

Is Replication Enough? Site Selection Bias in Program Evaluation

Hunt Allcott*

October 30, 2013

Abstract

Program evaluation often involves generalizing site-specific estimates to different or larger target populations. Given the potential for site-specific treatment effect heterogeneity, replication is viewed as important because it gives a sense of the distribution of possible effects. However, randomized control trials (RCTs) are only feasible in limited situations, and “site selection bias” can result if the same factors that make RCTs feasible also moderate treatment effects. I formalize the econometric assumptions that underlie this intuition, which I call “external unconfoundedness.” I then test external unconfoundedness in the context of a series of energy conservation RCTs involving 6.8 million households in 98 sites across the U.S. I show evidence of two positive site selection mechanisms. First, early Opower partners were in areas where consumers have more environmentalist tastes, which is associated with larger treatment effects. Second, utilities typically targeted initial interventions at the most responsive consumer subpopulations, meaning that subsequent interventions within the same utility often performed more poorly. I augment these results by showing suggestive evidence from two other domains: microfinance institutions (MFIs) that partner on academic experiments differ on observables from the global population of MFIs, and clinical trials for drugs and surgical procedures take place at hospitals that differ from the national population of hospitals.

JEL Codes: C93, D12, L94, O12, Q41.

Keywords: Randomized control trials, external validity, selection bias, energy conservation.

This paper replaces an earlier version which was titled "External Validity and Partner Selection Bias." That earlier draft was jointly authored with Sendhil Mullainathan, and the project has benefitted substantially from his insight and collaboration. I thank Josh Angrist, Amitabh Chandra, Lucas Davis, Meredith Fowlie, Xavier Gine, Chuck Goldman, Matt Harding, Joe Hotz, Guido Imbens, Larry Katz, Dan Levy, Jens Ludwig, Konrad Menzel, Emily Oster, Rohini Pande, Todd Rogers, Piyush Tandia, Ed Vytlačil, and seminar participants at the ASSA meetings, Berkeley, Columbia, Harvard, MIT, NBER Labor Studies, NBER Energy and Environmental Economics, NEUDC, the UCSB/UCLA Conference on Field Experiments, and the World Bank for insights and helpful advice. Thanks also to Tyler Curtis, Marc Laitin, Alex Laskey, Nate Srinivas, Dan Yates, and others at Opower for fruitful discussions. Christina

*New York University and National Bureau of Economic Research. Email: hunt.allcott@nyu.edu.

Larkin provided timely research assistance. I am grateful to the Sloan Foundation for financial support of this paper and related research on the economics of energy efficiency.

1 Introduction

Program evaluation is often used to make a policy decision: should a treatment be implemented in some "target" population? In some cases, evaluations are carried out in the full target population of policy interest, or in a randomly-selected subset thereof. In most cases, however, an evaluation is performed at one or more sample sites, and the results are generalized to make an implementation decision in a different and often larger set of target sites. This raises questions of external validity: how well does a parameter estimate generalize across sites?

When generalizing empirical results, we often implicitly or explicitly make one of two assumptions. First, if extrapolating from one sample site to one different target, we might assume that heterogeneity across sites is minimal, so results can be meaningfully generalized. Frequently, however, this assumption is unrealistically strong, meaning that it is important to replicate in additional sites. After enough replications, we might make a second assumption: that the distribution of treatment effects in the sample sites is a reasonable predictor of the distribution of effects in other target sites. Put simply, if an intervention works well in a few different places, we might advocate that it be scaled up. This second assumption would hold if sample sites had been selected randomly from the population of targets.

In practice, there are many reasons why sites are selected for empirical study. For example, because randomized field experiments require an implementing partner with managerial ability and operational expertise, the set of actual partners may be able to run more effective programs than the typical potential partner. As another example, partners that are already running programs that they know are effective are more likely to be open to independent impact estimates (Pritchett 2002). Both of these mechanisms would cause a positive *site selection bias*: Average Treatment Effects (ATEs) from sample sites would be larger than in the full set of targets. Alternatively, partners that are particularly innovative and willing to test new programs may also be running many other effective programs in the same area. If there are diminishing returns, the additional program with an actual partner might have less impact than with a typical potential partner. This would cause negative site selection bias.

While the idea of site selection bias might seem intuitive, there is little empirical evidence on this issue or the mechanisms through which might act. The reason is simple: since this type of selection operates at the level of the site instead of the individual treated unit, one needs a statistically meaningful sample of internally valid studies of the same intervention from different sites. Then, one must define a population of potential partner sites and somehow infer treatment effects in sites where studies have not yet been carried out. Given the cost of evaluating an intervention at one site, it is unusual for the same intervention to be rigorously evaluated at more than one site, or perhaps a small handful of sites. By contrast, providing evidence on individual-level selection bias and the internal validity of an estimator is much less onerous, as this requires a comparison of experimental to non-experimental results in only one setting, as in LaLonde (1986) and the literature that follows.

I study the Opower energy conservation program, a remarkable series of 98 randomized control trials involving 6.8 million households in sites dispersed across the United States. Opower mails "Home Energy Reports" to residential electricity consumers that provide energy conservation tips and compare their energy use to that of their neighbors. Electric and natural gas utilities partner with Opower largely because the program helps to comply with state-level energy conservation mandates. Aside from being a case study of broader issues of external validity, the generalizability of Opower's impact estimates is of great interest *per se*, as policymakers use efficacy at other sites to help determine whether or not to test the intervention locally.

I begin by studying treatment effect heterogeneity. I exploit 20.5 million observations of electricity use at 495,000 households in nine initial sites to test factors that could moderate treatment effects. Most importantly, effects are conditionally associated with usage and with measures of environmentalist preferences. Using data from all 98 RCTs, I also show that site-level ATEs over the first year of each program vary substantially, ranging from 0.074 to 1.012 kilowatt-hours per day, or 0.2 to 3.3 percent of national average electricity usage. This dispersion is economically significant in the sense that it should affect program adoption decisions: the support of the distribution of cost effectiveness spans the range of estimates for competing energy conservation programs,

I then test for site selection bias using two different approaches. The first approach compares Opower partner utilities to a natural target population, the set of all large electric utilities in the United States. Partners differ on site-level characteristics that are also associated with treatment effects in the set of sample sites. In particular, partners seem to be strongly selected through a mechanism associated with consumer preferences: areas that prioritize energy conservation (typically wealthier coastal regions of the U.S.) promulgate stronger energy conservation regulations, which then encourage utilities to partner with Opower. These same pro-environment populations are also more responsive to the intervention. However, this positive selection mechanism is offset by the fact that sample sites tend to be in more temperate climates with lower electricity demand, which is associated with smaller treatment effects. Looking across the entire set of observed site-level characteristics that might have influenced selection, Opower sample sites are neutrally or perhaps slightly positively selected on observables from the population of potential partner utilities.

The second approach exploits the timing of program implementation. Opower's first programs were implemented in 2008, and early evidence of cost effectiveness led to expansions to additional partner utilities and to additional sub-populations of households within the service territories of existing partner utilities. Some of these later programs have had large impacts. However, there are a mass of programs that have performed worse than the initial sites, and the trend over time has been toward smaller impacts. In other words, the initial sites were positively selected from the set of eventual sites.

I show that this positive site selection occurred both within and between partner utilities. When choosing populations of households within a new partner utility, Opower and the partner targeted higher-usage households where the intervention would likely be more cost effective. After the inter-

vention proved cost effective in the initial site-level population, some utilities then implemented the program in other, lower-usage populations where the effects have been smaller. Furthermore, early partners appear to be selected from later partners through the same mechanisms that all partners were selected from the population of utilities: early sites were in areas with more environmentally-conscious consumers and stronger conservation regulation. These populations are more responsive to treatment but have lower baseline energy demand.

The central qualitative result of these analyses is that even extensive replication did not solve the external validity problem in this context. ATEs in early sites were not an unbiased measure of ATEs in later sites. Furthermore, evidence of site selection on some observables suggests that there could equally be selection on unobservables.

After studying Opower, I also provide brief suggestive evidence from two other domains on how RCT sites differ systematically on observables from natural populations of target sites. First, I study microfinance institutions (MFIs) that have partnered to carry out randomized trials with three academic groups: the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Financial Access Initiative. I show that partner MFIs differ from the global population of MFIs on characteristics that might moderate effects of various treatments, including average loan size, cost per borrower, for-profit status, size, and share of borrowers that are female. Second, I study hospitals that are the sites of clinical trials for new drugs and surgical procedures. I show that clinical trial sites tend to be larger, more experienced in surgical procedures, offer a wider range of technologies and patient services, and are generally higher-quality than the average US hospital. Because both microfinance RCTs and clinical trials test a variety of different "treatments," one cannot correlate selection probability or timing with treatment effects as one can for the Opower experiments. However, these additional examples suggest that site selection bias is probably not unique to energy conservation RCTs.

I emphasize that this analysis simply cannot be used to argue that randomized control trials are not useful and important in this context. As shown in Allcott (2011), non-experimental approaches to evaluating the Opower programs that would necessarily be used in the absence of experimental data perform dramatically worse than experimental estimators in the same population. Furthermore, the working paper version showed that non-experimental estimates from the correct target population also perform substantially worse than treatment effects predicted for the target using experimental data from different sample populations. I use the conclusion to suggest several simple steps that can be taken to partially address site selection bias when designing and writing up RCTs.

The paper proceeds as follows. Section 2 lays out the theoretical framework. Section 3 introduces the Opower interventions and data, and Section 4 assesses treatment effect heterogeneity across individuals and sites. Section 5 studies site selection bias in the Opower context. Section 6 presents additional suggestive evidence from microfinance RCTs and clinical trials. Section 7 concludes.

2 Theoretical Framework

This section lays out the theoretical framework, beginning with the individual-level Rubin (1974) causal model and aggregating to site-level effects and site selection mechanisms. A central theme of this section is the analogy between individual-level and site-level selection processes.

2.1 Setup

There is a population of individual units indexed by i . Of interest is a binary treatment that affects observed outcome Y_i . Following Rubin (1974), each individual unit has two potential outcomes, $Y_i(1)$ if exposed to treatment and $Y_i(0)$ if not. For expositional simplicity, we assume that Y_i is a linear and additively-separable function of observed and unobserved characteristics X_i and Z_i :

$$Y_i(0) = \beta X_i + \zeta Z_i \tag{1a}$$

$$Y_i(1) = (\alpha + \beta)X_i + (\gamma + \zeta)Z_i \tag{1b}$$

The linear functional form is not central to the argument, and this could easily be re-written more generally. Individual i 's treatment effect is the difference in Y_i between the treated and untreated states:

$$\tau_i = Y_i(1) - Y_i(0) = \alpha X_i + \gamma Z_i \tag{2}$$

2.2 Site-Level Effects

Imagine now that the population of individual units is divided mutually exclusively and exhaustively into "sites." A "site" is a set of individual units, often grouped by geography, where one program evaluation might be carried out. For example, this might be a school or school district, a job training center, a microfinance institution, or an electric utility. Index sites by r , and define an integer variable R_i that indicates the site of which individual i is a member. Denote the population of sites by script \mathcal{R} . Within each site is also a "site-level population" of individual units.

Assume that each unit in the site-level population is treated with equal probability. This conforms to the Opower empirical examples and keeps the analysis simple; other analyses such as Heckman and Vytlacil (2007b) discuss the implications of differential selection across sites for external validity. Define $Y_r \equiv E[Y_i | R_i = r]$, $X_r \equiv E[X_i | R_i = r]$, and $Z_r \equiv E[Z_i | R_i = r]$. The Average Treatment Effect at site r is:

$$\tau_r = \alpha X_r + \gamma Z_r \tag{3}$$

X_r and Z_r could reflect differences in individual-level demographics across sites, such as income or education. They could also represent differences in economic environments that vary only at the site level: for example, site s could have a higher unemployment rate than site g , and unemployment could moderate the treatment effect. However, whether a characteristic is observable or unobservable in this model depends on whether its effect on τ can be empirically estimated. Observable site-level features such as unemployment rates are part of Z if they do not vary across the observed individual units.

In program evaluation, one might formally think of an estimator as "externally valid" if one can use data from one or more Sample sites to consistently estimate treatment effects in other Target sites. I consider two different situations in which results may be extrapolated. First, suppose that an analyst wants to extrapolate from one Sample site, indexed $r = s$, to one Target site, indexed $r = g$. The ATE in the Target is:

$$\begin{aligned} \tau_g &= \tau_s \\ &+ \alpha(X_g - X_s) + \gamma(Z_g - Z_s) \end{aligned} \tag{4}$$

The second line is the difference between Target and Sample ATEs. The first term is a function of observables X and can be estimated empirically. The second term is what generates unexplained site-level treatment effect heterogeneity.

Consider now a second situation in which a program has been evaluated in many Sample sites drawn from the population of sites \mathcal{R} , and the analyst wants to extrapolate to all sites in \mathcal{R} . Denoting $D_r \in \{1, 0\}$ as an indicator for whether site r is a Sample site, the expected treatment effect is.

$$\begin{aligned} E[\tau_r] &= E[\tau_r | D_r = 1] \\ &+ \alpha(E[X_r] - E[X_r | D_r = 1]) + \gamma(E[Z_r] - E[Z_r | D_r = 1]) \end{aligned} \tag{5}$$

The terms in this equation are analogous to those in Equation (4), except with expectations. The equation is also closely analogous to the usual equation illustrating individual-level selection bias.¹ The second line reflects differences in expected effects between Sample and Target sites,

¹Denote $T_i \in \{1, 0\}$ as the indicator variable for individual i 's treatment assignment. Comparing the mean outcomes of treated vs. untreated units gives:

$$\begin{aligned} E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] &= E[\tau_i | T_i = 1] \\ &+ \beta(E[X_i | T_i = 1] - E[X_i | T_i = 0]) + \zeta(E[Z_i | T_i = 1] - E[Z_i | T_i = 0]) \end{aligned} \tag{6}$$

The right hand side of the first line is the Average Treatment Effect on the Treated (ATT). The second line is

which I call *site selection bias*.

Of course, site selection bias does not mean that the estimated Sample ATEs are biased away from the true Sample ATEs. The model simply captures heterogeneous Conditional Average Treatment Effects that could vary across sites. The reason to use the phrase "site selection bias" is that it underscores that the Sample CATEs could be *systematically* different from Target CATEs. As I will discuss momentarily, these potential systematic differences arise from site selection processes that can be theoretically understood and observed in practice.

2.3 External Unconfoundedness

Denote $D_i \in \{1, 0\}$ as an indicator variable for whether individual i is a member of a Sample site. To extrapolate treatment effects, D_i must be independent of the difference in potential outcomes, an assumption I call *external unconfoundedness*.

Definition 1 *External Unconfoundedness*: $D_i \perp (Y_i(1) - Y_i(0)) | X_i$

This is closely analogous to the *unconfoundedness* assumption required for internal validity (Rosenbaum and Rubin 1983): $T_i \perp (Y_i(1), Y_i(0)) | X_i$. It is a weaker version of *unconfounded location* (Hotz, Imbens, and Mortimer 2005): $D_i \perp (Y_i(1), Y_i(0)) | X_i$.

It is useful to further refine external unconfoundedness into two assumptions which parallel the two situations above, with one and many Sample sites, respectively. In the one-Sample situation, the analyst would use Equation (4) to extrapolate from one Sample to one Target. This gives an unbiased estimate of τ_g only under an assumption I call *strong external unconfoundedness*.

Definition 2 *Strong External Unconfoundedness*: $D_i \perp (Y_i(1) - Y_i(0)) | (X_i, R_i \in \{s, g\})$

In words, this is that external unconfoundedness holds in a *pair* of sites, the Sample and the Target. If strong external unconfoundedness does not hold, treatment effects vary unobservably across sites, and there is no unbiased estimator of τ_g .

As an example of how strong external unconfoundedness has been used, consider analyses of the GAIN job training program that attribute differences in outcomes between Riverside County and other sites only to an emphasis on Labor Force Attachment (Dehejia 2003, Hotz, Imbens, and Klerman 2006). These analyses formally require that there are no unobservable factors that moderate the treatment effect and differ across sites. In fact, any impact evaluation from one site that argues that its results generalize to some different or broader population implicitly or explicitly assumes strong external unconfoundedness.

In many contexts, one expects unobservables to vary across sites, and strong external unconfoundedness is unrealistically restrictive. As a result, one may wish to replicate an experiment across multiple sites, or perform a meta-analysis. Consider now the second situation introduced

individual-level selection bias.

above, in which there are many Sample sites and the Target sites are the entire population of sites \mathcal{R} . Imagine that the researcher could randomly select Sample sites from \mathcal{R} . As the number of randomly-selected Sample sites increases, the distribution of treatment effects in the set of Sample sites would asymptotically equal the distribution of treatment effects in the Target sites. This motivates an assumption I call *external unconfoundedness in distribution*.

Definition 3 *External Unconfoundedness in Distribution: $D_i \perp (Y_i(1) - Y_i(0)) | (X_i, R_i \in \mathcal{R})$*

In words, this is that external unconfoundedness holds in a *population* of sites. When there is exactly one Sample site and one Target site, this assumption is identical to strong external unconfoundedness. However, when there are replications in many sites, external unconfoundedness in distribution is a weaker assumption. Under this assumption, distributions of unobservables may differ between any pair of Sample and Target sites, as long as the unobservables in the sets of Sample and Target sites converge in distribution as the number of sites grows large. As a result, the CATEs from any one Sample may not equal the CATEs from any one Target. However, the mean CATE from Sample sites is a consistent estimator of the expected CATE in Target sites as the number of sites increases. This highlights the logic behind replication: once a program is replicated in enough sites, we know the distribution of Target treatment effects. If external unconfoundedness in distribution does not hold, we have site selection bias.

2.4 Assignment Mechanisms

Imbens and Wooldridge (2009) specify three classes of mechanisms through which individuals are assigned to treatment or control: random assignment, quasi-random assignment, and other mechanisms under which unconfoundedness does not hold. In this section, I discuss three analogous classes of mechanisms that assign a potential Sample site to being an actual Sample site for a randomized control trial.

The "partner assignment mechanism" could represent two situations. First, sites could represent potential program implementation partners that would adopt a *new* program and evaluate it using a randomized trial. This is the case with Opower, as they approach additional utilities about adopting their Home Energy Report program. Second, sites could represent potential program evaluation partners that are already running an *existing* program and must decide whether to run a randomized trial for impact evaluation. This was eventually the case with the Job Training Partnership Act (JTPA) evaluations: the researchers approached job training centers that were already running the program and tried to convince them to implement randomized evaluations.

The first site assignment mechanism is random assignment: Sample sites are randomly selected from \mathcal{R} , the population of Target sites. As the number of sites grows large, external unconfoundedness in distribution holds. Of course, with a small number of sites, unobservables may not be balanced between Sample and Target even if the Sample sites were randomly selected from the

population of sites. In finite sample, just as stratified randomization can improve balance between treatment and control groups, stratified sampling of sites can improve balance between Sample and Target sites. For example, the JTPA evaluation initially hoped to randomly select sites for evaluations within 20 strata defined by size, region, and a measure of program quality (Hotz 1992).

The second class of assignment mechanisms includes non-randomized processes under which external unconfoundedness in distribution holds by assumption. This might arise when the program evaluator can choose the set of Sample sites without restrictions and does so to maximize external validity, but does not have enough sites for the asymptotics of random assignment to be useful. For example, the Moving to Opportunity experiment (Sanbonmatsu *et al.* 2011) was implemented in five cities chosen for size and geographic diversity. Similarly, the RAND Health Insurance Experiment (Manning *et al.* 1988) was implemented in six sites that were chosen for diversity in geographic location, city size, and physician availability.

The third class of assignment mechanisms comprises all mechanisms under which external unconfoundedness in distribution does not hold. In the absence of random assignment or other processes intentionally designed to maximize external validity, there are economic processes that drive selection into partnership. One natural way to model these processes is through a Roy-like selection equation in which decisionmakers at each potential Sample site decide whether to adopt or evaluate a program based on whether the costs outweigh the benefits. Potential Sample sites incur some positive or negative net cost C_r of adopting or evaluating the treatment and weight average outcomes Y_r by ω . I assume that the decisionmaker knows the treatment effect τ_r , although the intuition would be similar under imperfect information as long as the decisionmaker has some informative signal of γZ . The potential partner site becomes an actual partner if its net benefits are positive:

$$\begin{aligned} D_r &= 1 [\omega\tau_r - C_r > 0] \\ &= 1 [\omega(\alpha X_r + \gamma Z_r) - C_r > 0] \end{aligned} \tag{7a}$$

If this process determines selection into partnership, external unconfoundedness in distribution only holds only if it happens to be the case that the selection decision is independent of unobservables, i.e. $(\omega(\alpha X + \gamma Z) - C) \perp \gamma Z$. Otherwise, there is positive or negative *site selection bias*: the Sample ATEs may be larger or smaller than the Target ATEs, as illustrated by Equation (5). Here again, there is a close analogy to how individual-level Roy-style selection into treatment causes selection biases which threaten internal validity.

2.5 Example Site Selection Mechanisms

What is the practical meaning of ω and C ? How might different real-world factors generate positive or negative partner selection bias? Table 1 gives an example set of positive and negative site

selection mechanisms that might be relevant in different settings.

There are two classes of mechanisms through which $\text{corr}((\omega(\alpha X + \gamma Z) - C), \gamma Z)$ could be non-zero. One is driven by selection on gain: the mechanical correlation between $\omega\gamma Z$ and γZ . If potential partners care about outcomes when they decide whether to implement a new program or evaluate an existing program, and if they have some unobserved information about outcomes, then actual partners will be selected on unobservables. One specific mechanism within this class is "targeting": potential partners that think their populations would be more likely to benefit from a new intervention would be more likely to choose to implement it. Similarly, intervention designers such as Opower may want to immediately showcase the efficacy of a new program, giving them incentives to focus initial partner recruitment on sites with particularly responsive populations. A related mechanism results from the fact the "it pays to be ignorant," as argued by Pritchett (2002). Because rigorous evaluations are publicized and affect funding from foundations and governments, potential partners that believe their existing programs are effective may be more willing to have them rigorously evaluated, while those that believe they are running ineffective programs strategically choose to avoid randomized evaluations. The cost of RCTs plus other sources of imperfect information about program quality keep this equilibrium from unraveling into an equilibrium in which RCTs are run at all sites.

The second class of mechanisms is driven by selection on cost: the potential correlation between C and γZ . If the net costs of running an RCT are positively or negatively correlated with unobservable moderators of the treatment effect, then actual partners will differ from non-partners on unobservables. Several mechanisms in this category result from the fact that implementing randomized trials requires managerial ability and operational effectiveness. One mechanism is a form of "ability bias": potential partners that are most able to implement RCTs, and thus have low C , also implement the intervention itself most effectively. This form of positive partner selection bias is related to the idea of "gold plating" (Duflo, Glennerster, and Kremer 2008): in order to cleanly measure efficacy, treatments are often implemented with much greater precision and quality in experimental settings than they would be elsewhere.²

Many types of organizations run multiple programs: hospitals offer a variety of patient services, utilities run many different energy efficiency programs, and social services centers might offer health clinics, translation, and job training. Able and effective RCT partners may also offer more or better programs in addition to the intervention being evaluated. If these other programs are complements to the intervention, then this causes positive site selection bias. On the other hand, these additional programs could be substitutes that address the same outcome, and there may be diminishing returns to additional interventions. This would cause negative site selection bias.

²"Ability bias" and gold plating are issues of treatment fidelity (differences in the treatment itself) instead of differences in population responses to the same treatment. Taken literally, the theoretical framework above captures only the latter. To accommodate these ideas in the existing framework, one could use T to represent treatment by a "class of interventions" differentiated by a quality measure Q , and Q could be an element of X or Z .

2.6 The Magnitude of Site Selection Bias

What is the magnitude of site selection bias? More precisely, how much does the expected Sample ATE differ on unobservables compared to the expected ATE in the population of Target sites \mathcal{R} ? This can be derived analogously to Heckman's (1979) exposition of individual-level selection bias. For this derivation, assume that γZ_r and $\omega\tau_r - C_r$ are jointly normally distributed in the population of sites, with standard deviations $\sigma_{\gamma Z}$ and $\sigma_{\omega\tau - C}$, respectively, and correlation coefficient ρ . I define $\psi = E[\omega\tau_r - C_r]$ and, without loss of generality, impose that $E[\gamma Z_r] = 0$, because X can include a constant. If selection is governed by Equation (7a), then the expected ATE in the Sample sites is:

$$E[\tau_r | (X_r, D_r = 1)] = \alpha E[X_r] + E[\gamma Z_r | \omega\tau_r - C_r > 0] = E[\tau_g] + \sigma_{\gamma Z} \cdot \rho \cdot \frac{\phi\left(\frac{-\psi}{\sigma_{\omega\tau - C}}\right)}{\Phi\left(\frac{\psi}{\sigma_{\omega\tau - C}}\right)} \quad (8)$$

This equation shows that the expected ATE in the Sample sites is the expected Target ATE $E[\tau_g]$ plus an additional term, which reflects partner selection bias from unobservables. Unobservable partner selection bias is more severe when $\sigma_{\gamma Z}$, ρ , or $\frac{\phi\left(\frac{-\psi}{\sigma_{\omega\tau - C}}\right)}{\Phi\left(\frac{\psi}{\sigma_{\omega\tau - C}}\right)}$ is large. What does this mean from a practical perspective?

When $\sigma_{\gamma Z}$ is large, this means that there is significant variation in treatment effects across sites that cannot be explained by observables. On the other hand, as $\sigma_{\gamma Z}$ approaches zero, there will be no partner selection on unobservables, even if there is selection on observables. This motivates our empirical test in the next section of the extent of explained variation in treatment effects across Opower sites.

The correlation coefficient ρ is large when selection mechanisms such as the examples discussed above are stronger. This occurs when ω is large relative to C , meaning that there is powerful selection on expected ATEs, or when costs C are highly correlated with γZ . On the other hand, partner selection bias would not be severe if selection is largely driven by costs and costs are uncorrelated with unobservables that moderate the treatment effect. In the extreme, one could imagine a "natural experiment" in which sites choose to run RCTs due to costs and benefits that are fully independent of Z .

The inverse Mills ratio $\frac{\phi\left(\frac{-\psi}{\sigma_{\omega\tau - C}}\right)}{\Phi\left(\frac{\psi}{\sigma_{\omega\tau - C}}\right)}$ is a monotonically decreasing function of $\frac{\psi}{\sigma_{\omega\tau - C}}$. When $\psi = E[\omega\tau_r - C_r]$ is small, meaning that the net costs of being a partner are large, then only a few sites will elect to be partners. The sites that do become partners would be more likely to have large draws of the unobservable γZ , implying more severe partner selection bias. On the other hand, when the average net benefit of experimentation ψ is large, then many sites will elect to be partners, and partner selection bias is not severe. Therefore, as with individual-level selection into treatment, the ratio of the number of Sample to Target sites is a useful diagnostic. With Opower

and in many other contexts, only a small number of sites that theoretically could run RCTs actually do.

3 The Opower Experiments

3.1 Experimental Design

Opower is a private company that partners with utilities to mail Home Energy Reports to residential electricity and natural gas consumers. Utilities partner with Opower for several reasons. Most importantly, there are 27 states with Energy Efficiency Resource Standards (EERS), which require utilities to run programs that reduce energy use relative to counterfactual by a given amount, often on the order of one percent per year. In the absence of an EERS or other regulatory mechanism, for-profit investor-owned utilities (IOUs) do not have the incentive to reduce demand for the product they sell. Rural electric cooperatives and other utilities owned by municipalities or other government entities often run energy conservation programs if they believe they can benefit customers. Aside from conserving energy, some utilities also have found that the home energy report program can help improve consumers' positive perception of the utility brand.

To implement a program, Opower and the partner utility first identify a set of residential consumers to target. Some small utilities choose to target the entire residential consumer base, while others target heavy users who might be most responsive to the intervention, and others target local geographic areas where conservation could help to delay costly upgrades to distribution infrastructure. To be eligible for the program, a customer must have at least one year of valid pre-experiment energy use data and satisfy some additional technical conditions.³ The resulting site-level population is then randomized into treatment and control groups.

Figure 1 shows an example report. The two-page letter has two key components. The Neighbor Comparison Module on the first page compares the household's energy use to its 100 geographically-nearest neighbors that have similar house sizes and heating types. The Action Steps Module, which is typically on the second page, includes energy conservation tips targeted to the household based on its historical energy use patterns and observed characteristics.

The treatment group is sent reports at frequencies that vary within and between households and sites. For example, two early programs randomized treated households into either monthly or quarterly frequencies. Several other early programs targeted higher users with more frequent

³Typically, households in Opower's experimental populations need 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 to construct the neighbor comparisons. Households that have special medical rates or photovoltaic panels are sometimes also excluded. Utility staff and "VIPs" are sometimes automatically enrolled in the reports, and I exclude these non-randomized report recipients from any analysis. These technical exclusions eliminate only a small portion of the potential population. These exclusions do not contribute to site selection bias if one believes that the excluded households would never receive the intervention and are thus not part of a Target population.

reports. More recently, a common structure is three consecutive monthly reports followed by bimonthly.

Although the reports' basic structure is highly consistent, with two pages of neighbor comparisons, additional personalized energy use information, and energy conservation tips, the reports do vary. The envelope and the report are branded with each local utility's name, and the information and tips are updated each month to reflect the customer's most recent energy bills and seasonal factors - for example, customers are more likely to see information about air conditioners in the summer. Despite this variation, there is a remarkably high degree of treatment fidelity compared to other treatments of interest in economics. For example, "job training" often takes different forms at different sites (Dehejia 2003, Hotz, Imbens, and Klerman 2006), and the quality of "remedial education" could depend markedly on the teacher's ability.

Aside from treatment fidelity, there are two other useful features of the Opower experiments. First, in the taxonomy of Levitt and List (2009), these are "natural field experiments," meaning that people are in general not aware that they are being studied. Therefore, there are no Hawthorne effects. Second, because opting out of the letters requires active effort, there is very little non-compliance. This means that there is no need to model essential heterogeneity or the individual-level selection into the experimental treatment (Heckman, Urzua, and Vytlačil 2006), and the treatment effect is a Policy-Relevant Treatment Effect in the sense of Heckman and Vytlačil (2001).

3.2 Data

I use three kinds of data: characteristics of the population of potential utility partners, site-level metadata from all Sample sites, and household-level microdata from a subset of sites.

3.2.1 Site-Level Characteristics

Part of the analysis will model the selection of electric utilities into partnership with Opower. I define the population of potential partner sites \mathcal{R} to include the 882 large electric utilities in the United States. This excludes small utilities with fewer than 10,000 residential consumers and power marketers in states with deregulated retail markets, as Opower has no clients in these two categories. About five percent of utilities operate in multiple states. In order to model how state and local policies affect utilities' decisions, a utility is defined as a separate observation for each state in which it operates.

Opower has 63 partners where programs have begun as of July 2013, meaning that the selection probability is approximately seven percent. Recall from Equation (8) that, other things equal, lower selection probabilities result in larger site selection bias.

Table 2 describes the dataset of utility characteristics that I hypothesized might be associated with the selection process. The primary source of these data is the Energy Information Administration (EIA) Form 861 for calendar year 2007, the year before the first Opower programs began

(U.S. EIA 2013). From these data, I construct each utility's ownership structure (investor-owned, municipal, or other, which includes rural electric cooperatives and other government entities such as the Tennessee Valley Authority), number of residential consumers, average residential electricity usage and price, and the share of consumers that have voluntarily enrolled in "green pricing programs" that sell renewably-generated energy at a premium price. I also construct two measures of the extent of other existing energy efficiency programs: the ratio of estimated electricity conserved in residential energy conservation programs to total residential electricity sold and the ratio of total spending on energy conservation programs to total revenues.

Form 861 includes a list of the counties in each utility's service territory, which can be matched to county-level demographic information. Using the U.S. Election Atlas (Leip 2013), I construct the share of all votes in the 2004 and 2008 presidential elections that were for the Green party candidate, as well as the share of Democratic and Republican votes in those elections that were for the Democratic candidate. County mean income per capita is from the U.S. Bureau of Economic Analysis (2013). From the 2010 U.S. Census, I gather the share of the county's population residing in urban areas (metropolitan statistical areas) and the percent of the population over 25 years old that has a college degree. In a handful of cases (primarily in Alaska) where counties could not be matched between datasets, I substituted state-level averages. Finally, I include whether the state in which the utility operates has a Renewables Portfolio Standard (RPS), which requires utilities to procure a certain proportion of electricity from renewable sources, or an Energy Efficiency Resource Standard (EERS). The RPS data are from U.S. Department of Energy (2011), while the EERS data are from the Pew Center (2011).

The first three columns of Table 2 present means and standard deviations for the population of utilities, Opower partners, and non-partners, respectively. The fourth column tests whether the means differ between the two groups. Notice that this table is structured similarly to commonly-presented tables that provide suggestive evidence of internal validity by comparing observable characteristics of treatment and control groups. Twelve out of 15 are unbalanced with more than 90 percent confidence, and an F-test easily rejects the hypothesis that the observables are jointly uncorrelated with partner status. Opower's partners clearly differ on observables.

Figures 2a-2d presents geographical intuition for the site selection process. Figure 2a shows that Opower partner utilities are concentrated along the west coast, the upper midwest, and the north-east. Figure 2b highlights states that have either a quasi-governmental energy efficiency agency (Maine, Vermont, Oregon, and Hawaii) or an Energy Efficiency Resource Standard. The extremely high degree of overlap suggests that either EERS regulations are important immediate causes of partnership or underlying variation in concern for environmental issues drives both EERS regulations and utility management interest in the intervention. Figure 2c presents one potential correlate of this: state-level Democratic vote shares. Democratic states are more likely to have Opower sites, and more strongly Democratic states (California, Washington, New York, and Massachusetts in particular) tend to have started programs relatively early. Figure 2d illustrates a negative cor-

relation which will also become important: Opower sites tend to be in areas with lower average electricity usage. In particular, southern states with high summer air conditioning demand are less likely to have Opower sites, and the programs that have been implemented started relatively recently.

3.2.2 Site-Level Metadata

Due to contractual restrictions, Opower cannot share microdata from some recent partners. Instead, I observe their site-level metadata, including average treatment effects and standard errors, number of reports sent, and attrition for each post-treatment month of each RCT. Consistent with the theoretical framework, I define a "site" as a group of households where one experiment takes place. Some utilities have multiple "sites," because they began with one customer sub-population and then added other sub-populations in separate randomized control trials at a later date. There are 98 sites with at least one year of post-treatment data at 52 different utilities.

For this analysis, I study only the first 12 post-treatment months at each site. Considering full instead of partial years averages over seasonal differences in ATEs, whereas comparing programs that have been in effect over different seasons would require location-specific seasonal adjustments. Comparing programs that have been in effect for different durations would require duration controls, given that effect sizes tend to grow over time (Allcott and Rogers 2013). Using one year instead of two or more full years allows the analysis to include the largest number of sites. This comes at little cost in terms of precision: although the standard errors are somewhat wider, the one-year ATE explains 94.5 percent of the variation in the two-year ATE.

Opower's analysts estimated the ATEs using mutually-agreed procedures and code. I define m_0 as the month when the first reports are generated. The 12 months before m_0 are the "baseline" period, while the "post-treatment" period begins the first day of the month after m_0 . The month m_0 is excluded from the analysis, as it often will include days both before and after the first reports arrive. Y_{irt} is daily average electricity usage (in kilowatt-hours per day) for household i in site r for the period ending in date t . This comes from meter reads, which for most households happen approximately once each month. Y_{0ir} is a vector of three baseline usage controls: average daily usage over the entire baseline period, the baseline winter (December-March), and the baseline summer (June-September). π_{rm} is a set of month-of-sample indicators. The first year ATE is estimated using the following equation:

$$Y_{irt} = -\tau_r T_{ir} + \phi_{rm} Y_{0ir} + \pi_{rm} + \varepsilon_{irt} \quad (9)$$

The intervention causes a decrease in energy use. By convention, I multiply τ_r by -1 in Equation (9), so that reported τ are positive and larger values imply better efficacy.

Table 3 presents descriptive statistics for the metadata. The 98 site-level populations average about 70,000 households, of which an average of 47,000 are assigned to treatment. The total sample

size for this analysis is thus approximately 6.8 million households, or about one in every 14 in the United States. Control group post-treatment average usage ranges from 12.5 to 72.9 kilowatt-hours (kWh) per day. For context, one kilowatt-hour is enough electricity to run either a typical new refrigerator or a standard 60-watt incandescent lightbulb for about 17 hours. The average U.S. household consumes 11,280 kWh/year, or 30.9 kWh/day (U.S. EIA 2011). The ATEs also vary substantially, as I will discuss in the next section.

There are two types of attrition. First, an average of 10 percent of households move and close their utility accounts each year. The sites with the two highest one-year move rates (31 and 52 percent) are both within a utility in college town where most households are rentals that change hands each academic year. After an account closes, Opower ceases to send reports, and we no longer observe energy bills, so the households attrit from the sample.

The second type of attrition is when a household actively calls the utility and asks to opt out of the program. An average of 0.5 percent of households opt out during the first year. These households' utility bills are observed, and they remain in the sample. I define the "treatment" as "being mailed a Home Energy Report or actively opting out." This definition of "treatment" gives a treatment effect of policy interest: the effect of attempting to mail Home Energy Reports to an entire site-level population. In practice, because opt-out rates are so low, the ATE is the almost exactly the same when the "treatment" is defined as "being mailed a Home Energy Report."

3.2.3 Microdata

In addition to the metadata, I also have household-level microdata through the end of 2010 for all Opower programs that began before the end of 2009. Table 4 provides an overview. Due to confidentiality restrictions, utility names and locations are masked and the sites are numbered from one to nine. The dataset includes 20.5 million electricity meter reads from 495,000 households at nine sites, of which 5.2 million occur in the first year post-treatment. The rightmost column shows that treatment and control groups at all sites are statistically balanced on baseline usage; a tenth site is excluded due to mild imbalance.

Opower, and the utilities they work with, gather demographic data for each customer from surveys, public records, and private-sector marketing data providers. In this analysis, I consider a subset of variables that theory suggests could moderate the treatment effects and might also vary across sites. I focus on four potential moderators. The first is scale: higher levels of energy demand could allow larger quantities of energy conserved. The second is social norms: social inference, conditional cooperation, and conformity suggest that households who learn that they use more energy than the norm should conserve more energy than those who learn that they use less.

The third moderator is increasing marginal cost of conservation: households that have already participated in other energy efficiency programs may have already exploited the lowest-cost energy conservation opportunities, and any additional opportunities may be more costly. For example, a natural way that consumers might respond to the Opower treatment is to be more assiduous

about turning off lights when not in use. Many utilities also run programs to replace standard incandescent lightbulbs with energy efficient Compact Fluorescent Lightbulbs (CFLs). Because CFLs use one-fourth the electricity of an incandescent, a household that has participated in one of these programs and then responds to the Opower treatment by turning off the lights would save one-fourth the electricity of household that still had incandescents.

The final moderator is population preferences: some households are more environmentally conscious or interested in conservation, which might make them more responsive to the intervention. For example, Costa and Kahn (2012) show that households that vote Democratic, donate to environmental groups, or pay more for green energy have stronger treatment effects. The working paper version also documents that physical house features such as size and use of electric instead of oil or gas heat moderate the treatment effect, but I omit them here because they are not central to the site selection bias story.

Table 5 presents the means and standard deviations of variables at each site. The first three columns represent scale. The first two columns are heating and cooling degree-days, which measure how much temperatures deviate from 65 degrees during each meter read period, and thus how much electricity might be required to heat or cool a house to a comfortable temperature.⁴ Baseline Usage is the mean usage in kilowatt-hours per day over the baseline period. Most site-level averages are close to the national average of 30 kWh/day. First Comparison is drawn directly from the Social Comparison Module: it is the usage difference in kWh/day between household i and its mean neighbor on the first comparison report. (This is also observed for the control group because Opower generates placebo reports for these households.)

The next two variables are only available in site number 9. EE Program Participant is an indicator for whether the household had received a loan or rebate for an energy efficient appliance, insulation, or a heating, ventilation, and air conditioning system through another utility program before the Opower program began. This may also be associated with population preferences. A more direct measure of preferences is an indicator for whether the household had enrolled in the utility's green pricing program before the Opower program began. Wealth may also be associated with preferences for environmental conservation. As a measure of wealth, I include the home's assessed value, which is observed in four sites.

⁴More precisely, the average Cooling Degree-Days for an observation 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 degrees. A day with average temperature 75 has 10 CDDs, while a day with average temperature 30 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 degrees and the day's average temperature. A day with average temperature 75 has zero HDDs, while a day with average temperature 30 has 35 HDDs.

4 Treatment Effect Heterogeneity

For there to be site selection bias, the same factors that moderate treatment effects must also be associated with selection. Above, I hypothesized that four classes of variables might moderate treatment effects. The first part of this section empirically tests these hypotheses using the microdata.

Strong External Unconfoundedness is a stronger assumption than External Unconfoundedness in Distribution. In other words, if there is no site-level heterogeneity, there can be no site selection bias. Furthermore, Equation (8) shows that other things equal, wider dispersion of site-level heterogeneity implies larger site selection bias. This motivates the second part of this section, which documents the dispersion of site-level treatment effects.

4.1 Heterogeneous Treatment Effect Estimates Using Microdata

Heterogeneous treatment effects can be estimated by adding controls for observables X to Equation (9):

$$Y_{irt} = -(\alpha X_{irt} + \mu_r)T_{ir} + \beta_r X_{irt} + \phi_r Y_{0ir} + \pi_{rm} + \varepsilon_{irt} \quad (10)$$

As in the model, the α parameters capture how observables X moderate the treatment effect. The variable μ_r reflects site-level unexplained heterogeneity. The equation pre-multiplies $(\alpha X_{irt} + \mu_r)$ by -1 to maintain the convention that more positive effects imply better efficacy. The equation also includes site-specific controls for the main effects of X , baseline usage, and month-of-sample. Standard errors are robust and clustered by household.

Table 6 presents the results. Column 1 estimates the association between report frequency (in number of reports per month) and the ATE. This uses only data from site numbers 2 and 6, which randomized households into receiving reports either every month or every quarter. The coefficient in column 1 implies that 1 report per month compared to 1/3 report per month increases energy conservation by $0.163 \cdot (1 - 1/3) \approx 0.109$ kWh/day. This is consistent with other (unreported) regressions which show that monthly instead of quarterly reports increase impacts by 0.10 kWh/day in site 2 and 0.12 kWh/day in site 6. The coefficient in column 1 will be useful in controlling for frequency differences when comparing ATEs across different sites.

Column 2 controls for the weather and energy use variables, which are available in all nine sites. Although the β coefficients for heating degrees are positive (colder weather is strongly associated with more electricity demand), heating degrees are not statistically associated with the treatment effect in these nine sites. Cooling degrees, which primarily reflect electricity demand for air conditioning, are associated with larger treatment effects. The coefficient estimate suggests that increasing daily average temperatures from (for example) 65 to 75 degrees increases the Opower conservation effect by 0.20 kWh/day. This is a very large effect given that the average first-year τ_r

reported in Table 3 is 0.43 kWh/day. This suggests that implementing the program in sites with hotter summers could meaningfully improve efficacy.

Column 2 also includes interactions of Baseline Usage and First Comparison with the treatment indicator. Because these variables are highly correlated, it is important to include them both in this regression in order to separately interpret their conditional correlations. The coefficients show that households with baseline electricity use that is (for example) 10 kWh/day higher have treatment effects that are 0.1 kWh/day larger. Households that are informed on their first comparison report that their energy use is 10 kWh/day larger than their mean neighbor have treatment effects that are 0.21 kWh/day larger. These coefficients are simply conditional correlations, and they do not identify the causal impact of independently manipulating the perceived social norm. However, these coefficients do make clear that conservation effects are economically significantly larger for households that use more energy.

The third column uses only data from site 9, where green pricing and energy efficiency program participation are observed. Green pricing program participation is conditionally associated with 0.23 kWh/day larger treatment effects. This suggests that consumers with environmentalist preferences have stronger treatment effects. This is consistent with the previous work of Costa and Kahn (2012) and will play an important role in one potential site selection mechanism. Previous energy efficiency program participation is not statistically significantly associated with the treatment effect, although the standard error is too wide to reject an economically large association. This regression also includes a control for the association between T and Baseline Usage, in order to ensure that the preference-related coefficients do not act entirely through an association between environmentalism and baseline usage. When the T and Baseline Usage interaction is excluded, the Green Pricing indicator is much more strongly associated with the treatment effect.

Columns 4 and 5 test the association between house value and the treatment effect, using data from the four sites where this is observed. Results show that house value is unconditionally associated with the treatment effect, but the statistical association goes away when conditioning on the interaction of T and Baseline Usage.

Because the weather-related X_{irt} variables vary within household over time but do not vary much across households within a site, the α for weather are not identified if the data are collapsed to a single post-treatment observation for each household. However, the α and standard errors for time-invariant household characteristics are almost exactly identical if estimated with collapsed data. Because the sample is so large and the randomizations are balanced, the results are also not sensitive to alternative specifications such as including month-specific ϕ_r , excluding Y_{0ir} , excluding month-of-sample indicators, or using a fixed effect estimator with pre- and post-treatment data. The R^2 is so high because baseline usage and month-of-sample controls explain much of the variation in Y : a regression of Y_{irt} on $\phi_r Y_{0ir}$ using the full sample in column 2 has $R^2 = 0.57$, and adding π_{rm} increases the R^2 to 0.66.

4.2 Dispersion of Site Effects

Figure 3 illustrates the density of estimated average treatment effects across sites. Some of this variation is due to sampling error. The dotted line reflects the distribution that would be observed if the ATEs all had the same true value and the estimates differed only due to sampling variation. This was calculated by simulating draws from the sampling distribution of each site's estimates around a "true" coefficient of 0.43 kWh/day, which equals the mean across sites reported in Table 3. As illustrated by the solid line, the unconditional distribution of estimated effects has much more variation than can be explained simply by sampling error. The standard deviations of the sampling error and unconditional distributions are 0.055 and 0.22 kWh/day, respectively.

The longer-dashed line plots the distribution of treatment effects after adjusting for different average frequencies across sites. The Frequency-Adjusted ATE $\tilde{\tau}_r$ is $\beta_{\text{Reports/Month}} \cdot \overline{(\text{Reports/Month} - \text{Reports/Month}_r)}$, where $\beta_{\text{Reports/Month}} = 0.163$ kWh/day, from Column 1 of Table 6, and the average reports per month across sites $\overline{(\text{Reports/Month})}$ is 0.55, as reported in Table 3. This explains little of the variation, reducing the standard deviation to 0.21 kWh/day. The shorter-dashed line plots the distribution of frequency-adjusted treatment effects after conditioning linearly on each site's average control group usage. Usage does explain a meaningful part of the impact dispersion across sites, reducing the standard deviation of $\tilde{\tau}$ from 0.21 to 0.16.

Is this variation economically significant? I consider two measures of economic significance: predicted effects at scale and cost-effectiveness. Variation in predicted effects at scale is particularly relevant for Opower because policymakers are setting state and national energy efficiency requirements and want to know how much impact could likely be expected from "behavior-based" programs like Opower. Figure 4 presents the national-level total predicted electricity cost savings if the program were expanded to all US households. Each dot reflects the prediction using the ATE from each observed site, multiplied by 100 million households and using a national average electricity price of 10 cents/kWh. The predicted savings vary by a factor of 14, from \$271 million to \$3.73 billion per year.

Figure 5 presents the variation in cost effectiveness, with each dot analogously representing one site's cost effectiveness. While there are many ways to calculate this statistic; I present the simplest: the ratio of program cost to kilowatt-hours conserved during the first two years. Notice that by convention in evaluating energy efficiency programs, lower cost effectiveness is better. I make a boilerplate cost assumption of \$1 per report. The variation is again quite substantial. The most cost effective (0.54 cents/kWh) is 21 times better than the least cost effective, and the 90th percentile is 2.8 times worse than the 10th percentile. The site on the right of the figure with outlying poor cost effectiveness is a small program with extremely weak ATE and moderately frequent reports. The dispersion is somewhat wider when calculating cost effectiveness using only the programs' first year.

This variation is economically significant in the sense that it can cause program adoption errors: program managers at a Target site might make the wrong decision if they extrapolate cost

effectiveness from another site to that Target in order to decide whether to implement the program. Alternative energy conservation programs have been estimated to cost approximately five cents per kilowatt-hour (Arimura, Li, Newell, and Palmer 2011) or between 1.6 and 3.3 cents per kilowatt-hour (Friedrich *et al.* 2009). These three values are plotted as horizontal lines on Figure 5. Whether an Opower program at a new site has cost effectiveness at the lower or upper end of the range illustrated in Figure 5 therefore could change whether a manager would or would not want to adopt. Extrapolating cost effectiveness from other Sample sites could lead a Target to implement when it is in fact not cost effective, or fail to implement when it would be cost effective. As a concrete example, I note that in one early site with a small ATE and therefore poor cost effectiveness, the partner utility ended the program.

5 Site Selection Bias

This section studies site selection bias in three parts. First, I propose a specific set of selection mechanisms. Second, I study how utilities select into partnership with Opower, testing whether partners differ from non-partners on observables that moderate treatment effects. Third, I study the timing of selection, testing whether early sites differ from later sites.

5.1 Potential Site Selection Mechanisms

In the Opower setting, site selection occurs on two levels. First, Opower partners with a utility. It is simplest to think of utility management as the agent in this decision, because as a for-profit company, Opower will partner with any utility client that chooses to pay. However, one could also think of a partnership as an equilibrium result, as Opower makes pricing decisions and can allocate sales effort. Second, Opower and the utility choose a population of residential consumers within the utility service territory. Different selection mechanisms operate at each level, and both levels contribute to the eventual choice of sites from the nationwide population of utility consumers.

As illustrated in Equation (7a), a site selection mechanism is a process through which factors that moderate treatment effects are also associated with the partner’s decision to contract with Opower. I focus on five specific mechanisms suggested by theory, the above empirical correlations, and anecdotal evidence.

- **Usage Targeting.** Theory predicts, and initial evaluation reports results showed, that heavier electricity users have larger treatment effects. This made utilities with higher average usage, as well as within-utility sub-populations with relatively high usage, more likely to have cost effective results.
- **Population Preferences.** Some site-level populations have stronger preferences for environmental conservation. Utility managers from these conservationist areas might be more likely

to adopt the program, and their customer base might be more responsive. This is consistent with results in the previous sections, and it would generate positive selection.

- **Complementary or Substitute Programs.** Utilities that place a priority on energy efficiency programs in general may be more likely to adopt the Opower program. The utility’s other programs could be complements, because one way that consumers respond to the Opower treatment is by participating in other energy efficiency programs such as appliance rebates or weatherization (Allcott and Rogers 2013). The more effective these other programs are, the larger the Opower treatment effect. These programs could also be substitutes, because the marginal program could have diminishing returns.
- **Partner structure.** Larger utilities have economies of scale in implementing the Opower program. For example, the utility faces fixed costs of management time to implement and statistically evaluate the program, regardless of the number of households involved. Separately, Energy Efficiency Resource Standards are more likely to apply to investor-owned utilities, and EERS policies are key drivers of program adoption. Customers of large utilities and of IOUs may be more or less engaged with their utility, and more or less likely to read and react to mail from the utility.
- **Price.** Where electricity prices are higher, the program is potentially more cost effective from the utility’s perspective, making management more likely to adopt the program. Simultaneously, consumers may be more responsive to the intervention because energy conservation has higher returns.

5.2 Selection into Partnership

5.2.1 Procedure

The test of partner selection on observables has two steps. First, I estimate a probit model of utility selection into partnership with Opower, using the data from Table 2:

$$\Pr(D_r = 1|W_r) = \Phi(\rho W_r) \tag{11}$$

In this equation, Φ is the CDF of the standard normal distribution, and W_r is a vector of utility-level characteristics. Standard errors are robust. Observations are weighted by the utility’s number of residential consumers. This is because external unconfoundedness in distribution is a statement about individual i ’s probability of being in a treated site, not site r ’s probability of being treated. Intuitively, a set of Sample sites can still be highly representative of the overall population of individual units even if there are a large number of non-Sample sites that have only a small number of individuals.

I then regress the frequency-adjusted ATEs from the 98 observed sites on the fitted selection probabilities:

$$\tilde{\tau}_r = \theta \widehat{\Pr}(D_r = 1|W_r) + \eta + \varepsilon_r \quad (12)$$

Recall that some utilities have multiple "sites," i.e. multiple distinct RCTs. Each of the 98 observed sites is an observation, but standard errors are clustered by utility to reflect the fact that $\Pr(D_r = 1|W_r)$ varies only at the utility level. Standard errors are then further adjusted to reflect uncertainty in the first-step estimate of $\widehat{\Pr}(D_r = 1|W_r)$, per Murphy and Topel (1985). If $\theta > 0$, the utilities with higher selection probabilities given a particular set of W_r have larger ATEs, giving positive site selection through that mechanism. If $\theta < 0$, this implies negative partner selection on a particular set of observables.

5.2.2 Results

Table 7 presents the tests of utility-level site selection on observables. The top panel presents the first-step estimates of Equation (11), while the bottom panel presents the second-step estimates of Equation (12). I group site-level variables into distinct subsets that reflect each of the five proposed site selection mechanisms.

Column 1 fits the selection equation using Utility Mean Usage as the only W . This tests the usage targeting mechanism. The probit estimates show that higher-usage utilities are actually *less* likely to partner with Opower. This is consistent with the maps in Figure 2, which show that Opower has few partners in southern states with high summer electricity demand. However, higher-usage utilities do have larger treatment effects, so the $\hat{\theta}$ coefficient in the bottom panel implies negative selection on this variable. Column 6 shows that Utility Mean Usage remains negatively associated with partnership even after conditioning on other observables. These results suggest that there are some other unobserved regulatory or preference-related factors associated with Utility Mean Usage that overwhelm any potential usage targeting mechanism. For example, when Column 6 is re-estimated with four census region indicator variables, the $\hat{\rho}$ for Utility Mean Usage becomes statistically insignificant, while the other $\hat{\rho}$ remain statistically the same.

Figure 6a plots the data and fitted regression line for Equation (12) using the selection probabilities from column 1. The downward slope reflects a negative $\hat{\theta}$ and negative selection on Utility Mean Usage. At the right of the graph is a mass of low-usage utilities, none of which have large ATEs.

Column 2 tests the population preferences mechanism by including a set of W that should be correlated with environmental preferences and interest in energy conservation. While four of the six are unconditionally correlated with selection, they are correlated, so their conditional correlations are less likely to be statistically significant. Jointly, they are very powerful in predicting partnership. The estimates of Equation (12) in columns 2-6 also condition on the control group's mean post-

treatment usage, which effectively gives treatment effects in percent terms instead of kilowatt-hour levels. The estimated $\hat{\theta}$ in the bottom panel implies strong positive selection on population preferences. Figure 6b is analogous to Figure 6a, except using the predicted selection probabilities from column 2 and partialling out Control Mean Usage.

Column 3 tests the complementary or substitute programs mechanism. Both measures of pre-existing energy efficiency programs are positively associated with selection. The point estimate of $\hat{\theta}$ is positive, but not statistically significant. It is possible that these variables partially capture selection on population preferences along with the effects of complementary or substitute programs. Columns 4 and 5 test the partner structure and price mechanism, finding no evidence of site selection bias through these mechanisms. Column 6 tests for selection on the full set of observables. The point estimate is positive but not statistically significant, meaning that although Opower partners differ on these observables, the more likely partners conditional on these observables do not have different treatment effects than the less likely partners.

Appendix Table A1 presents two types of additional results. First, I estimate the two selection equations using each of the 15 variables individually. Only two give statistically significant negative site selection bias: Utility Mean Usage, which we saw in column 1 of Table 7, and the investor-owned utility indicator. Eight variables, many of which could be associated with population preferences, give positive partner selection bias. The second additional result is to estimate the two selection equations 15 times, each time with all variables other than one of the observables. None of the $\hat{\theta}$ coefficients change statistically from column 6 of Table 7, suggesting that no one individual variable drives the overall results.

5.3 Early vs. Late Adopters

An alternative approach to understanding site selection bias is to exploit the timing of program adoption, looking at early vs. late adopters. Formally, think of the set of eventual sites as the target sites, while the early adopter sites are the sample. One benefit of this approach relative to using predicted selection probabilities is that we can also observe selection on unobservables, given that we actually observe treatment effects in the full set of target sites.

Figure 7 illustrates the main result. Every single one of the 14 programs that began before June 2010 had a frequency-adjusted first year ATE of 0.36 kWh/day or larger. Forty four of the 84 programs that began on or after that date had frequency-adjusted first year ATE less than that value. There continue to be programs with high efficacy, including some with larger ATEs than at any of the early sites. Overall, however, there is a statistically significant downward slope: early adopter sites had larger ATEs than later adopters.

Denote the site start date (in days after January 1, 2008, divided by 365 to normalize to years) as S_r . Table 8 breaks down this trend into within vs. between utility trends, estimated using the following equation:

$$\tilde{\tau}_r = \theta S_r + \eta + \varepsilon_r \quad (13)$$

Column 1 is the regression version of Figure 7, showing that frequency-adjusted ATEs average 0.046 kWh/day lower for sites that start one year later. Column 2 conditions on the control group mean usage. Because later sites have higher control group mean usage and this is associated with stronger ATEs, the unexplained drop in ATEs is faster: 0.070 kWh/day each year. Columns 3 and 4 look only at each utility’s first site, which is one way to test for trends only between utilities. The time trends are still downward, although the unconditional trend in column 3 is not statistically significant.

Columns 5 and 6 look only within utility, conditioning on utility fixed effects and estimating the association between ATE and the order of start dates within the utility. For example, column 5 shows that the second start date relative to the first has 0.041 kWh/day smaller ATEs. Figure 8 illustrates this regression, showing a scatterplot of $\tilde{\tau}$ demeaned by utility against the within-utility start number. Column 6 shows that this positive selection result goes away when conditioning on usage, suggesting that the primary difference between early and later sites is due to targeting of initial programs toward high usage populations.

What explains the trend from column 4 of lower ATEs conditional on control group usage for utilities’ initial sites? To test this, I implement a two-step procedure similar to the partner selection estimates above. In the first equation, I predict the utility’s start date based on observables W_r :

$$U_r = \rho W_r + \epsilon_r \quad (14)$$

The second equation takes the frequency-adjusted ATE and regresses it on the site’s start date, controlling for \hat{U}_r , the utility-level start date predicted by observables:

$$\tilde{\tau}_r = \lambda S_r + \theta \hat{U}_r + \eta + \varepsilon_r \quad (15)$$

Standard errors are robust and clustered by utility, and they are further adjusted to account for uncertainty in the first step fit of \hat{U}_r , using formulas from Murphy and Topel (1985).

Table 9 presents the results, using format analogous to Table 7. Two basic insights arise that parallel the earlier results. First, early utilities have lower mean usage, implying negative selection. Second, some positive selection exists based on preferences. However, there is still unexplained selection on observables: none of the predicted selection times \hat{U}_r fully explain the downward trend in ATEs conditional on control mean usage.

6 Site Selection in Other Contexts

Of course, the Opower experiments are only one example in one setting. In this section, I examine two other contexts of broad interest: microfinance and clinical trials. I follow the same procedure

in each case. First, I define a population of target sites. Second, I gather site-level observables that could moderate the effects of different interventions. Third, I compare sample to non-sample sites on these potential moderators. Unlike with Opower, the interventions vary, so it is not possible to take the next step of correlating selection probability or timing with a consistently-defined treatment effect. Thus, this section is merely intended to briefly present suggestive evidence that site selection bias is not unique to the Opower context.

6.1 Microfinance

In the past ten years, there have been a large set of microfinance field experiments with many partner microfinance institutions (MFIs). These include studies of the effects of group vs. individual liability on default rates (Gine and Karlan 2010), using different variation in interest rates to identify moral hazard and adverse selection (Karlan and Zinman 2009), and estimating effects of credit availability on income, consumption, and other outcomes (Banerjee, Duflo, Glennerster, and Kinnan 2009). Do the types of MFIs that carry out RCTs differ on observables from other non-partner MFIs?

I define the population of sites \mathcal{R} as all MFIs covered by the Microfinance Information Exchange (MIX). MIX is a global database which includes information on the characteristics and performance of 1903 MFIs in 115 countries. Partner MFIs are defined as all MFIs listed as RCT partners on the Jameel Poverty Action Lab, Innovations for Poverty Action, and Financial Access Initiative websites. Roughly two percent of MFIs listed on MIX have partnered with one of these groups on randomized control trials. I consider characteristics that might be correlated with the outcomes of different field experiments. Average loan balance, percent of portfolio at risk of default, and the percent of borrowers who are female could be correlated with default rates, which are a common outcome variable in microfinance RCTs. The MFI's age, non-profit status, size, staff availability, and cost per borrower could be associated with ability to implement or monitor an intervention.

Table 10 presents the means and standard deviations of these characteristics by partner status. All data are weighted by the MFI's number of borrowers. This means that the statistics are valid for the population of borrowers, not the population of MFIs, consistent with the individual-level definition of external unconfoundedness. The rightmost column presents differences in mean characteristics of partners vs. non-partners. Larger MFIs are over-represented, perhaps because RCTs require larger, well-managed partners and large sample sizes. In some situations, a larger MFI might implement a treatment more or less effectively, or might have more or less trust from borrowers. MFIs with more female borrowers and smaller average loan balances are also over-represented, and both of these factors could be associated with lower baseline default rates. Default rates in many microfinance studies are very low, one wonders if effects of various interventions on default rates would be larger against a larger baseline. Finally, partner MFIs have lower costs per borrower. Overall, partner MFIs differ statistically on four of these eight individual characteristics, and an F-test of a regression of partner status on all characteristics rejects the hypothesis that partners and non-partners do not differ jointly on observables.

6.2 Clinical Trials

What types of hospitals carry out clinical trials for new drugs and procedures? If clinical trials are more common at research hospitals with more urban patient populations, higher-ability doctors, and better medical technologies, and if these factors moderate the efficacy of medical interventions, this would suggest systematic site selection bias.

Wennberg *et al.* (1998) provide a motivating example. In the 1990s, there were two large trials that tested carotid endarterectomy, a surgical procedure which treats hardening of the carotid artery in the neck. In order to participate, institutions and surgeons had to be experienced in the procedure and have low previous mortality rates. After the trials found the procedure to be preferred to alternative approaches, its use nearly doubled. Wennberg *et al.* (1998) use a broader sample of administrative data to show that mortality rates were significantly higher at non-trial hospitals, and for some classes of patients and hospitals, treatment with drugs instead of the surgical procedure might have been preferred.

The database for Aggregate Analysis of ClinicalTrials.gov (AACT) gives comprehensive information on clinical trials registered in the official database operated by the U.S. National Institute of Health. I separately consider two types of clinical trials: "Drug" trials, which includes drugs, biological interventions, and dietary supplements, and "Procedure" trials, which include both surgical and radiation procedures. The site names and zip codes can be matched to the set of hospitals in the U.S., which form the "target" population of hospitals where the interventions will be implemented.

Table 11 compares characteristics of hospitals that have been the site of at least one clinical trial to hospitals that have never hosted a registered trial. Hospital characteristics are drawn from the Medicare Hospital Compare database, the American Hospital Association (AHA) Annual Survey, and the NBER Medicare Provider of Services (POS) files. Appendix I presents more details on data preparation.

The first three rows show that clinical trial sites are at hospitals in urban areas and in counties with higher income and education. Remaining characteristics are grouped using Donabedian's (1988) triad of clinical quality measures: structure, process, and outcomes.

Clinical trial sites have significantly different structures. They take place at hospitals that are significantly larger, in terms of both beds and admissions. Furthermore, trial site hospitals perform many more surgeries per year. This is particularly important in light of evidence from Chandra and Staiger (2007), who show that due to productivity spillovers, surgical interventions are more effective in areas that perform more surgeries. In their conclusion, Chandra and Staiger (2007) point out that this may compromise the external validity of randomized control trials.

Clinical trial site hospitals are much more likely to have adopted electronic medical records, and the average site has five to six more of the 21 advanced technologies identified by US News in their Hospital Quality Rankings. Trial sites also offer three more of the 13 patient services scored by US News. If these technologies and services are complements to drugs, or more likely to surgical and radiation procedures, then these interventions will be less effective at non-trial sites.

Clinical trial sites also differ in the processes they use. They perform 0.16 to 0.19 standard deviations better on five surgical process measures included in the Hospital Safety Score (HSS) methodology. If surgeons' adherence to accepted procedures is associated with better outcomes, surgical procedures will tend to be more effective at trial hospitals compared to non-trial hospitals. On the other hand, patient surveys show that doctors and nurses at trial site hospitals are no better or worse at communication, including explaining medicines and what to do during recovery. Although patients' understanding of how to take a drug might affect adherence, and understanding of what to do during recovery might affect how well people recover from surgical procedures, effects of drugs and procedures would not differ through this channel at trial vs. non-trial hospitals.

The next four measures in the table capture outcomes. The measures are only slightly correlated, usually with correlation coefficients under 0.1, which likely reflects some combination of measurement error and true independence in the data. Clinical trial sites perform worse on two outcome measures: they have 0.13 standard deviations higher rates of four hospital-acquired conditions included in the HSS, and 0.34 standard deviations higher rates of six complications included in the HSS patient safety indicator index. Clinical trial sites do not differ on the rate of infections during colon surgery, and they have substantially lower mortality rates when treating patients suffering from heart attack, heart failure, and pneumonia.

Finally, clinical trial sites are 13 to 17 percentage points more likely to appear in the top 50 hospitals in 12 specialties rated by the US News Hospital Quality Rankings, and they have an average of 0.74 to 1.0 additional specialties ranked. Overall, these results point to "ability bias" as a site selection mechanism in clinical trials: almost mechanically, clinical trials take place at higher-quality hospitals because technology, size, and skill are complements to clinical research.

7 Conclusion

Given the potential for site-specific treatment effect heterogeneity, replication is viewed as important because it gives a sense of the distribution of possible effects. I formalize two external unconfoundedness assumptions that underlie two different assumptions that researchers might make when extrapolating results, and I show that site selection bias is a natural result of economic processes underlying RCT research. To quantify these issues, I study a series of energy conservation RCTs involving 6.8 million households in 98 sites across the U.S. I show evidence of two positive site selection mechanisms. First, early Opower partners were in areas where consumers have more environmentalist tastes, which is associated with larger treatment effects. Second, utilities typically targeted initial interventions at the most responsive consumer subpopulations, meaning that subsequent interventions within the same utility often performed more poorly. I augment these results by showing suggestive evidence from two other domains: microfinance institutions (MFIs) that partner on academic experiments differ on observables from the global population of MFIs, and clinical trials for drugs and surgical procedures take place at hospitals that differ from the

national population of hospitals.

While the Opower context offers substantial insight into site selection processes and substantial ability to control for selection on observables, one rarely has the luxury of a 98-site program evaluation. In more typical conditions, researchers can take several steps. At the design phase, the researcher can devote additional resources to recruiting the least likely partners. For example, if Opower needed to gather more information on what nationwide effects might be, it could focus marketing efforts or price reductions toward potential partners in less environmentally-conscious regions.

At the reporting phase, analysts can do two things. First, we can clearly define the target site or population of interest. Second, just as it is common to provide evidence on internal validity by comparing observable characteristics of treatment and control groups, we can provide suggestive evidence on external validity by comparing the observable characteristics of the sample population and the target population of policy interest. Similarly, we can compare the observable characteristics of the experimental partner to the observable characteristics of other organizations that might implement a scaled program. Evidence from the Opower energy conservation programs suggests that these and any other steps to understand and address site selection bias could be important.

References

- [1] Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas Kane, and Parag Pathak (2009). "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." NBER Working Paper No. 15549 (November).
- [2] AHA (American Hospital Association) (2012). "AHA Annual Survey Database." See <http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml>
- [3] Aigner, Dennis (1984). "The Welfare Econometrics of Peak-Load Pricing for Electricity." *Journal of Econometrics*, Vol. 26, No. 1-2, pages 1-15.
- [4] Allcott, Hunt (2011). "Social Norms and Energy Conservation." *Journal of Public Economics*, Vol. 95, No. 9-10 (October), pages 1082-1095.
- [5] Allcott, Hunt and Michael Greenstone (2012). "Is There an Energy Efficiency Gap?" *Journal of Economic Perspectives*, Vol. 26, No. 1 (Winter), pages 3-28.
- [6] Allcott, Hunt, and Sendhil Mullainathan (2010). "Behavior and Energy Policy." *Science*, Vol. 327, No. 5970 (March 5th).
- [7] Angrist, Joshua (2004). "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal*, Vol. 114, No. 494 (March), pages C52-C83.
- [8] Angrist, Joshua, and Ivan Fernandez-Val (2010). "ExtrapolATE-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper No. 16566 (December).
- [9] Angrist, Joshua, Victor Lavy, and Anatalia Schlosser (2010). "Multiple Experiments for the Causal Link between the Quantity and Quality of Children." *Journal of Labor Economics*, Vol. 28 (October), pages 773-824.
- [10] Angrist, Joshua, Parag Pathak, and Christopher Walters (2011). "Explaining Charter School Effectiveness." NBER Working Paper No. 17332 (August).
- [11] Angrist, Joshua, and Jorn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives*, Vol. 24, No. 2 (Spring), pages 3-30.
- [12] Arimura, Toshi, Shanjun Li, Richard Newell, and Karen Palmer (2011). "Cost-Effectiveness of Electricity Energy Efficiency Programs." Resources for the Future Discussion Paper 09-48 (May).
- [13] Ashby, Kira, Hilary Forster, Bruce Ceniceros, Bobbi Wilhelm, Kim Friebel, Rachel Henschel, and Shahana Samiullah (2012). "Green with Envy: Neighbor Comparisons and Social Norms in Five Home Energy Report Programs." <http://www.aceee.org/files/proceedings/2012/data/papers/0193-000218.pdf>
- [14] Ayres, Ian, Sophie Raseman, and Alice Shih (2009). "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." NBER Working Paper 15386 (September).
- [15] Banerjee, Abhijit (2009). "Big Answers for Big Questions." In Cohen, Jessica, and William Easterly (Eds.), *What Works in Development? Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.
- [16] Banerjee, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden (2007). "Remedying Education: Evidence from Two Randomized Experiments in India." *Quarterly Journal of Economics*, Vol. 122, No. 3, pages 1235-1264.
- [17] Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan (2009). "The Miracle of Microfinance? Evidence from a Randomized Evaluation." Working Paper, MIT (May).
- [18] Belot, Michele, and Jonathan James (2012). "Selection into Policy Relevant Field Experiments." Working Paper, Oxford University (June).

- [19] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). "How Much Should We Trust Difference-in-Differences Estimates?" *Quarterly Journal of Economics*, Vol. 119, No. 1, pages 249-275.
- [20] Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2010). "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics*, Vol. 125, No. 1 (February), pages 263-306
- [21] Bloom, Howard, Larry Orr, George Cave, Stephen Bell, and Fred Doolittle (1993). "The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months." U.S. Department of Labor Research and Evaluation Report Series 93-C.
- [22] Bobonis, Gustavo, Edward Miguel, and Charu Puri-Sharma (2006). "Iron Deficiency Anemia and School Participation." *Journal of Human Resources*, Vol. 41, No. 4, pages 692-721.
- [23] Campbell, Donald (1957). "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin*, Vol. 54, No. 4 (July), pages 297-312.
- [24] Card, David, Jochen Kluge, and Andrea Weber (2009). "Active Labor Market Policy Evaluations: A Meta-Analysis." IZA Discussion Paper No. 4002 (February).
- [25] Cartwright, Nancy (2007), "Are RCTs the Gold Standard?" *Biosocieties*, Vol. 2, No. 2 pages 11–20.
- [26] Cartwright, Nancy (2010). "What are randomized trials good for?" *Philosophical Studies*, Vol. 147, 59–70.
- [27] Center for Climate and Energy Solutions (2011). "Energy Efficiency Standards and Targets." <http://www.c2es.org/us-states-regions/policy-maps/energy-efficiency-standards>
- [28] Chassang, Sylvain, Gerard Padro I Miquel, and Erik Snowberg (2012). "Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments." *American Economic Review*, Vol 102, No. 4 (June), pages 1279-1309.
- [29] Chattopadhyay, Raghabendra, and Esther Duflo (2004). "Women as Policy Makers: Evidence from a Randomized Policy Experiment in India." *Econometrica*, Vol. 72, No. 5, pages 1409-1443.
- [30] CMS (Center for Medicare & Medicaid Services) (2013). "Hospital Compare Data." Available from <https://data.medicare.gov/data/hospital-compare>
- [31] CTTI (Clinical Trials Transformation Initiative) (2012). "Database for Aggregate Analysis of ClinicalTrials.gov. Available from <http://www.trialstransformation.org/what-we-do/analysis-dissemination/state-clinical-trials/aact-database>
- [32] Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora (2012). "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." Working Paper, Centre de Recherche en Economie et Statistique (June).
- [33] Costa, Dora, and Matthew Kahn (2012). "Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." Working Paper, UCLA (July).
- [34] Davis, Matthew (2011). "Behavior and Energy Savings." Working Paper, Environmental Defense Fund (May). <http://blogs.edf.org/energyexchange/files/2011/05/BehaviorAndEnergySavings.pdf>
- [35] Deaton, Angus (2010a). "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, Vol. 48, No. 2 (June), pages 424–455.
- [36] Deaton, Angus (2010b). "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives*, Vol. 24, No. 3 (Summer), pages 3-16.
- [37] Dehejia, Rajeev (2003). "Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data." *Journal of Business and Economic Statistics*, Vol. 21, No. 1, pages 1–11.

- [38] Dehejia, Rajeev, and Sadek Wahba (1999). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, Vol. 94, pages 1053–1062.
- [39] Donabedian, Avedis (1988). "The Quality of Care: How Can It Be Assessed?" *Journal of the American Medical Association*, Vol. 260, No. 12, pages 1743-1748.
- [40] Duflo, Esther (2004). "Scaling Up and Evaluation." Conference Paper, Annual World Bank Conference on Development Economics.
- [41] Duflo, Esther, Rachel Glennerster, and Michael Kremer (2007). "Using Randomization in Development Economics Research: A Toolkit." Centre for Economic Policy Research Discussion Paper No. 6059 (January).
- [42] Greenberg, David, and Mark Schroder (2004). The Digest of Social Experiments; Third Edition. Washington, DC: Urban Institute Press.
- [43] Friedrich, Katherine, Maggie Eldridge, Dan York, Patti Witte, and Marty Kushler (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).
- [44] Gine, Xavier, and Dean Karlan (2010). "Group versus Individual Liability: Long Term Evidence from Philippine Microcredit Lending Groups." Working Paper, Yale University (May).
- [45] Heckman, James (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, Vol. 47, No. 1 (January), pages 153-161.
- [46] Heckman, James (1992). "Randomization and social policy evaluation". In Charles Manski and Irwin Garfinkel (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press: Cambridge, MA, pages 201-230.
- [47] Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing Selection Bias Using Experimental Data." *Econometrica*, Vol. 66, No. 5 (September), pages 1017-1098.
- [48] Heckman, James, Hidehiko Ichimura, and Petra Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, Vol. 64, No. 4, (October), pages 605-654.
- [49] Heckman, James, Robert Lalonde, and Jeffrey Smith (1999). "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card (Eds.) Handbook of Labor Economics, Chapter 31, pages 1865-2097.
- [50] Heckman, James, and Jeffrey Smith (1995). "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, Vol. 9, No. 2 (Spring), pages 85-110.
- [51] Heckman, James, and Jeffrey Smith (1997). "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study," NBER Working Paper No. 6105 (July).
- [52] Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity." *The Review of Economics and Statistics*, Vol. 88, No. 3 (August), pages 389-432.
- [53] Heckman, James, and Edward Vytlacil (2001). "Policy-Relevant Treatment Effects." *American Economic Review*, Vol. 91, No. 2 (May), pages 107–111.
- [54] Heckman, James, and Edward Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, Vol. 73, No. 3 (May), pages 669–738.
- [55] Heckman, James, and Edward Vytlacil (2007a). "'Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In James Heckman and Edward Leamer (Eds), Handbook of Econometrics, Vol. 6B. Amsterdam: Elsevier, pages 4779-4874.

- [56] Heckman, James, and Edward Vytlacil (2007b). "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 Heckman and Edward Leamer (Eds), *Handbook of Econometrics*, Vol. 6B. Amsterdam: Elsevier, pages 4875-5144.
- [57] Hotz, Joseph (1992). "Designing Experimental Evaluations of Social Programs: The Case of the U.S. National JTPA Study." University of Chicago Harris School of Public Policy Working Paper 9203 (January).
- [58] Hotz, Joseph, Guido Imbens, and Jacob Klerman (2006). "Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program." *Journal of Labor Economics*, Vol. 24, No. 3, pages 521-66.
- [59] Hotz, Joseph, Guido Imbens, and Julie Mortimer (2005). "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics*, Vol. 125, No 1-2, pages 241-270.
- [60] Imbens, Guido (2010). "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, Vol. 48, No. 2 (June), pages 399-423.
- [61] Imbens, Guido, and Jeffrey Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, Vol. 47, No. 1 (March), pages 5-86.
- [62] Integral Analytics (2012). "Sacramento Municipal Utility District Home Energy Report Program." <http://www.integralanalytics.com/ia/Portals/0/FinalSMUDHERSEval2012v4.pdf>
- [63] Karlan, Dean, and Jonathan Zinman (2009). "Observing Unobservables: Identifying Information Asymmetries With a Consumer Credit Field Experiment." *Econometrica*, Vol. 77, No. 6, pages 1993-2008 (November).
- [64] KEMA (2012). "Puget Sound Energy's Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation." Madison, Wisconsin: DNV KEMA Energy and Sustainability.
- [65] LaLonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, Vol. 76, No. 4, pages 604-620.
- [66] Lee, David, and Thomas Lemieux (2009). "Regression Discontinuity Designs in Economics." NBER Working Paper 14723 (February).
- [67] Leip, David (2013). "Dave Leip's Atlas of U.S. Presidential Elections." Available from <http://uselectionatlas.org/>
- [68] Levitt, Steven D. and John A. List (2009). "Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review*, Vol. 53, No. 1 (January), pages 1-18.
- [69] Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan (2011). "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*, Vol. 25, No. 3 (Summer), pages 17-38.
- [70] Manning, Willard, Joseph Newhouse, Naihua Duan, Emmett Keeler, Bernadette Benjamin, Arleen Leibowitz, Susan Marquis, and Jack Zwanziger (1988). "Health Insurance and the Demand for Medical Care." Santa Monica, California: The RAND Corporation.
- [71] Manski, Charles (2011). "Policy Analysis with Incredible Certitude." *The Economic Journal*, Vol. 121, No. 554 (August), pages F261-F289.
- [72] Meyer, Bruce (1995). "Lessons from U.S. Unemployment Insurance Experiments." *Journal of Economic Literature*, Vol. 33, No. 1 (March), pages 91-131.
- [73] Miguel, Edward, and Michael Kremer (2004). "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica*, Vol. 72, No. 1, pages 159-217.

- [74] Murphy, Kevin M., and Robert Topel (1985). "Estimation and Inference in Two-Step Econometric Models." *Journal of Business and Economic Statistics*, Vol. 3, No. 4 (October), pages 370-379.
- [75] NBER (National Bureau of Economic Research) (2013). "CMS Medicare Provider of Services Files." Available from <http://www.nber.org/data/provider-of-services.html>
- [76] NHGIS (National Historical Geographic Information System) (2013). "NHGIS Data Finder." Available from <https://www.nhgis.org>
- [77] Nolan, Jessica, Wesley Schultz, Robert Cialdini, Noah Goldstein, and Vldas Griskevicius (2008). "Normative Influence is Underdetected." *Personality and Social Psychology Bulletin*, Vol. 34, pages 913-923.
- [78] Opinion Dynamics (2012). "Massachusetts Three Year Cross-Cutting Behavioral Program Evaluation Integrated Report." Waltham, MA: Opinion Dynamics Corporation.
- [79] Pritchett, Lant (2002). "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." Working Paper, Kennedy School of Government (April).
- [80] Rodrik, Dani (2009). "The New Development Economics: We Shall Experiment, but How Shall We Learn?" In J. Cohen and W. Easterly, Eds., *What Works in Development? Thinking Big and Thinking Small*. Washington, DC: Brookings Institution Press.
- [81] Rosenbaum, Paul, and Donald Rubin (1983). "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika*, Vol. 70, No. 1, pages 41-55.
- [82] Rothwell, Peter (2005). "External validity of randomised controlled trials: "To whom do the results of this trial apply?" *The Lancet*, Vol. 365, pages 82-93.
- [83] Rubin, Donald (1974). "Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies." *Journal of Educational Psychology*, Vol. 66, No. 5, pages 688-701.
- [84] Sanbonmatsu, Lisa, Jens Ludwig, Lawrence Katz, Lisa Gennetian, Greg Duncan, Ronald Kessler, Emma Adam, Thomas McDade, and Stacy Tessler Lindau (2011). "Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation." Available from http://isites.harvard.edu/fs/docs/icb.topic964076.files/mto_final_exec_summary.pdf
- [85] Schultz, Wesley, Jessica Nolan, Robert Cialdini, Noah Goldstein, and Vldas Griskevicius (2007). "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science*, Vol. 18, pages 429-434.
- [86] Smith, Jeffrey, and Petra Todd (2004). "Does Matching Address LaLonde's Critique of Nonexperimental Estimators?" *Journal of Econometrics*, Vol 125 pages 305-353.
- [87] Steg, P., J. Lopez-Sendon, E. Lopez de Sa, S. Goodman, J. Gore, F. Anderson Jr, D. Himbert,
- [88] Allegrone, J. and F. Van de Werf (2007). "External validity of clinical trials in acute myocardial infarction." *Archives of Internal Medicine*, Vol. 167, No. 1, pages 68-73.
- [89] Stuart, Elizabeth, Stephen Cole, Catherine Bradshaw, and Philip Leaf (2011). "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society*, Vol. 174, Part 2, pages 369-386.
- [90] U.S. Bureau of Economic Analysis (2013). "Regional Data: GDP & Personal Income." Available from http://www.bea.gov/iTable/index_regional.cfm
- [91] U.S. Census (2010a). "Money Income of Households by State Using 2- and 3-Year-Average Medians: 2006 to 2008." http://www.census.gov/hhes/www/income/income08/statemhi3_08.xls.

- [92] U.S. Census (2010b). "American Community Survey: GCT1502. Percent of People 25 Years and Over Who Have Completed a Bachelor's Degree." http://factfinder.census.gov/servlet/GCTTable?_bm=y&-context=gct&-ds_name=ACS_2008_3YR_G00_&-mt_name=ACS_2008_3YR_G00_GCT1502_US9T&-CONTEXT=gct&-tree_id=3308&-geo_id=&-format=US-9T&-_lang=en
- [93] U.S. Census (2010c). "Table 391. Vote Cast for United States Representatives, by Major Political Party – States." <http://www.census.gov/compendia/statab/2010/tables/10s0391.xls>
- [94] U.S. Department of Energy (2011). "Renewables Portfolio Standards." Available from http://apps1.eere.energy.gov/states/maps/renewable_portfolio_states.cfm
- [95] U.S. EIA (Energy Information Administration) (2013). "Form EIA-861 data files." Available from <http://www.eia.gov/electricity/data/eia861/>
- [96] U.S. EIA (Energy Information Administration) (2011). "Table 5A. Residential Average Monthly Bill by Census Division, and State." Available from http://www.eia.gov/electricity/sales_revenue_price/html/table5_a.html.
- [97] U.S. News (2013). "Methodology: U.S. News & World Report Best Hospitals 2013-2014." Available from http://www.usnews.com/pubfiles/BH_2013_Methodology_Report_Final_28August2013.pdf
- [98] Violette, Daniel, Provencher, Bill, and Mary Klos (2009). "Impact Evaluation of Positive Energy SMUD Pilot Study." Boulder, CO: Summit Blue Consulting.
- [99] Wennberg, David, F. L. Lucas, John Birkmeyer, Carl Bredenberg, and Elliott Fisher (1998). "Variation in Carotid Endarterectomy Mortality in the Medicare Population." *Journal of the American Medical Association*, Vol. 279, No. 16, pages 1278-1281.
- [100] Worrall, John (2007). "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass*, Vol. 2, No. 6, pages 981-1022.

Tables

Table 1: Example Site Selection Mechanisms

Selection Mechanism	Sign	Example
<i>Selection on Gain:</i>		
Targeting	Positive	Implementers target most responsive populations first.
It Pays to Be Ignorant	Positive	Partners with ineffective programs do not want to evaluate.
<i>Selection on Cost:</i>		
Ability Bias	Positive	RCTs require operational ability; ability also increases efficacy.
Complementary Programs	Positive	Able and effective partners also offer complementary programs.
Substitute Programs	Negative	Able and effective partners have other substitute programs, and the marginal intervention may have lower returns.

Table 2: Utility-Level Characteristics

	All	Partners	Non-Partners	Difference
Utility Mean Usage (kWh/day)	30.37 (9.04)	26.14 (6.99)	33.13 (9.16)	-6.99 (1.74)***
Renewables Portfolio Standard	0.63 (0.48)	0.79 (0.41)	0.52 (0.50)	0.27 (0.09)***
Green Pricing Market Share	0.0058 (0.0142)	0.0072 (0.0169)	0.0049 (0.0120)	0.0023 (0.0025)
Share Urban	0.80 (0.17)	0.85 (0.11)	0.77 (0.19)	0.08 (0.03)***
Democrat Vote Share	0.51 (0.10)	0.54 (0.08)	0.49 (0.11)	0.05 (0.02)***
Green Vote Share	0.0047 (0.0028)	0.0050 (0.0028)	0.0044 (0.0029)	0.0006 (0.0005)
Income per Capita (\$000s)	39.26 (7.24)	41.00 (6.34)	38.12 (7.56)	2.88 (1.35)**
Share College Graduates	0.28 (0.06)	0.30 (0.05)	0.26 (0.07)	0.03 (0.01)***
Residential Conservation/Sales	0.0037 (0.0082)	0.0075 (0.0115)	0.0011 (0.0027)	0.0064 (0.0033)*
Conservation Cost/Total Revenues	0.0082 (0.0117)	0.0131 (0.0147)	0.0051 (0.0077)	0.0080 (0.0038)**
Municipality-Owned Utility	0.09 (0.28)	0.02 (0.15)	0.13 (0.34)	-0.11 (0.03)***
Investor-Owned Utility	0.75 (0.43)	0.95 (0.21)	0.62 (0.49)	0.34 (0.05)***
Energy Efficiency Resource Standard	0.78 (0.41)	0.94 (0.24)	0.68 (0.47)	0.26 (0.06)***
ln(Residential Customers)	13.17 (1.58)	14.00 (1.00)	12.62 (1.65)	1.38 (0.30)***
Electricity Price (cents/kWh)	11.01 (3.29)	11.78 (3.31)	10.51 (3.18)	1.27 (0.78)
N	882	63	819	
F Test p-Value				0.0000 ***

Notes: The first three columns of this table present the means of utility-level characteristics for all utilities, for Opower partners, and for Opower non-partners, respectively. Standard deviations are in parenthesis. The fourth column presents the difference in means between partners and non-partners, with robust standard errors in parenthesis. Observations are weighted by number of residential consumers. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 3: Site-Level Metadata

	Mean	Standard Deviation	Minimum	Maximum
Number of Households (000s)	69.7	61.1	4.7	435
Number of Treated Households (000s)	46.7	48.4	2.91	348
Reports/Month	0.55	0.11	0.33	0.97
Email Reports per Paper Report	0.156	0.22	0	0.77
Control Mean Usage (kWh/day)	35.6	13.8	12.5	72.9
Average Treatment Effect (kWh/day)	0.43	0.22	0.074	1.012
Standard Error (kWh/day)	0.049	0.028	0.0085	0.17
Move Rate	0.105	0.067	0.034	0.52
Opt-Out Rate	0.005	0.0036	0.00036	0.18
Number of Sites	98			
Number of Distinct Utilities	52			

Notes: This table presents descriptive statistics of the site-level Opower metadata.

Table 4: Microdata Experiment Overviews

Site Number	Region	Start Date	Households	Treated Households	Electricity Use Observations	Baseline Usage: T-C and SE
1	Midwest	July 2009	54,475	28,027	1,873,722	0.04 (0.05)
2	Midwest	January 2009	72,885	39,024	3,186,778	0.01 (0.12)
3	Mountain	October 2009	38,710	24,201	1,308,914	0.12 (0.14)
4	West	October 2009	33,506	23,906	570,582	0.09 (0.13)
5	Northeast	September 2009	49,522	24,808	1,712,713	-0.21 (0.13)
6	West	October 2008	79,017	34,893	3,121,959	0.02 (0.1)
7	West	January 2009	42,819	9,422	1,673,438	0.26 (0.27)
8	West	September 2009	39,334	19,663	672,687	0 (0.17)
9	West	March 2008	83,955	34,664	6,393,523	-0.42 (0.58)
Combined		March 2008	494,223	238,608	20,514,316	

Notes: This table presents overviews of the nine initial Opower programs. Electricity Use Observations includes all pre- and post-treatment data.

Table 5: Household-Level Characteristics

<i>Site Number</i>	Weather		Energy Use		Household		
	<i>Heating Degrees</i>	<i>Cooling Degrees</i>	<i>Baseline Usage (kWh/day)</i>	<i>First Comparison (kWh/day)</i>	<i>EE Program Participant</i>	<i>Green Pricing</i>	<i>House Value (\$000s)</i>
1	15.9 (14.8)	2.1 (3.2)	30.9 (5.7)	3.3 (14.6)	-	-	-
2	21 (18.3)	1.1 (1.5)	29.7 (16.4)	0 (18.5)	-	-	396.1 (149.8)
3	18.3 (14.7)	1.6 (2.6)	25.1 (13.2)	0.4 (11.1)	-	-	-
4	4.3 (3.7)	2.2 (2.6)	18.2 (10.7)	1.3 (9.9)	-	-	-
5	13.6 (12.6)	2.5 (3.7)	30 (14.8)	2.5 (13.3)	-	-	-
6	13.7 (9.4)	0.7 (1.4)	30.5 (13.8)	2.9 (15.4)	-	-	345.6 (171)
7	2.9 (4.4)	11.1 (10.9)	31.2 (22.5)	-1.1 (13.9)	-	-	-
8	12.9 (6.5)	0.4 (0.6)	36.4 (17.2)	3.5 (16.3)	-	-	437.4 (293.5)
9	7.8 (7.7)	3 (3.7)	30.8 (15.1)	1 (14.1)	0.06 (0.24)	0.24 (0.43)	214.5 (146.1)

Notes: This table presents the means of household characteristics for each of the nine initial Opower programs, with standard deviations in parenthesis. Heating and cooling degrees are observed at the household-by-month level; all other variables are at the household level. Dashes mean that a variable is not observed at a site.

Table 6: Heterogeneous Treatment Effects

	(1)	(2)	(3)	(4)	(5)
T x Reports/Month	0.163 (0.058)***				
T x Heating Degrees		0.002 (0.002)			
T x Cooling Degrees		0.020 (0.009)**			
T x Baseline Usage		0.009 (0.003)***	0.026 (0.003)***	0.029 (0.003)***	
T x First Comparison		0.020 (0.003)***			
T x Green Pricing			0.232 (0.097)**		
T x EE Program Participant			0.015 (0.128)		
T x ln(House Value)				-0.047 (0.040)	0.120 (0.038)***
R2	0.70	0.68	0.73	0.71	0.71
N	1,764,421	4,846,584	957,835	2,867,006	2,867,006

Notes: This table presents estimates of Equation (10) with different X characteristics. The outcome variable is average electricity use in kilowatt-hours per day. Robust standard errors, clustered by household, are in parenthesis. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 7: Opower Partner Selection

Mechanism:	Usage Targeting	Pop. Prefs.	Other Progs.	Partner Structure	Price	All
	(1)	(2)	(3)	(4)	(5)	(6)
Probit Selection Equation						
Utility Mean Usage (kWh/day)	-0.060 (0.018)***					-0.051 (0.025)**
Renewables Portfolio Standard		0.49 (0.30)				-0.01 (0.38)
Green Pricing Market Share		0.02 (6.21)				-1.83 (7.89)
Share Urban		1.80 (1.07)*				0.90 (1.29)
Democrat Vote Share		0.74 (1.47)				-1.01 (1.76)
Green Vote Share		34.1 (44.2)				5.8 (57.2)
Income per Capita (\$000s)		-0.040 (0.030)				-0.07 (0.04)*
Share College Graduates		5.34 (2.93)*				10.58 (3.96)***
Residential Conservation/Sales			72.7 (26.5)***			82.3 (31.2)***
Conservation Cost/Total Revenues			10.2 (17.1)			-6.4 (18.3)
Municipality-Owned Utility				0.17 (0.44)		-0.72 (0.53)
Investor-Owned Utility				0.83 (0.40)**		0.76 (0.38)**
Energy Efficiency Resource Standard				0.91 (0.48)*		0.62 (0.56)
ln(Residential Customers)				0.30 (0.157)*		0.25 (0.154)
Electricity Price (cents/kWh)					0.079 (0.052)	-0.102 (0.067)
Chi-Squared Test p-Value	0.0007	0.0008	0.0000	0.0001	0.1274	0.0000
Pseudo R^2	0.12	0.11	0.14	0.22	0.03	0.36
Regression of $\tilde{\tau}$ on $\widehat{Pr}(T = 1)$						
$\hat{\theta}$ Coefficient	-0.359 (0.193)*	0.424 (0.157)***	0.163 (0.110)	-0.035 (0.125)	0.231 (0.233)	0.055 (0.076)
SE (Murphy-Topel Adjustment)	(0.166)**	(0.120)***	(0.106)	(0.124)	(0.190)	(0.076)
SE (Robust, Clustered by Utility)	No	Yes	Yes	Yes	Yes	Yes

Notes: The top section of this table presents the results of estimating the Opower partner selection function from Equation (11), with observations weighted by the number of residential consumers. Robust standard errors, clustered by state, are in parenthesis. The bottom section of this table presents the estimation results from Equation (12). The "Murphy-Topel (1985) Standard Error" is robust, clustered by utility, with the Murphy-Topel (1985) adjustment for uncertainty in the first-step estimates of $\widehat{Pr}(D_r = 1|W_r)$. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 8: Efficacy Trends Within vs. Between Utility

	All Sites	All Sites	First Site	First Site	Within-Utility	Within-Utility
	(1)	(2)	(3)	(4)	(5)	(6)
Start Date (Years)	-0.046 (0.019)**	-0.070 (0.015)***	-0.023 (0.022)	-0.074 (0.018)***		
Control Mean Usage (kWh/day)		0.011 (0.001)***		0.011 (0.002)***		0.016 (0.002)***
Within-Utility Start Number					-0.041 (0.015)***	-0.010 (0.010)
R2	0.04	0.55	0.02	0.46	0.07	0.75
<i>N</i>	98	98	61	61	98	98
Utility Fixed Effects	No	No	No	No	Yes	Yes

Notes: *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 9: Early vs. Late Start Dates

Mechanism:	Usage Targeting	Pop. Prefs.	Other Progs.	Partner Structure	Price	All
	(1)	(2)	(3)	(4)	(5)	(6)
Predicted Utility Start Date						
Utility Mean Usage (kWh/day)	0.066 (0.016)***					0.048 (0.024)**
Renewables Portfolio Standard		-0.71 (0.36)*				-0.45 (0.34)
Green Pricing Market Share		-16.53 (8.53)*				-10.92 (8.07)
Share Urban		0.99 (1.41)				3.44 (1.35)**
Democrat Vote Share		1.09 (1.76)				2.47 (1.79)
Green Vote Share		39.0 (48.8)				72.7 (45.5)
Income per Capita (\$000s)		-0.011 (0.031)				-0.03 (0.03)
Share College Graduates		-4.05 (3.24)				-2.31 (3.30)
Residential Conservation/Sales			-43.6 (33.0)			-39.0 (30.9)
Conservation Cost/Total Revenues			-24.5 (18.4)			-3.6 (18.8)
Municipality-Owned Utility				-0.68 (0.53)		0.15 (0.70)
Investor-Owned Utility				1.20 (0.50)**		1.55 (0.53)***
Energy Efficiency Resource Standard				-0.97 (0.54)*		-0.37 (0.57)
ln(Residential Customers)				-0.42 (0.136)***		-0.26 (0.161)
Electricity Price (cents/kWh)					-0.038 (0.039)	-0.004 (0.056)
F Statistic	16.81	2.28	7.10	5.91	0.92	3.61
R2	0.20	0.20	0.17	0.26	0.01	0.50
Regression of $\tilde{\tau}$ on S and \hat{U}						
$\hat{\lambda}$ Coefficient	-0.054 (0.018)***	-0.046 (0.016)***	-0.065 (0.017)***	-0.058 (0.015)***	-0.072 (0.016)***	-0.048 (0.014)***
SE (Murphy-Topel Adjustment)						
$\hat{\theta}$ Coefficient	0.139 (0.054)**	-0.116 (0.033)***	-0.047 (0.027)*	-0.064 (0.037)*	-0.218 (0.138)	-0.076 (0.018)***
SE (Murphy-Topel Adjustment)						
Conditional on Control Usage	No	Yes	Yes	Yes	Yes	Yes

Notes: *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 10: MFI Site Characteristics

	All	Partners	Non-Partners	Difference
Average Loan Balance (\$000's)	0.71 (1.80)	0.19 (0.18)	0.92 (2.09)	-0.73 (0.25)***
Percent Portfolio at Risk	0.05 (0.07)	0.04 (0.04)	0.05 (0.09)	-0.01 (0.02)
Percent Women Borrowers	0.76 (0.26)	0.94 (0.12)	0.68 (0.26)	0.25 (0.05)***
MFI Age (Years)	22.85 (21.99)	21.73 (10.71)	23.29 (25.05)	-1.56 (7.24)
Non-Profit	0.44 (0.50)	0.24 (0.43)	0.52 (0.50)	-0.27 (0.20)
Number of Borrowers (10 ⁶)	2.51 (2.69)	4.72 (2.08)	1.64 (2.38)	3.09 (0.91)***
Borrowers/Staff Ratio (10 ³)	0.301 (0.351)	0.292 (0.067)	0.304 (0.413)	-0.012 (0.069)
Cost per Borrower (\$000's)	0.07 (0.10)	0.02 (0.04)	0.08 (0.11)	-0.06 (0.01)***
N	1595	34	1561	
F Test p-Value				0.000 ***

Notes: The first three columns present the mean characteristics for all MFIs, for field experiment partners, and for field experiment non-partners, respectively. Standard deviations are in parenthesis. The fourth column presents the difference in means between partners and non-partners, with robust standard errors in parenthesis. Observations are weighted by number of borrowers. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively. Currencies are in US dollars at market exchange rates. Percent of Portfolio at Risk is the percent of gross loan portfolio that is renegotiated or overdue by more than 30 days.

Table 11: Clinical Trial Site Characteristics

	Population Mean	Drug Trials Difference	Procedure Trials Difference
County Percent with College Degree	0.28 (0.10)	0.07 (0.00)***	0.07 (0.00)***
County Income per Capita	42.2 (13.2)	6.8 (0.6)***	7.0 (0.8)***
Urban	0.88 (0.33)	0.26 (0.01)***	0.24 (0.01)***
Bed Count	432 (348)	279 (21)***	297 (22)***
Annual Number of Admissions (000s)	19.81 (16.10)	13.69 (1.00)***	14.45 (1.02)***
Annual Number of Surgeries (000s)	14.21 (12.95)	9.97 (0.75)***	11.95 (1.22)***
Uses Electronic Medical Records	0.72 (0.28)	0.10 (0.01)***	0.13 (0.01)***
US News Technology Score	9.44 (5.69)	5.29 (0.25)***	5.74 (0.28)***
US News Patient Services Score	6.77 (3.38)	2.80 (0.15)***	3.02 (0.15)***
Surgical Care Process Score	0.17 (0.55)	0.19 (0.03)***	0.16 (0.02)***
Patient Communication Score	-0.33 (0.80)	0.02 (0.04)	-0.02 (0.03)
Hospital-Acquired Condition Score	0.10 (0.77)	0.13 (0.03)***	0.13 (0.04)***
Patient Safety Indicator Score	0.18 (1.16)	0.34 (0.06)***	0.34 (0.06)***
Surgical Site Infection Ratio: Colon	0.00 (0.90)	-0.03 (0.06)	0.03 (0.05)
Mortality Rate Score	-0.28 (1.05)	-0.42 (0.05)***	-0.44 (0.05)***
US News Top Hospital	0.12 (0.32)	0.13 (0.02)***	0.17 (0.02)***
Specialties in US News Top 50	0.60 (2.15)	0.74 (0.11)***	1.00 (0.15)***
N	4579		
F Test p-Value		0.0000 ***	0.0000 ***

Notes: The first column presents the mean characteristic for all US hospitals, with standard deviations in parenthesis. The second and third columns present differences in means between trial Sample sites and non-Sample sites, with robust standard errors in parenthesis. Observations are weighted by number of admissions. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

8 Figures

Figure 1: Home Energy Report, Front and Back



John Doe
1235 Main St.
Bellevue, WA 98006

Home Energy Report

Account number: 1234567890
Report period: 12/01/12-01/31/13

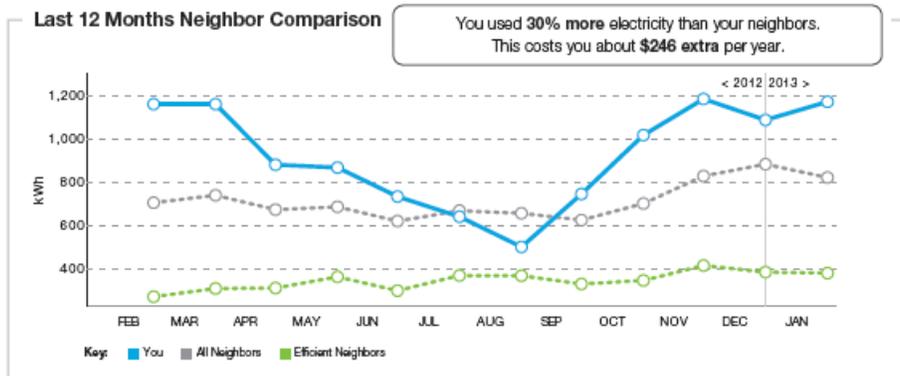
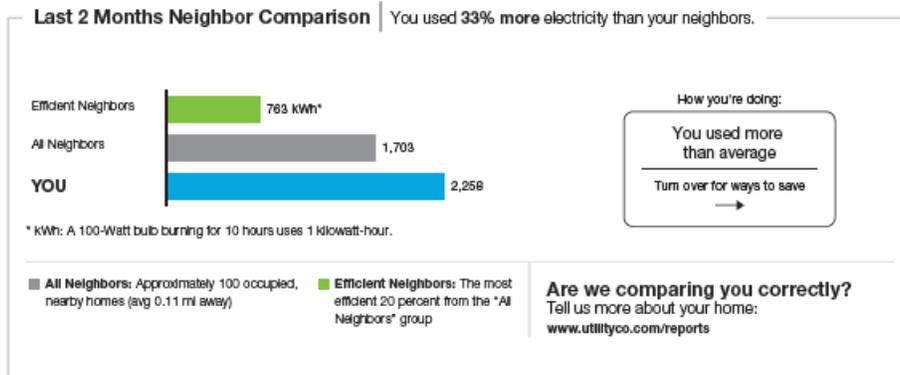
We are pleased to provide this personalized report to you as part of an energy savings program.

The purpose of this report is to:

- Provide information
- Track your progress
- Share energy efficiency tips



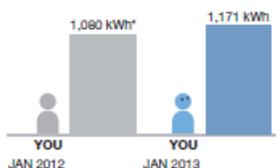
This information and more available at www.utilityco.com/reports



Turn over for savings →

Personal Comparison

How you're doing compared to last year:



So far this year, you used **8% MORE** electricity than last year.

Looking for ways to save? Visit www.utilityco.com/reports

* kWh: A 100-Watt bulb burning for 10 hours uses 1 kilowatt-hour.

Action Steps | Personalized tips chosen for your home

Smart Purchase

An affordable way to save more

Program your thermostat

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

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

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

SAVE UP TO
\$80 PER YEAR

Smart Purchase

An affordable way to save more

Check your air filters every month

You can improve the energy efficiency of your heating and cooling systems and improve your indoor air quality by checking your filters monthly.

First, remove the filter — it usually slides right out. Next, hold the filter up to a light to see if it is clogged.

You can find an inexpensive replacement for a clogged disposable filter at your local hardware store. Check your manual for cleaning instructions if you have a permanent filter.

SAVE UP TO
\$45 PER YEAR

Smart Purchase

An affordable way to save more

Seal air leaks

Gaps and cracks between the inside and outside of your home can allow heated or cooled air to escape. This forces your heating or cooling system to work harder, increases energy costs, and decreases comfort.

To find leaks, follow drafts to their source. Check where materials meet, like between the foundation and walls, the chimney and siding, and where gas and electricity lines exit your house.

Seal any small cracks you find with caulk and larger ones with polyurethane foam.

SAVE UP TO
\$215 PER YEAR

UtilityCo

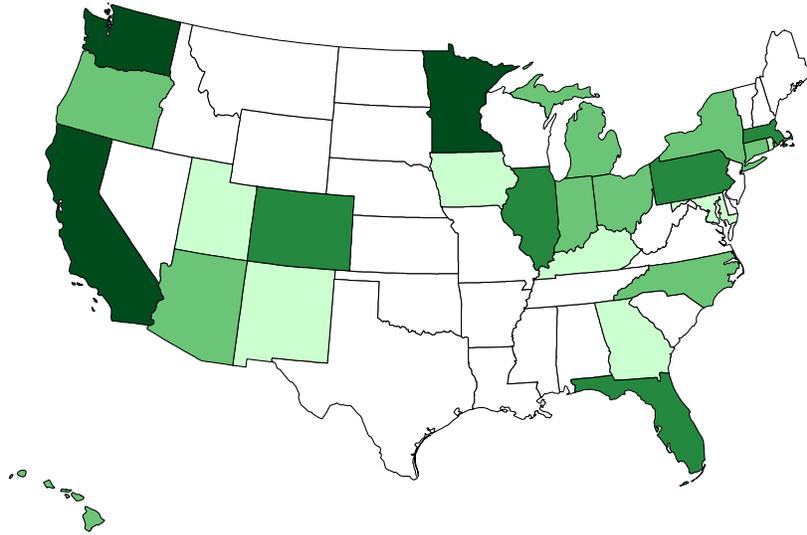
runs on OP@WER*

www.utilityco.com/reports | (555) 555-5555 | energyreports@example.com

© 2013 Opower

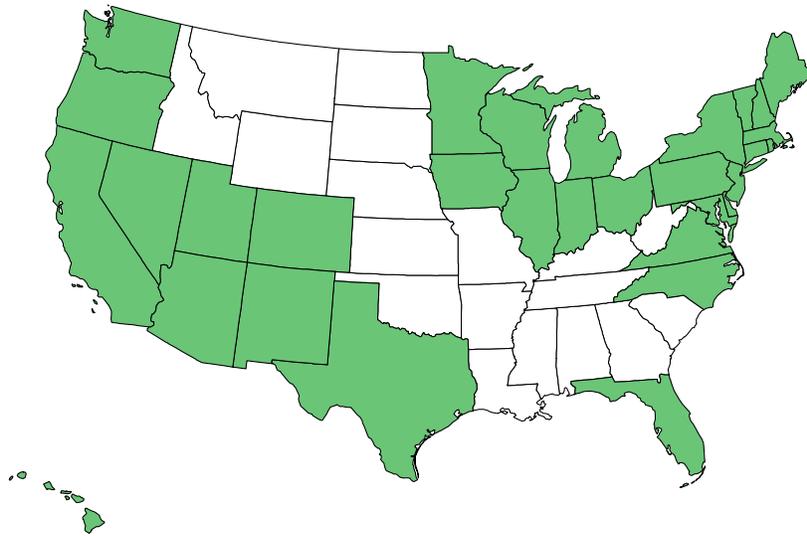
Printed on 100% post-consumer recycled paper using water-based inks.

Figure 2a: States with Opower Sites



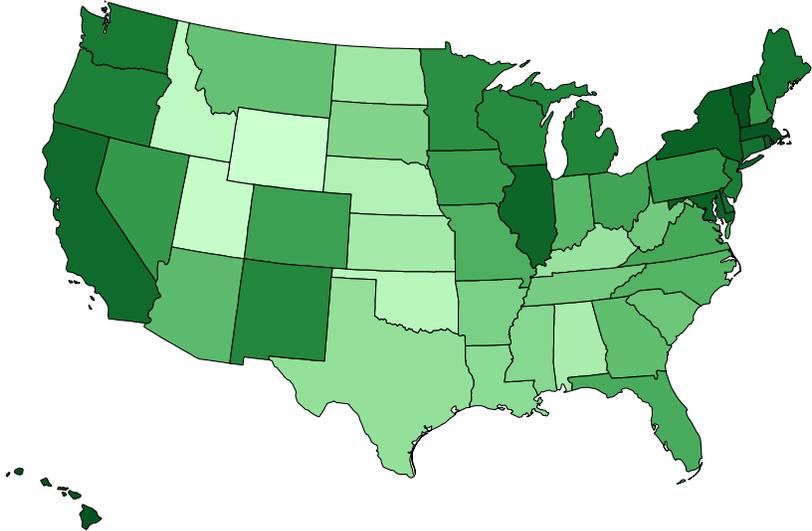
Notes: Shaded states have an Opower site. Darker colors indicate earlier program start dates.

Figure 2b: States with Energy Efficiency Resource Standards and Agencies



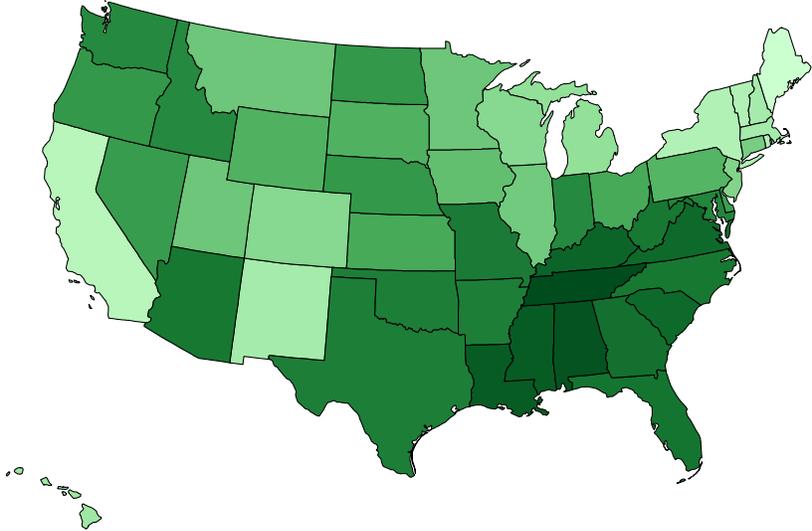
Notes: Shaded states have either a quasi-governmental energy efficiency agency (Maine, Vermont, Oregon, and Hawaii) or an Energy Efficiency Resource Standard.

Figure 2c: State Democrat Vote Shares



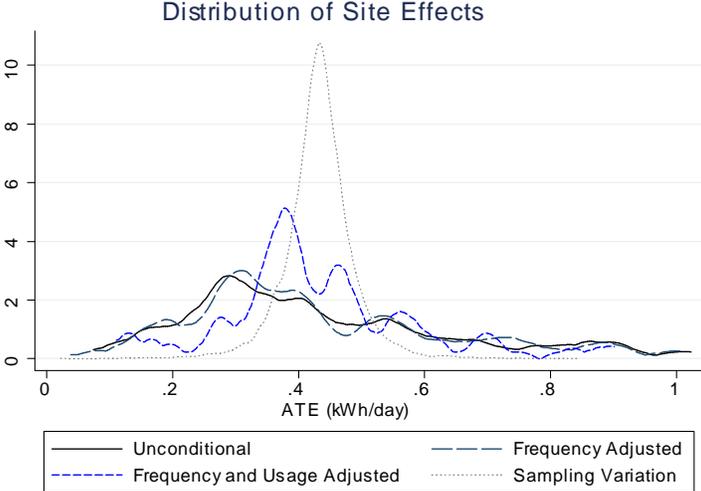
Notes: Darker shading represents a higher ratio of votes for the Democratic candidate to votes for the Democratic or Republican candidate in the 2004 and 2008 elections.

Figure 2d: State Average Residential Electricity Usage



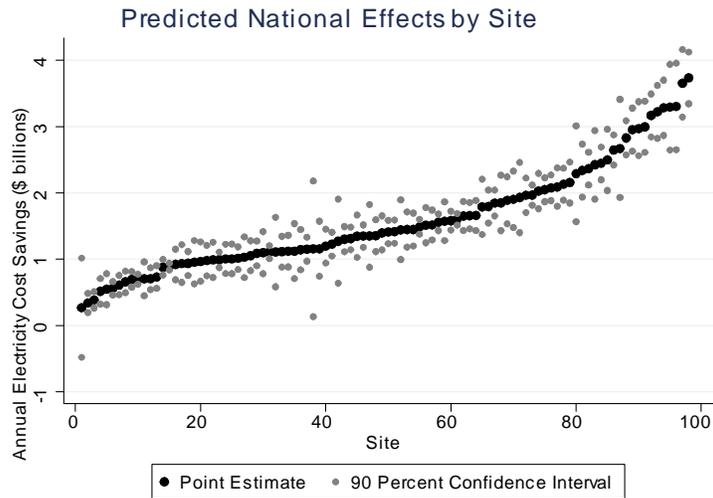
Notes: Darker shading indicates higher average residential electricity usage.

Figure 3: Distribution of Site Effects



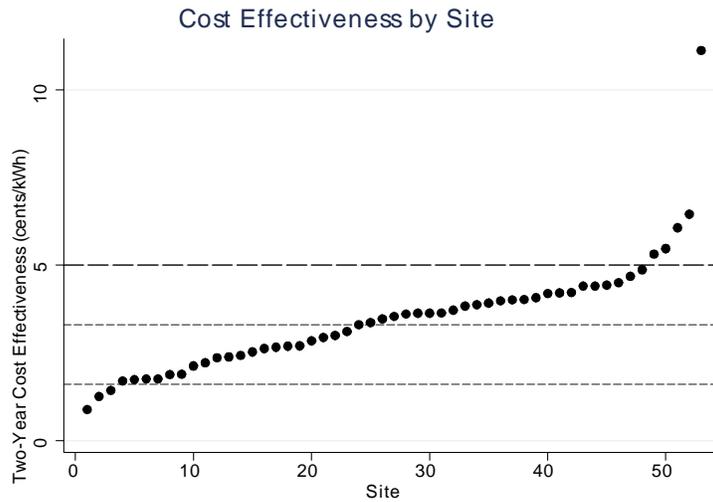
Notes: This figure presents kernel density plots of the distributions of estimated ATEs across sites. "Sampling Variation" represents the distribution of ATEs that would result if all true ATEs were the same, and the estimated ATEs differed only due to sampling variation.

Figure 4: Predicted Nationwide Effects by Site



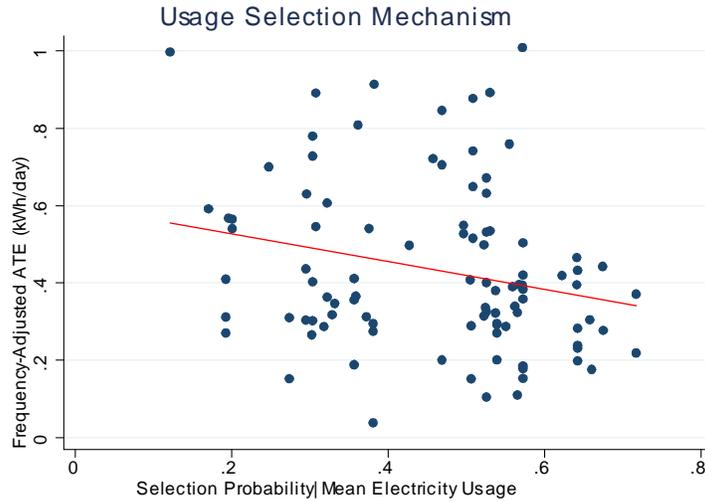
Notes: This figure presents the national annual electricity cost savings that would be predicted by extrapolating the ATE from each site to all US households. Each black dot represents a site, with 90 percent confidence intervals plotted in grey.

Figure 5: Cost Effectiveness by Site



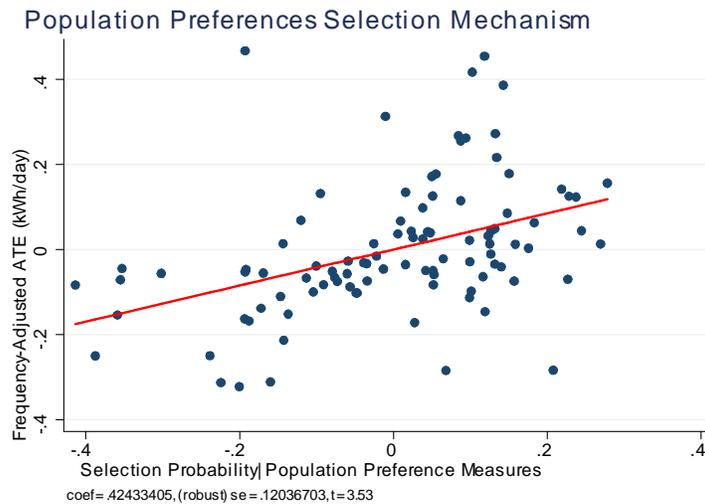
Notes: This figure presents the cost effectiveness over the first two years of each program.

Figure 6a: Site Selection on Utility Average Usage



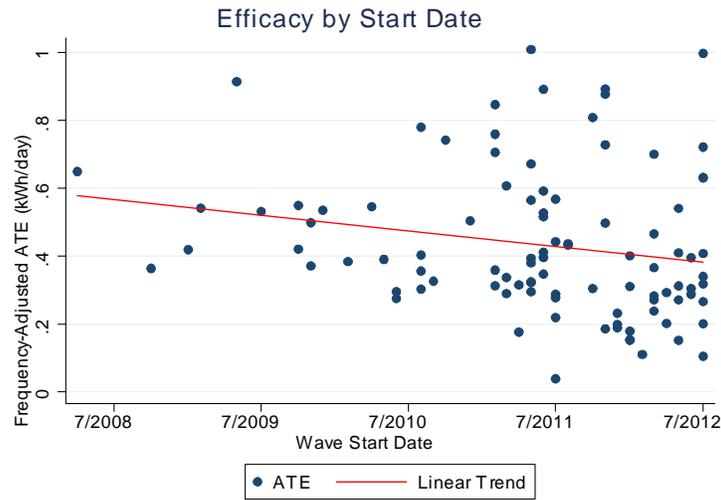
Notes: This figure plots the data and fitted regression line for Equation (12) when selection probability is estimated based on Mean Electricity Usage.

Figure 6b: Site Selection on Population Preferences



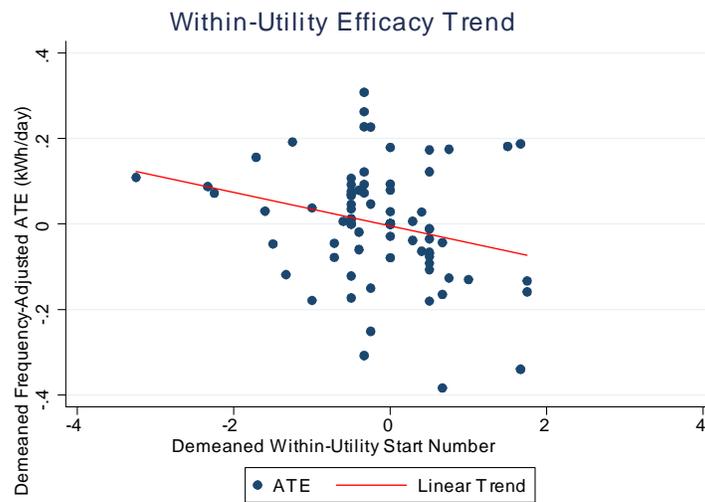
Notes: This figure plots the data and fitted regression line for Equation (12) when selection probability is estimated based on the population preferences measures: Renewables Portfolio Standard, Green Pricing Market Share, Share Urban, Democrat Vote Share, Green Vote Share, Income per Capita, and Share College Graduates. Stata's reported standard error is robust, clustered by utility.

Figure 7: Efficacy Trend



Notes: This figure plots the data and fitted regression line for Equation (13), matching column 1 of Table 8.

Figure 8: Within-Utility Efficacy Trend



Notes: This figure plots the data and fitted regression line for Equation (13), conditional on utility fixed effects. It matches column 5 of Table 8.

Appendix: For Online Publication

Is Replication Enough? Site Selection Bias in Program Evaluation

Hunt Allcott

Appendix: Preparation of Clinical Trial and Hospital Data

ClinicalTrials.gov is a registry and results database of clinical trials conducted in the United States and other countries. Although the registry does not contain all clinical studies, the number of studies registered has increased as policies and laws requiring registration have been enacted and as voluntary registration has caught on. The database for Aggregate Analysis of ClinicalTrials.gov contains records of each registered trial as of September 27, 2012 (CTTI 2012). There were 108,047 "interventional" studies (randomized control trials). Of these, 71 percent were "Drug trials," by which I mean that at least one treatment group was given a drug, biological intervention, or dietary supplement. Thirteen percent were "Procedure trials," by which I mean that at least one treatment group received a surgical or radiation procedure. Each trial takes place at one or more sites, and there are 480,000 trial-by-site observations for Drug trials and 72,000 trial-by-site observations for Procedure trials. Many trials take place at clinics, corporate research sites, or other institutions: 135,000 and 37,000 trial-by-site observations of Drug and Procedure trials, respectively, were matched to the hospital database using hospital name and zip code.

The hospital database combines three major data sources: the NBER Center for Medicare & Medicaid Services (CMS) Provider of Services (POS) files for 2011 (NBER 2013), the American Hospital Association (AHA) Annual Survey Database for 2011 (AHA 2012), and the CMS Hospital Compare database (CMS 2013). Hospitals are linked between the databases using the six-digit CMS provider identification number. From the POS files, I extract the hospital name, county, zip code, urban location indicator variable, and bed count.

From the AHA database, I extract number of admissions and number of surgical procedures, as well as information on electronic medical records and the US News Technology and Patient Services scores. The Electronic Medical Records variable takes value 1 if the hospital has fully implemented, 0.5 if partially implemented, and zero if there are no electronic medical records. In their Best Hospitals 2013-2014 rankings, U.S. News and World Report identifies 21 technologies as part of their Index of Hospital Quality (U.S. News 2013), from ablation of Barrett's esophagus to transplant services. The U.S. News Technology Score variable is simply the number of these technologies that the hospital offers on-site. U.S. News also identifies 13 patient services, from an Alzheimer's center to wound management services. Analogously, the U.S. News Patient Services Score is the number of these services that the hospital offers on-site.

The remainder of the measures are from the CMS Hospital Compare database. Each of the measures described below is normalized across hospitals to mean zero, standard deviation one. The Patient Communication Score combines four variables from the Survey of Patients' Hospital Experiences using the following formula:

$$\begin{aligned} & \text{Percent of patients who reported that their nurses "Always" communicated well} \\ & + \frac{1}{2} \cdot \text{Percent of patients who reported that their nurses "Usually" communicated well} \\ & \quad \text{Percent of patients who reported that their doctors "Always" communicated well} \\ & + \frac{1}{2} \cdot \text{Percent of patients who reported that their doctors "Usually" communicated well} \\ & \quad + \text{Percent of patients who reported that staff "Always" explained about medicines} \\ & + \frac{1}{2} \cdot \text{Percent of patients who reported that staff "Usually" explained about medicines} \\ & \quad + \text{Percent of patients who reported that YES they were given} \\ & \quad \quad \quad \text{information about what to do during recovery} \end{aligned}$$

The Mortality Rate Score variable is the sum of three components: the 30-day mortality rates from pneumonia, heart failure, and heart attack. Each component is normalized to mean zero, standard deviation one before being added together.

The next four variables from Hospital Compare were motivated directly from the Hospital Safety Score methodology, available from <http://www.hospitalsafetyscore.org>. The Surgical Care Process Score is the sum of five measures from the Surgical Care Improvement Project, which reports the percentage of times

that surgeons at the hospital followed accepted practices, from giving prophylactic antibiotic within one hour of surgical incision to giving appropriate venous thromboembolism. For each of the five specific measures, I normalized the percentages to have mean zero, standard deviation one across hospitals so as to not overweight variation coming from any one measure. I then summed the normalized measures and again normalized the sum to have mean zero, standard deviation one.

The Surgical Site Infection Ratio is the Standardized Infection Ratio for Colorectal Surgery.

The Hospital Safety Score includes the incidence rates of four Hospital Acquired Conditions: foreign object retained after surgery, air embolism, pressure ulcers, and falls and trauma. Each of these individual rates is normalized to mean zero, standard deviation one. The Hospital Acquired Condition Score is the sum of these four normalized measures.

The Hospital Safety Score incorporates six measures from the Agency for Healthcare Research and Quality Patient Safety Indicators (PSIs), which are again reported as incidence rates. These include surgical deaths, collapsed lungs, post-operative blood clots, post-operative ruptured wounds, and accidental lacerations.

Appendix Tables

Appendix Table A1: Additional Partner Selection Results

	Univariate Correlation with ATE	Univariate $\hat{\rho}$ Selection Coefficient	Univariate $\hat{\theta}$ Coefficient	$\hat{\theta}$ Coefficient when Omitted
	(1)	(2)	(3)	(4)
Probit Selection Equation				
Utility Mean Usage (kWh/day)	0.008 (0.004)**	-0.06 (0.02)***	-0.36 (0.17)**	0.059 (0.076)
Renewables Portfolio Standard	0.17 (0.04)***	0.79 (0.28)***	0.59 (0.15)***	0.056 (0.076)
Green Pricing Market Share	3.01 (0.77)***	7.66 (6.36)	1.00 (0.25)***	0.063 (0.078)
Share Urban	0.21 (0.20)	2.39 (0.86)***	0.29 (0.23)	0.057 (0.077)
Democrat Vote Share	0.47 (0.25)*	3.29 (1.48)**	0.38 (0.20)*	0.048 (0.075)
Green Vote Share	28.18 (5.59)***	47.34 (38.41)	1.54 (0.31)***	0.053 (0.076)
Income per Capita (\$000s)	0.01 (0.00)***	0.04 (0.02)**	0.42 (0.15)***	0.080 (0.079)
Share College Graduates	0.68 (0.31)**	5.85 (1.68)***	0.30 (0.14)**	0.068 (0.081)
Residential Conservation/Sales	2.42 (2.47)	85.19 (17.81)***	0.13 (0.11)	0.097 (0.086)
Conservation Cost/Total Revenues	4.75 (1.65)***	39.46 (12.68)***	0.33 (0.11)***	0.062 (0.077)
Municipality-Owned Utility	0.01 (0.05)	-1.07 (0.28)***	-0.02 (0.15)	0.045 (0.079)
Investor-Owned Utility	-0.11 (0.06)*	1.44 (0.25)***	-0.26 (0.14)*	0.085 (0.079)
Energy Efficiency Resource Standard	0.35 (0.04)***	1.20 (0.47)**	0.96 (0.10)***	0.059 (0.070)
ln(Residential Customers)	-0.001 (0.014)	0.471 (0.124)***	0.003 (0.088)	0.030 (0.083)
Electricity Price (cents/kWh)	0.007 (0.006)	0.079 (0.052)	0.231 (0.190)	0.051 (0.080)

Notes: Column 1 presents the coefficient of a regression of the frequency-adjusted ATE $\tilde{\tau}_r$ on each W_r variable, with robust standard errors clustered by utility. Column 2 presents the $\hat{\rho}$ coefficient from estimating the Opower partner selection function from Equation (11) using each W_r variable individually, with robust standard errors and observations weighted by number of residential consumers. Column 3 presents the $\hat{\theta}$ from Equation (12) using the predicted probabilities from column 2. Column 4 presents the $\hat{\theta}$ from the selection estimation from Equation (12) leaving out each W_r variable individually. Columns 3 and 4 have robust standard errors, clustered by utility, with the Murphy-Topel (1985) adjustment. Columns 1, 3, and 4 condition on Control Mean Usage for all variables other than Utility Mean Usage. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.