the experimenter's parameter of interest and the second links that parameter to the quantities estimated by the experimenter. I refer to this parameter as the $SATE$, which stands for the selected average treatment effect (Angrist and Imbens, 1991).[11]

2.2 Household Selection Under Social Pressure

In response to the experimenter, a household decides whether to select into the study. Suppose the experimenter recruits by approaching each household in the relevant population. If household i is home, $H_i = 1$, they answer their door and make the selection decision that maximizes their utility. As the decision is binary, this can be accomplished with the

[10] The relationship between these potential outcomes and the realized outcome, Y_i, is described by,

$$Y_i = Z_i Y_{1i} + (1 - Z_i) Y_{0i},$$

where Z_i corresponds to random assignment by the experimenter.

[11] When selection into the study approximates the real-world decision to select into treatment, the $SATE$ is also the average treatment effect on the treated, ATT (Heckman, 1992; Imbens and Angrist, 1994; Heckman and Vytlacil, 2007b).

following decision rule,

$$D_i = 1(U_i \geq 0), \tag{2}$$

where U_i is the net utility from selecting into the study.[12] For ease of exposition, I assume U_i is independent of H_i.[13]

Suppose the net utility a household obtains from selecting into the study is influenced by two channels. First, the effect of the treatment, R_i. I assume households correctly anticipate their effect, or return, upon being informed of the treatment. From the perspective of the experimenter, this return is distributed across households according to an unknown cumulative distribution function F, which has a corresponding probability distribution function f. Second, the social pressure, S, to please the experimenter and select into their study. I distinguish social pressure from other types of social preferences (e.g. Becker, 1974; Andreoni, 1989, 1990) by assuming the pressure a household experiences depends on the extent to which the experimenter scrutinizes their selection decision, with greater scrutiny creating additional pressure.[14] Furthermore, I assume social pressure is chosen by the experimenter and, as a result, is not indexed by i.

These two channels influence net utility through,

$$U_i = v(R_i, S), \tag{3}$$

where $v(\cdot)$ is a continuous and twice differentiable function. Relative to modeling frameworks in the evaluation literature, this formulation of net utility is closest to the Roy model and its extensions (Roy, 1951; Heckman and Vytlacil, 2007a).[15] A widely invoked modeling assumption has selection driven by gains, suggesting $v_R > 0$. In the context of medical trials, Malani (2008) reports evidence consistent with this assumption. With regards to social pressure, there is a great deal of evidence suggesting experimenter scrutiny increases the likelihood of observing socially desirable choices, suggesting $v_S > 0$ (see Levitt and List, 2007, for an overview of this evidence). While there is less evidence to guide assumptions

[12]The indicator function in equation (2) is defined as follows. For any real-valued scalar q,

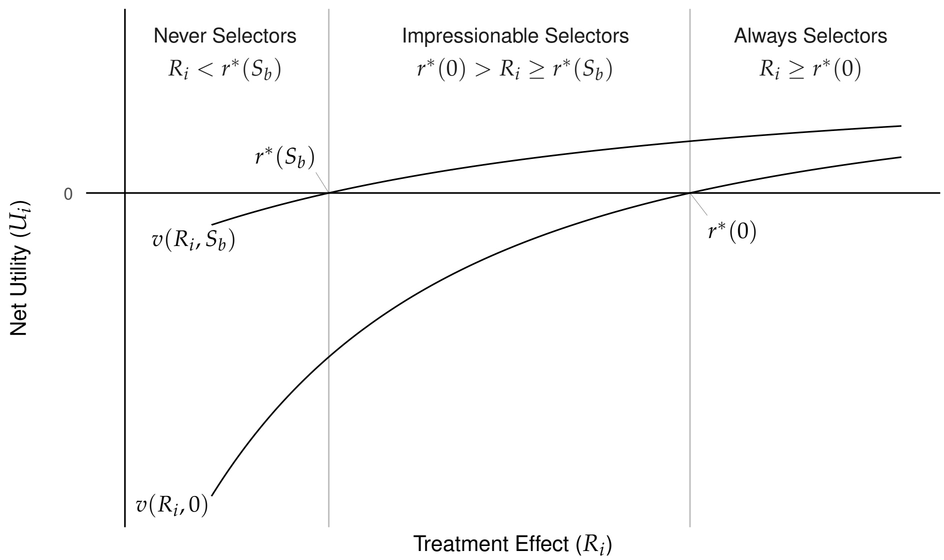
$$1(q \geq 0) = \begin{cases} 1, & q \geq 0 \\ 0, & q < 0. \end{cases}$$

[13]See DellaVigna et al. (2012, 2017) for frameworks that endogenize H_i in response to visits by a fundraiser and surveyor, respectively.

[14]This formulation of social pressure is broadly consistent with the predictions from models of status (Frank, 1985), conformity (Bernheim, 1994), social distance (Akerlof, 1997), identity (Akerlof and Kranton, 2000), social image (Benabou and Tirole, 2006), and moral costs (Levitt and List, 2007).

[15]In particular, the extended Roy model in Heckman and Vytlacil (2007a), which has $U_i = R_i - c$, where c is common across households.

Figure 2: Household Net Utility from Selecting into Study Under Social Pressure Model



Note: This figure plots net utility as a function of the treatment effect, R_i, for two levels of social pressure, S_b and 0, $S_b > 0$. The treatment effect at which net utility is zero characterizes a threshold, $r^*(S)$. Treatment effects above this threshold prompt a household to select into the study.

on the curvature of v, standard intuition suggests the greater the return, the less influential social pressure will be, $v_{RS} \leq 0$, and diminishing marginal net utility in its two inputs, $v_{RR} \leq 0$ and $v_{SS} \leq 0$.

To see how these two channel's influence a household's selection decision, it is convenient to characterize a threshold decision rule. Let $r^*(S)$ denote a return above which a household will select into the study. This threshold is defined as the solution to,

$$v(r^*(S), S) = 0. \tag{4}$$

The objective of this paper is to investigate how changes in the social pressure to self-select influence the composition of experimental samples. Differentiating equation 4 with respect to S and rearranging reveals a clear prediction with respect to this threshold decision rule. Increasing social pressure reduces the treatment effect required to convince a household to select into a study,

$$\frac{\partial r^*(S)}{\partial S} = -\frac{v_S}{v_R} < 0,$$

where the inequality follows from $v_S > 0$ and $v_R > 0$.

Figure 2 illustrates how $r^*(S)$ is characterized and how it responds to two different levels of social pressure. Let $S = S_b > 0$ denote a baseline level of social pressure that is compared to a social pressureless setting where $S = 0$. As the figure highlights, reducing social pressure from S_b to 0 shifts the net-utility from a given treatment effect down, reducing the treatment effect required for a household to select into the study. Figure 2 also illustrates how two levels of social pressure characterizes three types of households. The first type experiences a treatment effect that doesn't justify selecting into the study at either level of social pressure, $R_i < r^*(S_b)$. This type of household is deemed a Never Selector. The second type of household is marginal. If left to their own devices they would not select into the study, but if placed under social pressure by the experimenter, they will feel compelled to self-select, $r^*(0) \geq R_i > r^*(S_b)$. I call these households Impressionable Selectors. The third and final type of household experiences such a large treatment effect that they will always select into the study, $R_i \geq r^*(S_b)$. I refer to these households as Always Selectors.

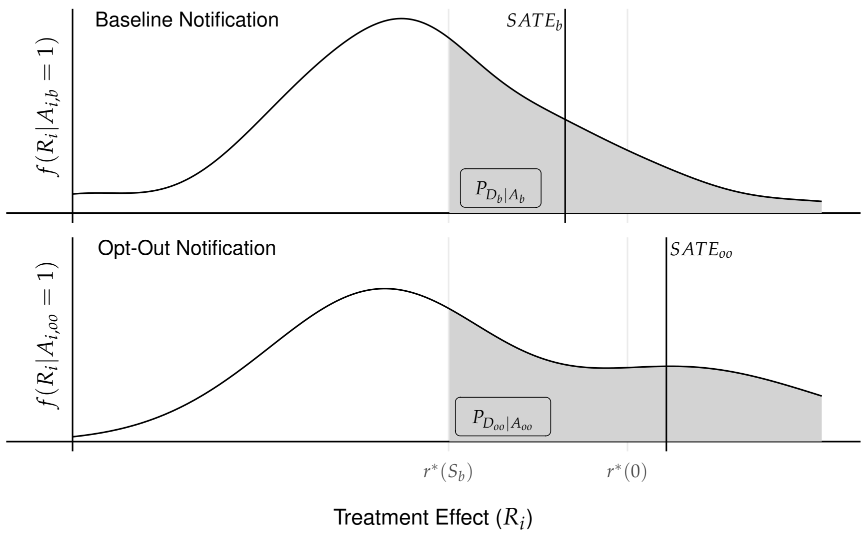
2.3 Testable Predictions

To determine falsifiable predictions for the self-selection model just presented, I place additional structure on the recruitment process employed by the experimenter. A day before their visit, suppose the experimenter notifies households about the study and treatment, with households observing this notification with strictly positive probability, $\rho \in (0, 1]$. Suppose there are two types of notifications. I deem the first a baseline notification and denote this condition as b. When a household receiving this notification answers their door, the level of social pressure they experience corresponds to $S = S_b$ from Figure 4. The second notification is identical to b except that it allows the household to opt-out of receiving the experimenter's recruitment visit. I refer to this notification by oo and assume a household opting out of the visit experiences no social pressure, which corresponds to $S = 0$ from Figure 4. In response to these two types of notifications, the self-selection model yields four testable predictions.[16]

First, relative to the baseline notification, the frequency of households answering their door is predicted to decrease in response to the opt-out notification. To see this prediction, recall that in response to the baseline notification, households answer their door when they are home, giving us $P_{A_b} = Pr(H_i = 1)$. In the opt-out condition, when the notification is observed, the Impressionable and Never Selectors will select out of the study by opting out

[16] Throughout this discussion, I assume the Always, Impressionable, and Never Selector sets are each non-empty.

Figure 3: Selection and Selected Average Treatment Effect Under Social Pressure Model



Note: This figure illustrates the probability density of the treatment effect for households answering their door, $A_{i,j}=1$, in response to the two notifications, $j \in \{b, oo\}$. With the baseline notification, a random subsample of the population answers their door and all households with a treatment effect above $r^*(S_b)$ select into the study causing the experimenter to measure a program impact of $SATE_b$. For the opt-out notification, households that observe the notification and have a treatment effect below $r^*(0)$ select out before answering their door. Households answering their door still select in if their effect is above $r^*(S_b)$, but the subsample opting out leads to a treatment effect distribution with a fatter right tail and a bigger program impact, $SATE_{oo}$.

in the pressureless privacy of their home instead of risk being pressured into the study or having to bear the cost of disappointing the experimenter in person. As a result, the share of households answering their door is $P_{A_{oo}} = \rho Pr(R_i \geq r^*(0))Pr(H_i = 1) + (1-\rho)P_{A_b}$, which is strictly less than P_{A_b}.

Second, compared to the baseline notification, the opt-out notification is predicted to reduce the share of households selecting into the study. In response to the baseline notification, both the Always and Impressionable Selectors participate when home, giving us, $P_{D_b} = Pr(R_i \geq r^*(S_b))Pr(H_i = 1)$. However, when the Impressionable Selectors observe a opt-out notification, they opt out of the study and a smaller share of households select in, $P_{D_{oo}} = \rho Pr(R_i \geq r^*(0))Pr(H_i = 1) + (1-\rho)P_{D_b}$, which gives us $P_{D_b} > P_{D_{oo}}$.

Third, conditional on answering the door, the opt-out notification is predicted to increase the share of households selecting into the study. Figure 3 illustrates this prediction by simulating the probability density of the treatment effect for the households answering their door in response to each notification. For the baseline notification, the conditional rate is simply the unconditional rate unweighted by the share of households answering their door, $P_{D_b|A_b} = Pr(R_i \geq r^*(S_b))$. With the opt-out notification, the density of the treatment effect distribution is skewed by the households opting out, which inflates the odds that a household will participate when they answer their door, $P_{D_{oo}|A_{oo}} = \rho + (1-\rho)P_{D_b|A_b}$. Thus, $P_{D_{oo}|A_{oo}} > P_{D_b|A_b}$.

Fourth, the average effect of the treatment in response to the opt-out notification is predicted to exceed the average effect found with the baseline notification. This prediction is also illustrated in Figure 3, which plots the $SATE$ that would be observed in the simulated data. The intuition for this result goes as follows. With the baseline notification, both the Always and Impressionable Selectors participate in the study, whereas for the opt-out notification, the Impressionable Selections that observe the notification opt out. With a sample that is more likely to have been motivated by the effect of the treatment, the average effect observed in response to the opt-out notification will then exceed the average effect for the baseline notification. More formally, the $SATE$ under the baseline notification is $SATE_b = E[R_i|R_i \geq r^*(S_b)]$, whereas for with the opt-out notification the experimenter measures $SATE_{oo} = \rho E[R_i|R_i \geq r^*(0)] + (1-\rho)SATE_b$, which yields $SATE_{oo} > SATE_b$.

3. Design

To test the self-selection model developed above, I conduct a natural field experiment (Harrison and List, 2004) layered over an evaluation experiment on an energy efficiency program. In this section, I describe the details of the treatment, the population recruited for the evaluation experiment, the notifications used to test for social pressure, and the recruitment procedures.

3.1 Treatment

The treatment evaluated in this paper is a free LED lighting program. The program allocates eight free LED light bulbs to a household, each of which is equivalent to a traditional 60 watt incandescent bulb. The key feature of the treatment is that each bulb uses substantially less electricity to provide light than incumbent technologies. While using a 60 watt incandescent light bulb for an hour consumes 60 watts of electricity and the compact fluorescent equivalent uses 15 watts, an LED equivalent consumes just 9 watts. The specific model used was chosen on the basis of a recommendations from a product testing company called Wirecutter.[17] On average, each pack cost $16.36 to acquire via the online retailer Amazon.com.

3.2 Population

The experimental sample was recruited from a population of 4,888 households in Evanston, Lincolnwood, Oak Park, River Forest, and Skokie, Illinois. These towns were chosen because of their proximity to the University of Chicago campus and their abundance of single family homes.[18] Single family homes in these towns were identified by scraping the Cook County Assessor's online database. This web scrape was conducted primarily for the purpose of locating large clusters of homes for recruitment, but it also allowed for the observation of several household characteristics.

Table 1 summarizes the characteristics of the population observed in the Cook County Assessor's database. Perhaps the most salient feature of the population is the age of the homes. Only 3 percent were built in the last half-century and nearly 40 percent were built more than a century ago. Beyond the old age of these homes, there is a good diversity across the other characteristics. In terms of house size, stories, number of bathrooms, central air conditioning, and status of basement there is good variation. Additionally, the vast majority of homes have a garage.

3.3 Notification

Households were notified of their recruitment visit a day beforehand with a flyer placed on their doorknob. The flyers were designed to inform households when the visit would occur, the treatment being studied, and the procedures associated with its receipt. Figure 4 provides examples of the baseline and opt-out flyers used. Panel A shows that across the two flyers, the only difference is the option in the opt-out condition to, "check this box if you do not want to be disturbed." This wording was taken from past research on the consequences of social pressure for charitable giving and survey response (DellaVigna et al., 2012, 2017; Giaccherini et al., 2019). The date and time of a recruitment visit was handwritten on each flyer. The promised time windows were either in the morning from 10am to Noon

[17] In particular, the treatment was an eight pack of Cree A19 60 watt equivalent light bulbs. The specific characteristics of the light bulbs included the production of soft white light, the ability to operate on a dimmer, the option of use in outdoor lighting fixtures, and no required time to warm-up. Additionally, they are expected to work for 13 years of use, at which point they can be easily disposed because they do not contain the dangerous chemicals found in fluorescent bulbs.

[18] The focus on single family homes is due to employment restrictions that forbid staff from entering a premise, such as an apartment complex.

Table 1: Characteristics of Population by Recruitment Flyer

	Baseline Flyer	Opt-Out Flyer
Age of Home: 0-50 Years	3.4	2.7
	(0.58)	(0.50)
	$p = 0.35$	
Age of Home: 50-100 Years	57.0	59.7
	(3.33)	(3.18)
	$p = 0.56$	
Square Footage: 0-1,500	31.6	33.3
	(2.55)	(2.67)
	$p = 0.63$	
Square Footage: 1,500-2,500	49.8	47.9
	(1.74)	(1.89)
	$p = 0.46$	
Stories: 2+	63.4	60.1
	(2.65)	(2.75)
	$p = 0.39$	
Air Conditioning: Central	52.0	51.7
	(1.73)	(1.83)
	$p = 0.91$	
Full Bathrooms: 2+	49.2	49.2
	(2.18)	(2.01)
	$p = 1.00$	
Basement: Finished	29.0	29.8
	(1.46)	(1.40)
	$p = 0.69$	
Garage	87.4	87.8
	(0.83)	(1.00)
	$p = 0.74$	

Note: This table summarizes the characteristics of households in the population recruited for the experimental sample. These characteristics were collected from a web scrape of the Cook County Assessor's online database. The first column reports the proportion of households with a given characteristic that received a baseline flyer, while the second column reports that same information for households receiving an opt-out flyer. Underneath each proportion is a standard error that is robust to heteroskedasticity and route-level correlations. Finally, a p-value is reported to assess the balance of recruitment flyer assignment.

or in the early afternoon from 1pm to 3pm. Additionally, two other flyers were used that duplicated the baseline and opt-out flyer in Figure 4, but included mention of a $10 incentive for participating in the study. Panel B highlights the wording used for the incentive.

In total, the population of 4,888 households received one of four flyers: A baseline or opt-out flyer, each of which was crossed with no incentive or with a $10 incentive for participation. Figure 5 shows the count of households receiving each flyer. Households were randomly assigned to a specific flyer in clusters, or routes, of approximately twenty homes. As a consequence, all inference in this paper will account for correlations in the relevant unobservable at the route level. To assess the quality of the random assignment, Table 1 reports the p-value from a test of balance across the various household characteristics observed in the sample receiving a baseline flyer versus the sample receiving an opt-out flyer. Consistent with claim that the flyers were randomly assigned, Table 1 reports that there were no statistically significant differences in household characteristics between the two samples.

The logistics of flyering largely revolved around maximizing the odds of households noticing their flyer. Towards that aim, flyering was conducted by a team of research assistants between, approximately, 9am and 2pm. This way households had plenty of time to notice a flyer ahead of their recruitment visit. Furthermore, research assistants used a mini stapler to secure each flyer on the assigned doorknob so as to avoid the flyers falling off and going unnoticed.

3.4 Recruitment

A total of 12 University of Chicago students worked as recruiters for the experiment. Students were paid $13 per hour for their time. Advertisements to work on a door-to-door study evaluating an energy efficiency program were shared on university listhosts and posted around campus. Students responding to the advertisements were then interviewed by the author and offered a position if they were interested. Before recruiting for the first time, students participated in a training session where the study materials and procedures were reviewed in detail. Additionally, the team of recruiters specialized in this task and did not also flyer.

Recruitment was conducted on weekends between mid-July and late-August of 2019. To signal the legitimacy of the study, recruiters wore a lanyard displaying their student photo identification card and were encouraged to wear University of Chicago apparel. Recruiters were assigned four routes per day over which they would recruit households receiving each type of flyer.

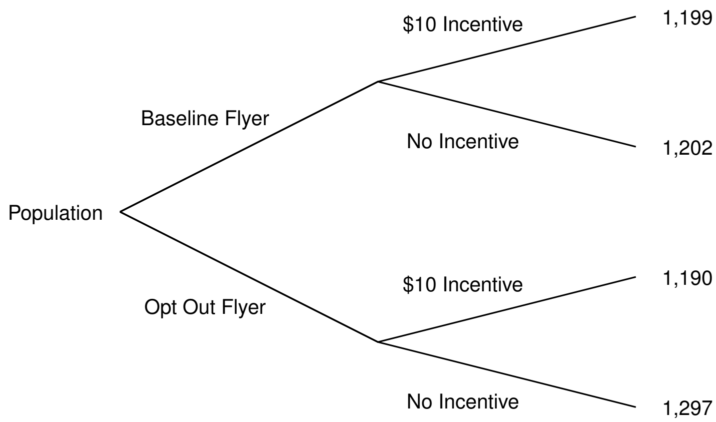
Figure 4: Flyers Used to Notify Households of Recruitment Visit

A. No Incentive

B. $10 Incentive

Note: Examples of the baseline and opt-out flyers used to notify households a day before their recruitment visit.

Figure 5: Assignment of Population to Recruitment Flyer

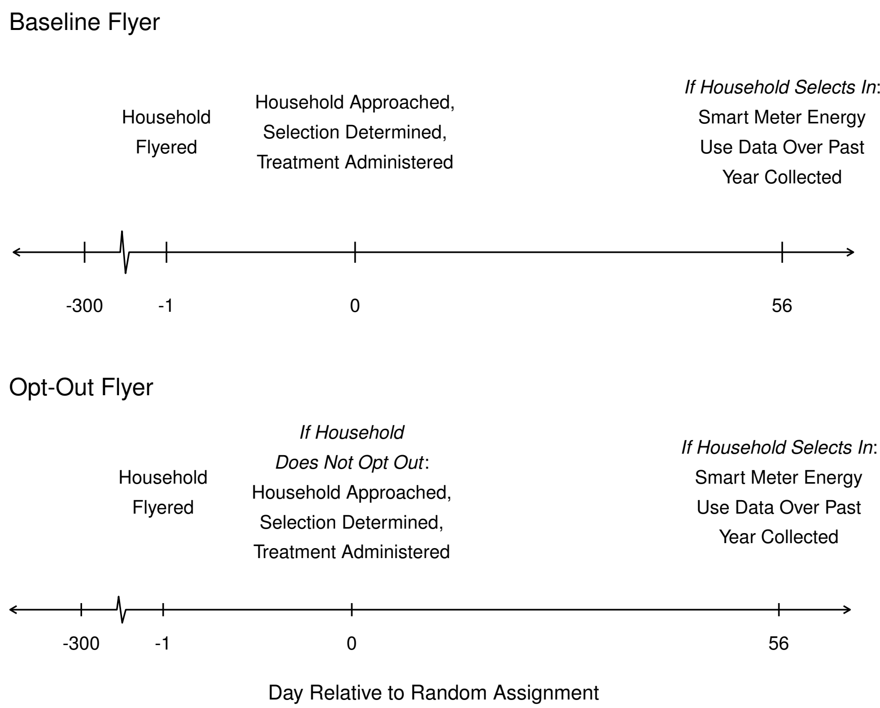


Note: This figure illustrates the allocation of the population of households recruited to the four types of flyers.

Recruiters approached between 50-100 households per day. Figure 6 summarizes the timeline of the study across the baseline and opt-out flyer types. When a household selected out of the study by checking their opt-out box, the recruiter would record this information and proceed to the next house. Otherwise recruiters would alert a household with a knock on their door or ring of their doorbell. When a household answered their door, the recruiter would read a paper script on a clipboard that inquired as to whether the household was interested in finding out more about a study on a free LED lighting program. In the $10 incentive condition, recruiters would also mention the payment and keep a $10 bill on top of their clipboard to signal the mode of payment.

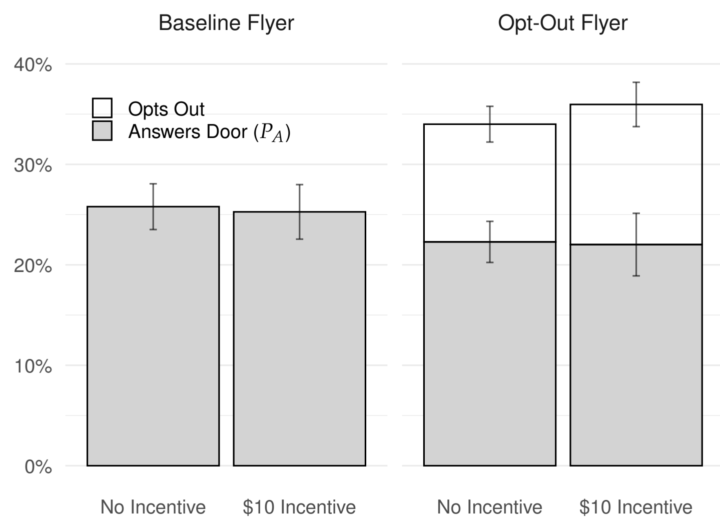
For households expressing interest in the study, recruiters would then activate an Apple iPad with a wireless data connection and read an informed consent form detailing the study procedures. Consenting households were then tasked with completing an online form that would authorize sharing information on their energy usage at the end of the study. While this online form was maintained by a company called UtilityAPI, it was hosted on a University of Chicago branded website. Households could complete this online form on the iPad or their own device. Households successfully completing this online form were then immediately randomized into the treatment or control group. Assignment to the treatment group led to the immediate receipt of the package of eight LED lightbulbs, whereas control assignment led to no light bulbs. In the $10 incentive condition, every household completing

Figure 6: Timeline of Recruitment, Selection, and Collection of Outcome



Note: Relative to the day selection and the randomization occurred, households received their flyer on day -1. On day 0 they received their recruitment visit. If they answered their door and selected into the study, then the randomization was conducted immediately and the treatment was administered. 56 days after the randomization occurred, energy data for the households selecting into the study was accessed for the past year, allowing observation of energy usage from day -300 to 56. The main distinction between the two types of flyers is that households opting out in response to the opt-out flyer were not recruited on day 0.

Figure 7: Door Answer by Recruitment Flyer



Note: Average door answer rates for each flyer type and, for the opt-out flyer, the proportion of households opting out. 95 percent confidence intervals are reported for each estimate, which account for route level correlations between households.

the online form immediately received their payment.

4. Results

4.1 Selection

To test the predictions developed in Section 2.3, I assess the effect of the recruitment flyers on the frequency of households in the population answering their door, P_A, participating in the study, P_D, and participating in the study conditional on answering their door $P_{D|A}$.

Figure 7 reports the share of households answering their door, P_A, for each recruitment flyer. Starting with the baseline flyer, 25 to 26 percent of households are found to answer their door in response to the recruiter. Moving to the opt-out flyer, two shares are plotted for each level of the incentive. First, the propensity for a household to check their opt-out box is reported, with approximately 12 to 14 percent undertaking this option. Second, the share answering their door is plotted. Across both incentive levels this share appears nearly identical at 22 percent. These trends are broadly consistent with the predictions of the self-selection model, with fewer households answering the door in response to the opt-out flyer than the baseline flyer, $P_{A_b} > P_{A_{oo}}$. Also of note is the effect of offering a \$10 incentive, which appears to reduce the share answering their door in response to the baseline flyer and increases the share checking the opt-out box in response to the opt-out flyer.

Figure 8 advances the analysis past the door answering phase of recruitment by plotting the share of households participating in the study across each recruitment flyer. In Panel A, the unconditional participation rate is plotted, P_D. Therein 1.5 to 2 percent of households are found to select into the study in response to the baseline flyer and about 2 percent are found to participate when assigned to the opt-out flyer. Interestingly, this runs counter to the prediction of the self-selection model, which predicted $P_{D_b} > P_{D_{oo}}$. However, evidence more in line with the model is reported in Panel B, with 6 to 7 percent of households opening their door selecting into the study when assigned the baseline flyer. This number increases to nearly 10 percent when, instead, the opt-out flyer is used. This trend is consistent with predicted dynamic of the self-selection model, $P_{D_b|A_b} < P_{D_{oo}|A_{oo}}$. Another dynamic worth mentioning in Figure 8 is that across these two panels the \$10 incentive is found to have little to no effect on selection for the opt-out flyer, but appears to crowd out selection for the baseline flyer.

To conduct inference on the trends highlighted in Figures 7 and 8, I consider the following empirical model for door answer,

$$A_i = \alpha_0 + \alpha_A O_i + W_i^A \tag{5}$$

where A_i is an indicator for household i answering their door, O_i is an indicator for whether they were recruited with an opt-out flyer, and W_i^A is a household specific unobservable that is orthogonal via randomization. I also consider a nearly identical model for the self-selection decision,

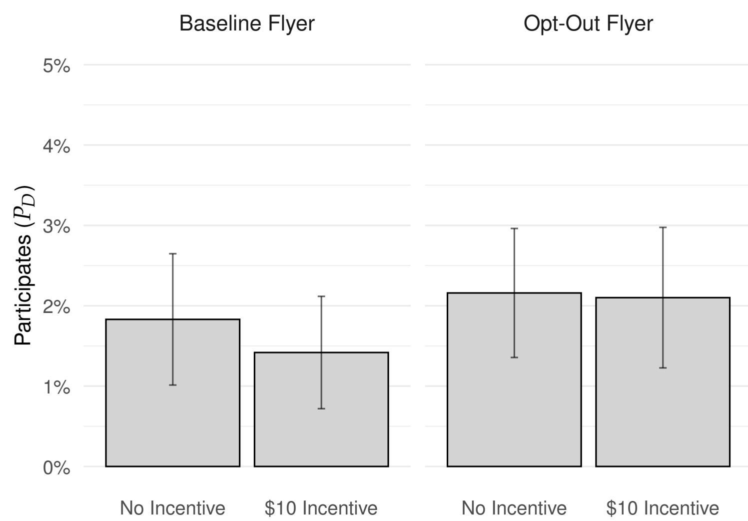
$$D_i = \alpha_0 + \alpha_D O_i + W_i^D \tag{6}$$

which is identical to equation 5 with the exception of the outcome, which is an indicator for household i self-selecting into the study.

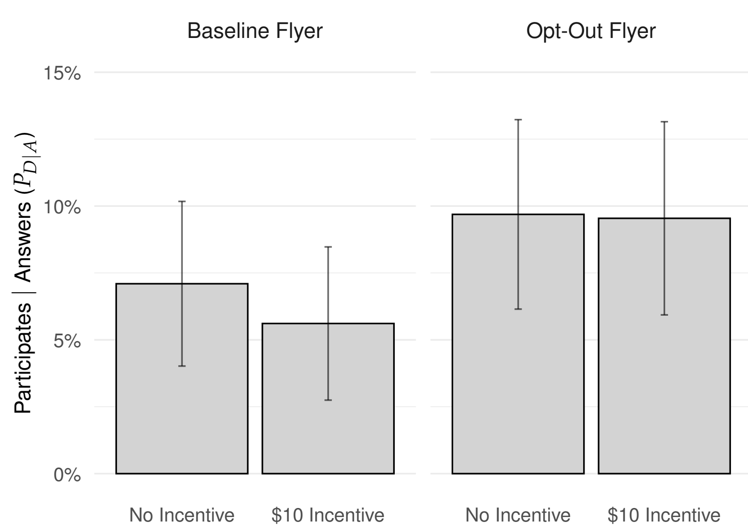
Table 2 reports the parameters in equations 5 and 6 estimated by ordinary least squares. Furthermore, Table 2 reports the standard errors for these parameters, which allow for arbitrary correlations across the households in a given route. Columns 1-3 focus on the door answer decision. Therein the trends observed in Figure 7 are confirmed, with the opt-out flyer causing a 3 percentage point reduction in the probability of households answering their door. Moving from columns 1 to 3, this effect is established to persist in spite of the control

Figure 8: Self-Selection into Experimental Sample by Recruitment Flyer

A. Participation Rate



B. Participation Rate Conditional on Answering Door



Note: Average door answer rates for each flyer type and, for the opt-out flyer, the proportion of households opting out. 95 percent confidence intervals are reported for each estimate, which account for route level correlations between households

Table 2: Self-Selection into Experimental Sample by Recruitment Flyer

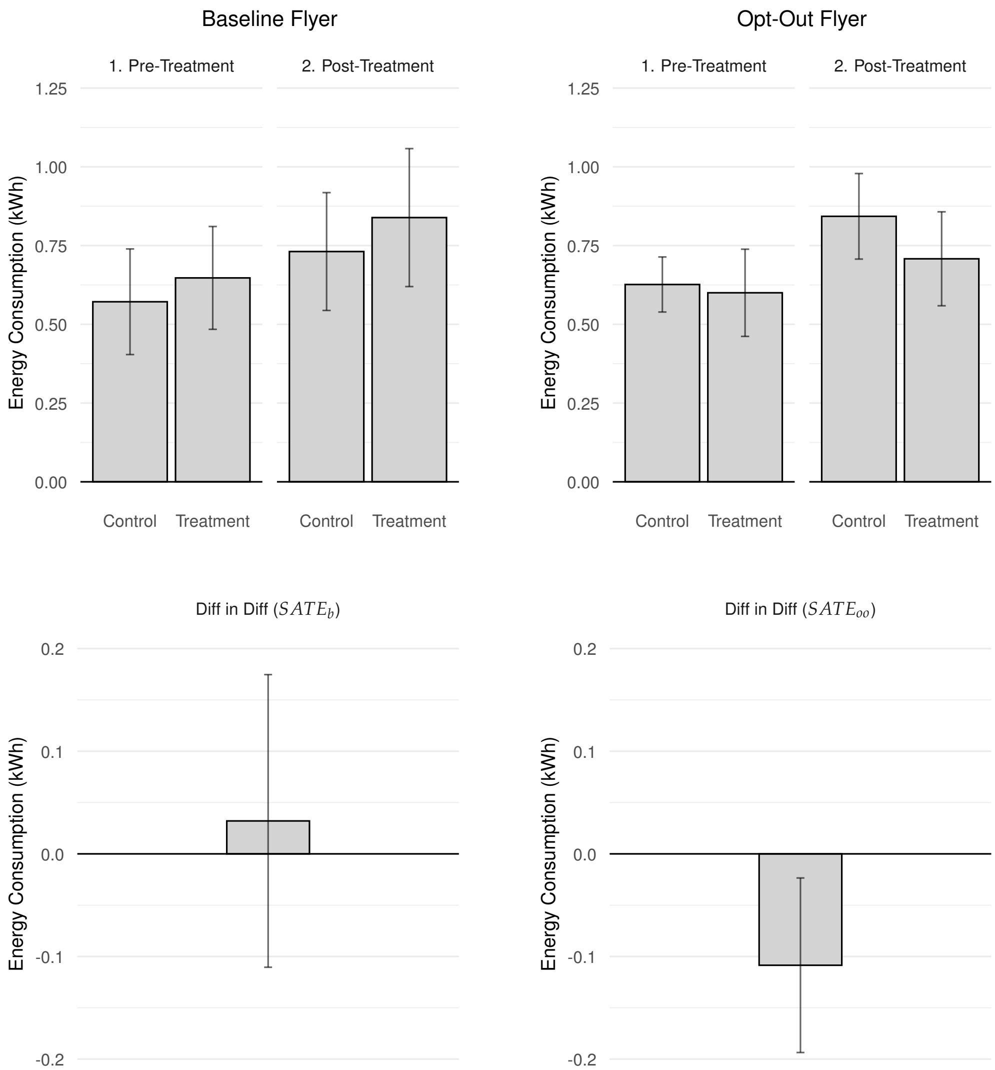
	Answer Door			Participate			Participate | Answer		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Opt-Out	-3.4	-3.4	-2.8	0.5	0.5	0.6	3.3	3.2	3.7
	(1.3)	(1.3)	(1.3)	(0.4)	(0.4)	(0.4)	(1.7)	(1.7)	(1.6)
$10 Incentive		-0.4	-0.5		-0.2	-0.2		-0.8	-0.8
		(1.3)	(1.2)		(0.4)	(0.4)		(1.7)	(1.6)
Constant	25.5	25.7		1.6	1.7		6.4	6.8	
	(0.9)	(1.1)		(0.3)	(0.4)		(1.1)	(1.4)	
Controls			X			X			X
Households	4,888	4,888	4,888	4,888	4,888	4,888	1,164	1,164	1,164
Routes	259	259	259	259	259	259	256	256	256
R^2	0.002	0.002	0.013	0.000	0.000	0.012	0.004	0.004	0.054

Note: This table reports the effect of the opt-out flyer and the $10 incentive on door answering, study participation, and study participation conditional on door answering. Effects are estimated with ordinary least squares and standard errors robust to correlations across households in a given route are reported. The controls used are fixed effects for the recruiter, the calendar day of recruitment, the time of recruitment, and the household characteristics in Table 1.

variables included in the empirical model, with the third column including a fixed effect for the recruiter, the calendar day, the time of recruitment, and each of the characteristics reported in Table 1. Furthermore, the size of the effect is statistically significant at conventional levels, with p-values ranging from 0.01 to 0.03. Also of note in Table 2 is the negative effect of the $10 incentive. While this effect is not statistically significant it runs counter to standard economic intuition.

Moving to the estimation of equation 6 in Table 2 in Columns 4-6, the effect of the opt-out flyer on participation is found to be approximately 0.5 percentage points. However across the specifications this estimate cannot reject the null of no effect at traditional levels of significance, with p-values equal to 0.22 without the full set of fixed effects and 0.11 with the controls included. In Columns 7-9, equation 6 is estimated on the subsample of households answering there door. Across these columns, the opt-out flyer is found to cause a 3 to 4 percentage point increase in the participation rate, with p-values of 0.05 and 0.06 for the specifications without the full set of controls and 0.02 with those controls added.

Figure 9: Evening Energy Use by Recruitment Flyer



Note: The top two panels of this figure plot average evening energy consumption for households assigned to the treatment and control group in the pre- and post-treatment periods. Evening is defined as 6pm to Midnight, a definition that was pre-registered. The bottom two panels plot the difference-in-difference estimator associated with the levels of energy consumption above. 95 percent confidence intervals are reported, which allow for arbitrary across household correlations at the route level.

4.2 Impact of Treatment

I next consider the effect of the LED lighting program on household energy consumption measured in the baseline and opt-out recruitment conditions. My investigation focuses on energy consumption during evening hours, a decision that was pre-registered in my pre-analysis plan on the American Economic Association's registry for randomized control trials.

For the baseline and opt-out flyers, Figure 9 plots average energy usage during this time period for households assigned to the treatment and control group in the pre- and post-treatment period. Energy usage measured in kilowatt hours and this is observed in thirty minute increments. Starting with the baseline flyer, similar gaps between treatment and control group energy usage are found in the pre- and post-treatment periods. Moving to the opt-out flyer, though, a very different dynamic emerges. In particular, a small gap between treatment and control households in the pre-treatment period is dramatically increased in the post-treatment period.

The bottom two panels of Figure 9 plots the effect of the LED lighting program with the difference-in-difference estimator observed for each flyer. For the sample recruited with the baseline flyer, the effect of the program is negligible, while there is a sizable energy savings for the households recruited with the opt-out flyer.

To more formally assess the effect of the LED lighting program, I conduct a regression analysis with a model that estimates the difference-in-difference during evening hours and non-evening hours. The results of this analysis are reported in Table 3. Across four specifications that vary the inclusion of key control variables, I find that the LED lighting program causes households recruited with the baseline flyer to save no energy during evening and non-evening hours. However, that same program causes a 0.09 to 0.11 reduction in the number of kilowatt hours consumed during evening hours amongst households recruited with the opt-out flyer. Importantly these estimates are measured with a sufficient level of precision to reject the null hypothesis, with p-values of approximately 0.02. This dynamic is consistent with the prediction of the self-selection model developed in this paper, which predicted an attenuated $SATE_b$ relative to $SATE_{oo}$.

5. Conclusion

This paper examines the consequences of a subtle change to recruitment in an evaluation experiment. The manipulation that is implemented is motivated by a model of social pres-

Table 3: Average Impact of Treatment by Recruitment Flyer

	Energy Usage per 30 Minute Interval (kWh)							
	(1)		(2)		(3)		(4)	
	Baseline	Opt Out	Baseline	Opt Out	Baseline	Opt Out	Baseline	Opt Out
Evening Hours (18-24)								
Trt X Post	0.03	-0.11	0.02	-0.09	0.04	-0.10	0.03	-0.11
	(0.07)	(0.04)	(0.07)	(0.04)	(0.07)	(0.04)	(0.07)	(0.04)
Trt	0.08	-0.03			0.07	-0.03	0.07	-0.03
	(0.12)	(0.08)			(0.12)	(0.09)	(0.12)	(0.09)
Post	0.16	0.22	0.16	0.21	0.01	0.09	0.10	0.16
	(0.04)	(0.04)	(0.04)	(0.04)	(0.09)	(0.08)	(0.04)	(0.04)
Constant	0.57	0.63						
	(0.09)	(0.04)						
Non-Evening Hours (0-18)								
Trt X Post	-0.02	-0.06	-0.03	-0.04	-0.01	-0.06	-0.02	-0.06
	(0.04)	(0.05)	(0.05)	(0.04)	(0.04)	(0.05)	(0.04)	(0.05)
Trt	0.04	0.01			0.03	0.00	0.04	0.01
	(0.08)	(0.07)			(0.08)	(0.07)	(0.08)	(0.07)
Post	0.08	0.12	0.07	0.12	-0.07	-0.01	0.02	0.07
	(0.03)	(0.03)	(0.03)	(0.03)	(0.08)	(0.07)	(0.03)	(0.03)
Constant	0.47	0.46						
	(0.06)	(0.04)						
Fixed Effects			Household		Day		Max Daily Temp.	
Observations	1,204,245		1,204,245		1,204,245		1,204,245	
Routes	59		59		59		59	
R^2	0.026		0.242		0.158		0.092	

Note: This table reports the effect of the LED lighting program on household energy consumption. Effects are estimated with ordinary least squares and standard errors robust to correlations across households in a given route are reported.

sure. I find that this subtle manipulation, which varies the social scrutiny associated with selecting out of an experimental sample for a study evaluating a free LED lighting program, has drastic consequences for the policy insights that are obtained. In particular, I find that allowing households to select out of the study without any scrutiny from the experimenter leads to an evaluation experiment that measures substantial energy savings from an LED lighting program, whereas a more traditional recruitment strategy finds no energy savings. These results establish that the manner in which evaluation experiments are conducted can dramatically influence the very results they attempt to measure. It also suggests that models in policy experimentation could more accurately isolate the parameter of interest and have greater descriptive ability if they factor in the social features of experimental research designs.

References

Akerlof, George A., "Social Distance and Social Decisions," *Econometrica*, 1997, *65* (5), 1005–1027.

— **and Rachel E. Kranton**, "Economics and Identity," *Quarterly Journal of Economics*, 2000, *115* (3), 715–753.

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

Andreoni, James, "Giving with Impure Altruism: Applications to Charity and Ricardian Equivalence," *Journal of Political Economy*, 1989, *97* (6), 1447–1458.

— , "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving," *Economic Journal*, 1990, *100* (401), 464–477.

— **, Justin M. Rao, and Hannah Trachtman**, "Avoiding the Ask: A Field Experiment on Altruism, Empathy, and Charitable Giving," *Journal of Political Economy*, 2017, *125* (5), 625–653.

Angrist, Joshua D. and Guido W. Imbens, "Sources of Identifying Information in Evaluation Models," *NBER Working Paper No. 117*, 1991.

Ashenfelter, Orley, "The Case for Evaluating Training Programs with Randomized Trials," *Economics of Education Review*, 1987, *6* (4), 333–338.

Bager, Simon and Luis Mundaca, "Making 'Smart Meters' Smarter? Insights from a Behavioral Economics Pilot Field Experiment in Copenhagen, Denmark," *Energy Research and Social Science*, 2017, *28*, 66–76.

Banerjee, Abhijit and Esther Duflo, "The Experimental Approach to Development Economics," *Annual Review of Economics*, 2009, *1* (1), 151–178.

Becker, Gary S., "A Theory of Social Interactions," *Journal of Political Economy*, 1974, *82* (6), 1063–1093.

Benabou, Roland and Jean Tirole, "Incentives and Prosocial Behavior," *American Economic Review*, 2006, *96* (5), 1652–1678.

Bernheim, B. Douglas, "A Theory of Conformity," *Journal of Political Economy*, 1994, *102* (5), 841–877.

Bloom, Howard S., "Accounting for No-Shows in Experimental Evaluation Designs," *Evaluation Review*, 1984, *8* (2), 225–246.

Brandon, Alec, Christopher M. Clapp, John A. List, Robert Metcalfe, and Michael K. Price, "Do Smart Technologies Deliver? Experimental Evidence on Smart Thermostats and Energy Conservation," *University of Chicago Working Paper*, 2019.

Broberg, Tomas, Tore Ellingsen, and Magnus Johannesson, "Is generosity involuntary?," *Economics Letters*, 2007, *94* (1), 32–37.

Burtless, Gary and Larry L. Orr, "Are Classical Experiments Needed for Manpower Policy," *Journal of Human Resources*, 1986, *21* (4), 606–639.

Dana, Jason, Daylian Cain, and Robyn M. Dawes, "What you don't know won't hurt me: Costly (but quiet) exit in dictator games," *Organizational Behavior and Human Decision Processes*, 2006, *100* (2), 193–201.

Deaton, Angus and Nancy Cartwright, "Understanding and Misunderstanding Randomized Controlled Trials," *Social Science and Medicine*, 2018, *210*, 2–21.

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

_ , _ , _ , and Gautam Rao, "Voting to Tell Others," *The Review of Economic Studies*, 2017, *84* (1), 143–181.

Duflo, Esther, Rachel Glennerster, and Michael Kremer, "Using Randomization in Development Economics Research: A Toolkit," in T. Paul Schultz and John Strauss, eds., *Handbook of Development Economics*, Vol. 4, Elsevier, 2007.

Eckel, Catherine C. and Philip J. Grossman, "Volunteers and Pseudo-Volunteers: The Effect of Recruitment Method in Dictator Experiments," *Experimental Economics*, 2000, *3* (2), 107–120.

Faruqui, Ahmad, Sanem Sergeci, and Lamine Akaba, "The Impact of Dynamic Pricing on Residential and Small Commercial and Industrial Usage: New Experimental Evidence from Connecticut," *Energy Journal*, 2014, *35* (1), 137–160.

Fowlie, Meredith, Catherine Wolfram, C. Anna Spurlock, Annika Todd, Patrick Baylis, and Peter Cappers, "Default Effects and Follow-On Behavior: Evidence from an Electricity Pricing Program," *NBER Working Paper No. 23553*, 2017.

Frank, Robert H., "The Demand for Unobservable and Other Nonpositional Goods," *American Economic Review*, 1985, *75* (1), 101–116.

Gandhi, Raina, Christopher Knittel, Paula Pedro, and Catherine Wolfram, "Running Randomized Field Experiments for Energy Efficiency Programs: A Practitioner's Guide," *Economics of Energy and Environmental Policy*, 2016, *5* (2), 7–26.

Giaccherini, Matilde, David H. Herberich, David Jimenez-Gomez, John A. List, Giovanni Ponti, and Michael K. Price, "The Behavioralist Goes Door-to-Door: Understanding Household Technological Diffusion Using a Theory-Driven Natural Field Experiment," *NBER Working Paper No. 26173*, 2019.

Harding, Matthew and Carlos Lamarche, "Empowering Consumers Through Data and Smart Technology: Experimental Evidence on the Consequences of Time-of-Use Electricity Pricing Policies," *Journal of Policy Analysis and Management*, 2016, *35* (4), 906–931.

Harrison, Glenn W. and John A. List, "Field Experiments," *Journal of Economic Literature*, 2004, *42* (4), 1009–1055.

Hausman, Jerry A. and David A. Wise, "Attrition Bias in Expeirmental and Panel Data: The Gary Income Maintenance Experiment," *Econometrica*, 1979, *47* (2), 455–473.

Heckman, James J., "Randomization and Social Policy Evaluation," in Charles Manski and Irwin Garfinkel, eds., *Evaluating Welfare and Training Programs*, Harvard University Press, 1992.

_ **and Edward Vytlacil**, "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Elsevier, 2007.

_ **and** _ , "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments," in James J. Heckman and Edward Leamer, eds., *Handbook of Econometrics*, Vol. 6B, Elsevier, 2007.

_ **and Jeffrey A. Smith**, "Assessing the Case for Social Experiments," *Journal of Economic Perspectives*, 1995, *9* (2), 85–110.

Imbens, Guido W. and Joshua D. Angrist, "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 1994, *62* (2), 467–475.

Ito, Koichiro, Takanori Ida, and Makoto Tanaka, "Moral Suasion and Economic Incentives: Field Experimental Evidence from Energy Demand," *American Economic Journal: Economic Policy*, 2018, *10* (1), 240–267.

Jessoe, Katrina and David Rapson, "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use," *American Economic Review*, 2014, *104* (4), 1417–1438.

LADWP, "LADWP Delivers Energy Efficiency Directly to Its Residential Customers," Press Release June 2019.

Lazear, Edward P., Ulrike Malmendier, and Roberto A. Weber, "Sorting in Experiments with Application to Social Preferences," *American Economic Journal: Applied Economics*, 2012, *4* (1), 136–163.

Levitt, Steven D. and John A. List, "What Do Laboratory Experiments Measuring Social Preferences Reveal About the Real World?," *Journal of Economic Perspectives*, 2007, *21* (2), 153–174.

Maguire, Kelly B., "Does mode matter? A comparison of telephone, mail, and in-person treatments in contingent valuation surveys," *Journal of Environmental Management*, 2009, *90* (11), 3528–3533.

Malani, Anup, "Patient enrollment in medical trials: Selection bias in a randomized experiment," *Journal of Econometrics*, 2008, *144* (2), 341–351.

Mannesto, Gregory and John B. Loomis, "Evaluation of Mail and In-person Contingent Valuation Surveys: Results of a Study of Recrational Boaters," *Journal of Environmental Management*, 1991, *32* (2), 177–190.

Orcutt, Guy H. and Alice G. Orcutt, "Incentive and Disincentive Experimentation for Income Maintenance Policy Purposes," *American Economic Review*, 1968, *58* (4), 754–772.

Peters, Jörg, Jörg Langbein, and Gareth Roberts, "Policy evaluation, randomized controlled trials, and external validity–A systematic review," *Economics Letters*, 2016, *147*, 51–54.

Rolnick, Sharon J., Cynthia R. Gross, Judith Garrard, and Robert W. Gibson, "A Comparison of Response Rate, Data Quality, and Cost in the Collection of Data on Sexual History and Personal Behaviors," *American Journal of Epidemiology*, 1989, *129* (5), 1052–1061.

Roy, Andrew D., "Some Thoughts on the Distribution of Earnings," *Oxford Economic Papers*, 1951, *3* (2), 135–146.