External Validity and Partner Selection Bias

Hunt Allcott and Sendhil Mullainathan[*]

August 26, 2011

Abstract

Program evaluation often involves generalizing internally-valid site-specific estimates to a different population or environment. While there is substantial evidence on the internal validity of non-experimental relative to experimental estimates (e.g. Lalonde 1986), there is little quantitative evidence on the external validity of site-specific estimates, because identical treatments are rarely evaluated in multiple settings. This paper examines a remarkable series of 14 energy conservation field experiments run by a company called OPOWER, involving 550,000 households in different cities across the U.S. Despite the availability of potentially-promising individual-level controls, we show that the unexplained variation in treatment effects across sites is both statistically and economically significant. Furthermore, we show that the electric utilities that partner with OPOWER differ systematically on characteristics that are correlated with the treatment effect, providing evidence of a "partner selection bias" that is analogous to biases caused by individual-level selection into treatment. We augment this result in a different context by showing that partner microfinancial institutions (MFIs) that carry out randomized experiments appear to be selected on observable characteristics from the global pool of MFIs. Finally, we propose a statistical test for parameter heterogeneity at “sub-sites” within a site that provides suggestive evidence on whether site-specific estimates can be generalized.

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

Keywords: Randomized field experiments, extrapolation, selection bias, energy conservation.

We thank Josh Angrist, Amitabh Chandra, Meredith Fowlie, Guido Imbens, Dan Levy, Jens Ludwig, Rohini Pande, Todd Rogers, and seminar participants at Harvard, MIT, and NBER for insights and helpful comments. Thanks also to Tyler Curtis, Marc Laitin, Alex Laskey, Dan Yates, and others at OPOWER for fruitful discussions. Piyush Tantia and Scott Gaul provided useful help in accessing the Microfinance Information Exchange data. We are grateful to the Sloan Foundation for financial support for this paper and our related research on the economics of energy efficiency.

[*] Allcott: New York University, NBER, and ideas42. Mullainathan: Harvard University, NBER, and ideas42.

1 Introduction

Program evaluation is a fundamental part of empirical work in economics. Evaluations are used to make a policy decision: should a program be implemented or not? In some cases, evaluations are carried out in the entire target population of policy interest, or in a randomly-selected subset thereof. In most cases, however, an evaluation is performed at some sample site, and the results are generalized to make an implementation decision in a different and often larger set of target sites. This raises the question of "external validity": how well does a parameter estimate generalize across sites?

While there is a substantial theoretical discussion of external validity[1] and the importance of the problem is broadly recognized[2], we know very little about the nature of external validity problems in practice. The reason is simple: providing evidence on the generalizability of any program's impact estimate requires a statistical comparison of results from multiple internally valid studies in multiple settings.[3] However, it is unusual for an identical treatment to be experimentally or quasi-experimentally evaluated multiple times, because randomized field experiments are costly and useful natural experiments are rare. By contrast, many papers provide evidence on selection bias and the "internal validity" of an estimator, as this requires a comparison of an internally valid estimate to the non-experimental results in only one setting.

In this paper, we empirically analyze a remarkable series of 14 randomized experiments involving more than one-half million households in different sites across the United States. The experiments are run by a company called OPOWER, which mails Home Energy Reports to residential electricity consumers that provide energy conservation tips and compare their energy use to that of their neighbors. Because these Reports are essentially the same in each site and because there is essentially no non-compliance with treatment assignment, we have the unusual opportunity to focus on one particular aspect of external validity: how well the effects of an identical treatment can be

[1] Formal theoretical analyses of external validity are included in Hotz, Imbens, and Mortimer (2005), Imbens (2009), and others. Especially important in this area is work by Heckman and coauthors, including Heckman and Vytlacil (2007a, 2007b) and a series of other papers that we discuss below.

[2] Other recent articles that contain discussions of the importance of external validity include Angrist and Pischke (2010), Banerjee (2009), Cartwright (2007a, 2010), Deaton (2010), Duflo (2004), Duflo, Glennerster, and Kremer (2007), Greenberg and Shroder (2004), Heckman and Smith (1995), Ludwig, Kling, and Mullainathan (2011), Manski and Garfinkel (1992), Manski (2011), Rodrik (2009), Rothwell (2005), Worrall (2007), and many others.

[3] There is some literature that compares impacts of programs implemented at multiple sites. In development economics, this includes Banerjee, Cole, Duflo, and Linden (2007), Chattopadhyay and Duflo (2004), and a pair of related papers by Miguel and Kremer (2004) and Bobonis, Miguel, and Sharma (2006). Quasi-experimental estimates can also be compared across locations or across groups to whom different instruments are "local," as in Angrist, Lavy, and Schlosser (2010). Of course, as one weakens the definition of what a "similar" treatment is, there are increasingly large literatures of meta-analyses that compare the effects of "similar" treatments in different settings, including Aigner (1984) on electricity pricing, Card, Kluve, and Weber (2009) on labor market policies, Holla and Kremer (2009) on education and health care in developing countries, and Meyer (1995) on unemployment experiments.

Aside from the closely-related Job Training Partnership Act program evaluations we discuss below, there are other analyses of multi-site job training programs. Hotz, Imbens, and Mortimer (2005) analyze the Work INcentive (WIN) job training program implemented at four separate locations in the 1980s, while Dehejia (2003) and Hotz, Imbens, and Klerman (2006) examine the Greater Avenues for Independence (GAIN) job training program, which was carried out in six California counties.

generalized across heterogeneous populations and economic environments.[4] The quantitative results will of course be specific to this program. However, just as Lalonde's (1986) study of selection bias in the context of one particular job training program was broadly informative about internal validity, some of the qualitative findings from this type of analysis may similarly be informative about aspects of external validity.

The generalizability of the OPOWER program's effects to potential future sites is of great interest *per se*. This is because a proliferation of new regulations mandating energy conservation, spurred partially by concern over climate change and high energy prices, will force additional utilities across the country to decide whether to adopt the program. OPOWER is also of special interest because we ourselves have extrapolated the results from one early experiment, making an implicit assumption about external validity which we can now formally test. We carried out this extrapolation in a short article in Science magazine, where we argued that the treatment effects from one OPOWER experiment in Minnesota suggested that a nationwide rollout of the program would be cost effective relative to other energy conservation programs and would generate billions of dollars in energy cost savings (Allcott and Mullainathan 2010).

In the OPOWER example, we can now show that Average Treatment Effects vary by 240 percent across the 14 existing sites, an amount which is both statistically and economically significant. In the context of the calculation in our Science magazine article, this means that depending on which experiment we had evaluated first, our estimate of total energy cost savings from a nationally-scaled program would have varied by billions of dollars. Furthermore, we show that despite having seemingly good household-level demographics, controlling for these observables does not reduce the dispersion of the experimental ATEs. The standard deviation of the unexplained site-level heterogeneity is twice the average standard error.

Of course, while the magnitude of OPOWER's site-level treatment effect heterogeneity is of specific interest, there is no sense in which the specific empirical result generalizes outside of the context of these 14 experiments. One result which we believe is qualitatively generalizable derives from another unusual feature of the OPOWER context: we observe the characteristics of the entire set of electric utilities in the United States, which forms a well-defined population of OPOWER's potential partner sites. We use these data to show that OPOWER's partners are selected on observables: partner utilities tend to have different ownership structure, are larger, and tend to be in wealthier states with stronger environmental regulation. More importantly, within the 14 experiments where results are available, there is statistical evidence of what we call *partner selection bias*: selection probabilities conditional on observables are systematically correlated with treatment

[4] It is well-understood that there are other classes of threats to a study's external validity. Randomized trials may suffer from Hawthorne effects, in which the subjects behave differently because they know they are being studied. Subjects who choose or are allowed to select into randomized trials may differ from the population of interest. Even if experimental participants are randomly selected from a population, that population may itself be different from other populations of interest. Treatment fidelity may be questionable, for example because scientific projects are "gold plated" or because programs must be adapted in order to be implemented at scale. Furthermore, when programs are scaled, there may be general equilibrium effects.

effects. This suggests that the average ATE for sites that have adopted the program is likely not an unbiased measure of the average ATE for sites that have not.

Partner selection bias is likely to be relevant to randomized controlled trials (RCTs) in other settings. From any population of potential partners, only some end up running an RCT. Because field experiments require managerial ability and operational efficacy, the set of actual partners may be running more effective programs than the broader population of potential partners. This would mean that program impact estimates from RCTs would be larger than the average impact in the full population of potential partner sites. As another example, partners that are already running programs that they know are effective are likely to be more open to independent impact estimates, generating a positive selection effect.

Alternatively, partners that are particularly innovative and willing to test new programs may also be running many other effective programs in the same population. If there are diminishing returns, the additional program with an actual partner might have lower impact than at the average potential partner site. This partner selection process is conceptually analogous to the decisions that individuals make when they decide whether or not to select into treatment, which generates the individual-level selection bias that motivates the use of RCTs. In Section 2, we use a simple model to show formally how the partner-level selection process generates a comparable form of positive or negative selection bias.

As with Lalonde's (1986) analysis of experimental vs. non-experimental estimators, the OPOWER field experiments are only one example in one setting. To provide one additional data point on the conceptual issue of partner selection bias, we turn to microfinance. We examine the characteristics of microfinancial institutions (MFIs) that have partnered to carry out randomized trials with three large academic initiatives: the Jameel Poverty Action Lab, Innovations for Poverty Action, and the Financial Access Initiative. We show that partner MFIs differ from the average MFI on characteristics that might be associated with effects of various treatments, including for-profit status, size, experience, share of borrowers that are female, and average loan size. Because microfinance field experiments study a variety of different "treatments," we do not correlate selection probabilities with treatment effects as we do for the OPOWER experiments. The microfinance example, however, provides additional suggestive evidence that partner selection bias is not unique to the OPOWER energy conservation programs.

Indeed, analyses of the Job Training Partnership Act (JTPA) of 1982 also provide closely-related existing evidence. The JTPA initiated job training programs at 600 sites, of which 16 were evaluated with randomized trials. These 16 experimental sites were those that agreed to participate out of more than 200 that were originally approached (Hotz 1992). Heckman (1992) discusses how "randomization bias" may have affected the selection of experimental sites in these evaluations. Even if the 16 sites were representative of the broader population of sites, Heckman and Smith (1997) simulate that because of the substantial variability in effects across sites, the aggregate experimental impact estimates would have differed substantially depending on which set of sites were evaluated. Our paper complements this work by providing clear empirical evidence of partner selection bias in a different context and by adding a simple theoretical model of the experimental partner selection process.

Practically, what more can be done to address concerns about external validity and partner selection bias? We make two proposals. First, 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 experimental population and the target population of policy interest. Similarly, we can compare the observable characteristics of the experimental partner to the distribution of observable characteristics of organizations that might implement a scaled program. These data can be combined with informal discussions of the partner selection process and how the experimental population and partner might differ on other characteristics that correlate with the treatment effect.

Our second proposal is an empirical test that can provide suggestive evidence on on whether an empirical result will be externally valid, which we call an "F-test of sub-site heterogeneity." The idea is very simple: the unconfounded location assumption only holds when observable characteristics can be used to adjust for all of the parameter heterogeneity across sites. But the definition of a "site" is arbitrary: for example, if an RCT is implemented within many schools in a particular city, is the entire city a "site," or is each school a "site"? The intuition for our test is that if there is unexplained heterogeneity across sub-sites within a Sample, there is likely to be unexplained heterogeneity between Sample and Target unless the distribution of sub-site heterogeneity happens to be identical across sites.

Our analysis cannot be used to argue that randomized field experiments 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. In fact, 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. Furthermore, while partner selection bias is a problem specific to RCTs, the rest of our discussion of the generalizability of site-specific parameter estimates is relevant to both structural and reduced form parameters estimated using either randomized experiments or natural experiments.

The paper proceeds as follows. Section 2 presents our formal model of treatment effects, partner selection, and two technical assumptions for external validity, "unconfounded location" and "expected unconfounded location." Section 3 introduces the OPOWER data, and Section 4 estimates the magnitude of site-specific heterogeneity. Section 5 presents empirical evidence of partner selection bias in the OPOWER context, and Section 6 analyzes similar evidence for field experiments with microfinancial institutions. Section 7 presents the F-test of sub-site heterogeneity, and Section 8 concludes.

2 Model

In this section, we write a simple model of potential outcomes, selection into treatment, and average treatment effects. We start by reviewing a simple individual-level model that builds on the Rubin (1974) Causal Model and incorporates generalized Roy-style selection into treatment analogous to the Marginal Treatment Effects framework (e.g. Heckman and Vytlacil 2007b). The review is useful for comparison, as we then build a population-level model of average treatment effects and selection into programs and evaluations.

2.1 Individual-Level Model

There is a population of individual units indexed by i. Of interest is a binary treatment that affects observed outcome Y_i. Each individual unit has two potential outcomes, Y_{1i} if exposed to treatment and Y_{0i} if not. For simplicity, we assume that Y_i is a linear and additively-separable function of observed and unobserved characteristics X_i and Z_i:

$$\begin{aligned} Y_{0i} &= \beta X_i + \zeta Z_i && \text{(1a)} \\ Y_{1i} &= (\alpha + \beta) X_i + (\gamma + \zeta) Z_i && \text{(1b)} \end{aligned}$$

The linearity assumption is not central to our argument; what is central is that there are unobservables Z that influence the treatment effect. Individual i's treatment effect is the difference in Y_i between the treated and untreated states:

$$\tau_i = Y_{1i} - Y_{0i} = \alpha X_i + \gamma Z_i \tag{2}$$

Individuals incur some positive or negative net cost C_i of treatment and weight outcome Y in their objective functions by weight ω. Denoting $T_i \in \{1, 0\}$ as the treatment indicator variable, the individual selects into treatment if the net benefits are positive:

$$\begin{aligned} T_i &= 1\left(\omega \tau_i - C_i > 0\right) && \text{(3a)} \\ &= 1\left\{\omega(\alpha X_i + \gamma Z_i) - C_r > 0\right\} && \text{(3b)} \end{aligned}$$

Comparing the mean outcomes of treated vs. untreated units gives:

$$E[Y_{1i}|T_i=1]-E[Y_{0i}|T_i=0]=E[\tau_i|T_i=1] \\ +\beta(E[X_i|T_i=1]-E[X_i|T_i=0])+\zeta(E[Z_i|T_i=1]-E[Z_i|T_i=0]) \quad (4)$$

The right hand side of the first line is the Average Treatment Effect on the Treated. The second line is selection bias. The first term in the second line is a function of observables X, which can be controlled for. The second term is a function of unobservables Z, and it equals zero only under the assumption of *unconfoundedness* (Rubin 1990):

$$T_i \perp (Y_{1i}, Y_{0i})\,|X_i \quad (5)$$

Of course, if assignment to treatment is governed by equation (3a) with non-zero ω and γ, unconfoundedness does not hold, and an internally-valid estimator of the ATT is not available.

2.2 Population-Level Model

There is an analogous selection process at the partner level. This process could take two forms. First, there could be a population of potential partners that would adopt a *new* program and evaluate it using a randomized trial. For example, this is the case with OPOWER, as they approach additional utilities about adopting their Home Energy Report program. Second, there could be a population of potential partners that are already running an *existing* program and must decide whether to run a randomized trial to evaluate it. For example, this was the case with the JTPA evaluations: the researchers approached job training centers that were already running the program and attempted to convince them to allow randomized evaluations. In either case, we consider a simple selection model where a decisionmaker at each potential partner decides whether to adopt or evaluate a program based on the expected site average treatment effect and some additional net cost.

Assuming for simplicity that the population is treated with equal probability, the treatment effect at site r depends on the observable and unobservable characteristics X_{ir} and Z_{ir} of the individuals within the site's population:

$$\tau_r = \alpha E[X_{ir}|r] + \gamma E[Z_{ir}|r]$$

Potential partners incur some positive or negative net cost C_r of adopting or evaluting the treatment and weight outcomes Y_{ir} in their objective functions by weight ω. Denoting $T_r \in \{1, 0\}$

as the site-level indicator variable, the decisionmaker for a potential partner adopts or evaluates the program if net benefits are positive:

$$\begin{aligned} T_r &= 1\{\omega\tau_r - C_r > 0\} && \text{(6a)} \\ &= 1\{\omega(\alpha E[X_{ir}|r] + \gamma E[Z_{ir}|r]) - C_r > 0\} && \text{(6b)} \end{aligned}$$

Suppose there is a Sample population $r = s$ where the treatment has been implemented using a randomized trial, and we wish to generalize treatment effects to a Target site $r = g$. The Target treatment effect is:

$$\begin{aligned} \tau_g = \alpha E[X_{ir}|r=g] + \gamma E[Z_{ir}|r=g] = \tau_s \\ + \alpha(E[X_{ir}|r=g] - E[X_{ir}|r=s]) + \gamma(E[Z_{ir}|r=g] - E[Z_{ir}|r=s]) \quad (7) \end{aligned}$$

The far right hand side of the first line is the Average Treatment Effect in the Sample. The second line reflects the difference between Target and Sample ATEs. The first term is a function of observables X, which can be controlled for. The second term is a function of unobservables Z.

2.3 Assumptions for External Validity

When extrapolating from one Sample to one Target population, an unbiased estimator can be constructed under the assumption of *unconfounded location* (Hotz, Imbens, and Mortimer 2005). Using $f_r(Z)$ to denote the marginal distribution of Z in site r, this is:

$$f_s(Z) = f_g(Z) \quad (8)$$

This assumption is simply that unobservables are balanced between Sample and Target. Some analyses explicitly or implicitly assume location unconfoundedness, including the 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. We test for unconfounded location in the OPOWER context in Section 4.

In many contexts, one expects that unobservables vary across sites, and location unconfoundedness is unrealistically restrictive. Imagine, however, that one could draw a random sample of many sites and estimate the treatment effect in each. The distribution of treatment effects in the set of experimental Sample sites would equal the distribution of treatment effects in the Target sites where no experiment had been run. This motivates the assumption of *expected unconfounded location*:

$$T_r \perp f_r(Z) \tag{9}$$

This assumption is simply that the distributions of unobservables are balanced between the sets of treated and untreated sites. Under this assumption, the treatment effect in any one Sample site is an unbiased estimator of the treatment effect in any other Target site, after controlling for differences in the distribution of observables across sites. The treatment effect in any one Target need not be exactly as predicted by data from any one Sample, but across many extrapolations from many different Samples to many different Targets, the mean predicted ATE would equal the true mean ATE.

If equation (6a) determines selection of sites into treatment, expected location unconfoundedness holds under one of two conditions. First, it holds if τ is homogeneous across sites after controlling for differences in the distribution of individual-level observables. In other words, location unconfoundedness is a sufficient condition for expected location unconfoundedness. Second, it holds if selection into partnership is driven only by C and C is independent of unobservables that influence the treatment effect: $\omega = 0$ and $f_r(Z) \perp C$. When the assumption is violated, we call this "partner selection bias." This effect is closely related to the discussion of "randomization bias" that originates in Heckman (1992) and continues in later work (e.g. Heckman and Smith 1995, Heckman and Vytlacil 2007b), although that discussion is also concerned with how randomized experiments affect the selection of individuals into programs at the partner sites.

The covariance between C and τ determines whether partner selection is positive or negative. Because implementing randomized trials requires managerial ability and operational efficacy, the potential partners that are best equipped to evaluate programs may also run the most effective programs. This form of positive partner selection bias has been called "gold plating" (Duflo, Glennerster, and Kremer 2008). Another form of positive partner selection bias results from the fact the "It Pays to Be Ignorant" (Pritchett 2002): because rigorous evaluations are publicized and affect funding from foundations, governments, and other sources, potential partners that believe they are running effective programs are willing to have them evaluated, while those that fear they are running ineffective programs strategically choose to remain ignorant by avoiding randomized evaluations.

Negative partner selection bias is also quite possible. For example, potential partners that might be most capable and interested in experimenting with new programs could also be running many other effective programs or could have already treated the parts of their population that have the largest treatment effects. This form of "diminishing returns bias" would generate negative selection.

Of course, partner selection bias does not mean that the estimated Sample ATEs are biased away from the true Sample ATEs. A symantically different but mathematically equivalent way of characterizing this issue would be to discuss the "the local nature of site-specific estimates." The reason why we use the phrase "partner selection bias" is to emphasize that these local estimates in the set of partner sites may be *systematically* different from the impact estimates in the set of non-partner sites. Furthermore, these systematic differences arise from a selection process that can be theoretically understood and observed in practice.

In the subsequent sections, we analyze the external validity of the OPOWER experimental results. After giving an overview of the experiments and data in Section 3, we test the assumption of unconfounded location in Section 4. After finding that this assumption does not hold, in Section 5 we test the weaker assumption of expected unconfounded location.

3 OPOWER Experiment Overview

The empirical focus of this paper is on a series of randomized field experiments run by a company called OPOWER. The "treatment" in these experiments is to mail Home Energy Reports (HERs) to residential electricity consumers, with the goal of causing them to use less energy. These experiments have been extensively studied, including by Allcott and Mullainathan (2010), Ayres, Raseman, and Shih (2009), Costa and Kahn (2010), Davis (2011), Nolan *et al.* (2008), Schultz *et al.* (2007), and Violette, Provencher, and Klos (2009). The programs been covered extensively in the popular press and are at the center of the energy industry's growing interest in "behavior-based" (as opposed to "technology-based") energy conservation programs that are evaluated using randomized controlled trials. See Allcott (2011) for a comprehensive program evaluation, including additional background, interpretation, and discussion of economic and policy issues.

The Reports have two key components. The Social Comparison Module, which is illustrated in Figure 1, compares the household's energy use to its 100 geographically-nearest neighbors that have similar house sizes and heating types. The Action Steps Module, illustrated in Figure 2, includes energy conservation tips targeted to the household based on its historical energy use patterns and observed characteristics. OPOWER takes a population of utility customers, randomizes them into Treatment and Control, and sends Reports to the Treatment group on a monthly, bimonthly, or quarterly basis.

Aside from the frequency with which the Reports are mailed, the treatment is almost identical across the sites we study. The envelope and the Home Energy Report it contains are branded with each local utility's name, and there are minor differences in graphics and presentation over time within an experiment and across experiments. Because these differences are so small, it is likely that the bulk of the treatment effect heterogeneity results from differences in the population and from differences in the economic environment such as weather-driven variability in energy use patterns, not by differences in the Reports. In any event, there is a remarkably high degree of treatment fidelity compared to other treatments of interest in economics. For example, "job training" takes different forms at different sites (Dehejia 2003, Hotz, Imbens, and Klerman 2006), and the quality of "remedial education" should depend on the teacher's ability. The degree of treatment fidelity across OPOWER's sites makes it more likely that the treatment effects will generalize.

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 effectively no 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 Vytlacil 2006), and the treatment effect is a Policy-Relevant Treatment Effect in the sense of Heckman and Vytlacil (2001).[5]

As of the end of 2010, OPOWER had contracts to work with 45 utilities in 21 states, as mapped in Figure 3. While the partners are spread throughout the country, they tend to be concentrated along the West Coast, the upper Midwest, and the Northeast - areas of the U.S. that are wealthier, better educated, often vote democratic, and have stronger environmental regulation. Among OPOWER's partners are 30 regulated Investor-Owned Utilities (IOUs), nearly all of which are subject to mandatory energy conservation targets called Energy Efficiency Resource Standards (EERS). OPOWER also has contracts with 13 municipal utilities and three local electricity "cooperatives." These 16 utilities are non-profits whose goals may include saving money for consumers or environmental conservation. In Section 5, we quantitatively analyze the characteristics of OPOWER's partners.

As of October 2009, experiments had begun at 10 of these utilities, giving at least one year of post-treatment data. Three more locations had begun pilots but were deemed too small to include randomized control groups, so they are excluded from the present analysis. At four of the ten utilities, the populations were divided into sub-populations with higher and lower baseline usage, and the Treatment groups in the high-usage subpopulation were sent HERs with higher frequency. As a result, our analysis considers 14 "experiments" at 14 "sites." Our qualitative results are similar if we define a "site" as a utility, and consider 10 separate sites.

3.1 Data

Table 1 provides an overview of the start date and size for each experiment. In total, we observe nearly 20 million monthly electricity bills from 550 thousand households. OPOWER has contractual obligations to keep some of its partners' identities confidential, so we mask utility names and

[5] In fact, following Allcott (2011), we actually define the "treatment" as "being mailed a letter or actively opting out," so there is precisely zero non-compliance. This definition of "treatment" does in this case produce a treatment effect of policy interest: the effect of attempting to mail Home Energy Reports to an entire population. In practice, because opt-out rates are on the order of one percent per year, the ATE is the same when the "treatment" is defined as "being mailed a letter" (Allcott 2011).

locations and number the experiments from 1 to 14. Experiment pairs 1 and 2, 4 and 5, 10 and 11, and 13 and 14 are the four involving different customer subpopulations at the same utility.

This study benefits from exceptionally good household-level data, which improves the likelihood that we might be able to use these data to explain differences in treatment effects across locations. OPOWER, and the utilities they work with, gather demographic data for each customer from surveys, public records, and private-sector marketing data providers. In addition, we have augmented the household-level data with Census Tract-level information from the 2000 U.S. Census. Although we observe a larger universe of covariates, we focus on a smaller set of covariates that theory predicts might be more likely to be associated with the treatment effect.

Table 2 details the means and standard deviations of the house occupant characteristics that we consider. "First Comparison" is a normalized measure of the household's baseline energy usage compared to its neighbors, as presented to them on the first Home Energy Report they receive. Zero corresponds to the mean of the neighbor distribution, and households with lower values used relatively more energy. As detailed in Allcott (2011), theory predicts that responses to these social comparisons depend on how individuals compare to their neighbors, and the treatment effects vary substantially with baseline energy usage. As documented in Costa and Kahn (2010), households that vote Democratic, donate to environmental groups, or voluntarily purchase renewable energy have different treatment effects. The next three variables in Table 2 are tract-level average characteristics which we hypothesized could be associated with these sorts of "cultural" differences that moderate the treatment effect.

The final two columns of Table 2 present the average heating and cooling degrees for the post-treatment observations. These weather variables are associated with electricity demand and therefore may be associated with the treatment effect.[6] Experiments 10 and 11 are in an especially warm climate, with low average heating degrees and high cooling degrees, while experiments 13 and 14 are in a moderate climate, and many other sites are relatively cold.

Table 3 details the house characteristics we use. These include variables known to be associated with energy use, and thus perhaps the marginal cost of energy conservation, including whether the household has electric heat, whether the house has a pool, type of dwelling (single family or multi-family), and the size, in thousands of square feet. Because older houses have less insulation and are more "drafty," they take more energy to heat and cool, and additional motivation for or information about energy conservation could have differential effects by house age. Finally, renters have less incentive to invest in the house's energy efficiency, so we consider whether the house is rented or owner-occupied. Some characteristics are not observed at all sites; for example, we observe House Value only in experiments 3, 9, 12, 13, and 14.

[6] More 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 OPOWER Site Effects

In this section, we examine the heterogeneity in treatment effects across OPOWER's experiments. We first show that there is economically significant variation in the ATEs across sites, without conditioning on the different distributions of individual-level observables X. We then test the extent to which controlling for observables increases or decreases the conditional variation in site-specific effects.

4.1 Unconditional Variation in ATEs Across Sites

The Average Treatment Effects (ATEs) for each experiment are calculated using a difference-in-differences estimator with household fixed effects υ_i and month-by-year dummies μ_{my}, with robust standard errors clustered by household:

$$Y_{it} = \tau T_i P_{it} + \pi P_{it} + \mu_{my} + \upsilon_i + \varepsilon_{it} \tag{10}$$

In this equation, P_{it} is a post-treatment indicator variable and Y_{it} is household i's average daily electricity consumption for period t, normalized by the control group average post-treatment consumption. Note that this normalization is different for each site, so reducing energy use by two percent in a site with high consumption entails a larger level of kilowatt-hour reduction than a reduction of two percent in a site with low consumption.

The estimated ATEs are shown in Table 4. As documented by Allcott (2011), the estimated ATEs are not very sensitive to different specifications of control variables and fixed effects. Notice that in experiments 2, 3, and 9, the site population was randomly assigned between monthly, bimonthly, and/or quarterly frequencies, while all other experiments involved only one frequency.[7] The unweighted mean ATE across experiments and frequencies is a 2.03 percent reduction in electricity use. The ATEs vary by a factor of 2.4, from 1.37 percent to 3.32 percent. While some variation in treatment effects is associated with the frequency of receiving Reports, Table 4 shows that there is still substantial variation within frequency across experiments.

Utilities typically compare energy conservation programs based on a particular measure of cost effectiveness: cents of program cost to the utility per kilowatt-hours of electricity conserved. OPOWER has provided confidential pricing data, which varies by the frequency of treatment. These cost data can be combined with total energy savings, which is the site's percent average treatment effect multiplied by average electricity consumption per year, to calculate cost effectiveness. The unweighted mean cost effectiveness across sites is 3.31 cents per kilowatt-hour. As shown in the

[7] In some of the more recent experiments, letters are sent each month for the first several months of the program and bimonthly or quarterly after that.

right panels of Table 4, cost effectiveness varies across experiments by a factor of 4.2, from 1.28 to 5.36 cents per kilowatt-hour.

4.1.1 "Economic Significance"

Average Treatment Effects vary by a factor of 2.4, and cost effectiveness by a factor of 4.2. Is this variation "economically significant"? We consider two measures of economic significance. The first measure is whether site-level heterogeneity causes *program adoption errors*: a decision to adopt the program in a new location where it is in fact not cost-effective, or to not adopt the program in a new location where it is in fact cost-effective. This measure is particularly relevant in the case of OPOWER because additional utilities are considering adopting the program, and their initial decisions are based substantially on the track record of ATEs from existing sites. Heckman and Smith's (1997) analysis of JTPA also aligns with this definition: the program efficacy estimated from the 16 experimental sites was an important criterion in determining whether the program would be extended, and Heckman and Smith (1997) showed that this estimated efficacy depended markedly on which sites were used for the evaluation.

Consider the typical case of a profit-maximizing utility with an energy conservation target and a portfolio of different energy conservation programs that it can fund. Such a utility should adopt the OPOWER program if it is more cost effective than the marginal energy conservation program currently in use. While this presumably varies across utilities, there are some benchmarks. 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).[8] Whether an OPOWER program at a new site has cost effectiveness at the lower end (1.28 cents per kilowatt-hour) or upper end (5.36 cents per kilowatt-hour) of the range for existing experiments detailed in Table 2 therefore does appear to change whether a utility would or would not want to partner with OPOWER. As a concrete example, note that Experiments 10 and 11, which have two of the smallest ATEs and the worst cost effectiveness, have been cancelled by the partner utility. In this sense, the variability in site effects is therefore economically significant.

A second measure of whether site-level heterogeneity is economically significant is *variation in predicted effects at scale*: how much do the total predicted impacts of a scaled program differ depending on which site is used for the prediction? This is relevant in the case of OPOWER in the sense that some policy analyses have considered the potential impacts of scaling up the program to utility customers nationwide (Allcott and Mullainathan (2010), Davis and Wagner (2011)). If this prediction were done with results from one initial site, how much would the results vary depending on which experiment had been implemented first?

[8] The Friedrich *et al.* (2009) analysis is published by energy efficiency research and advocacy organization called the American Council for an Energy Efficient Economy. It relies on electric utilities' estimates of cost effectiveness, which often use a non-experimental, accounting-based method of program evaluation called the "deemed savings approach." Some analysts believe that these estimates could be biased toward zero.

One key statistic of interest in these analyses is the total energy cost savings that would result from such a nationwide expansion. This is calculated by multiplying nationwide total residential electricity expenditures by the average treatment effect. As Figure 4 illustrates, the predicted savings would differ by several billion dollars per year depending on which site's ATE is used for the prediction. If Site 12 or 13 were used, with their relatively small ATE, the predicted energy cost savings would be just over \$2 billion per year. If Site 8 were used, with its relatively imprecisely-estimated and large ATE, the predicted national savings would be approximately \$5 billion per year. These values are mechanically connected to the dispersion in the ATEs, so the range from smallest to largest is again around 240 percent.

4.2 Site Effects Conditional on Observables

Without controlling for observables X, we have shown that there is economically significant variation in site effects. Controlling for X could either increase or decrease the remaining variation in site effects. Mathematically, if the site-level means of αX and γZ are negatively (positively) correlated, then controlling for αX increases (decreases) the dispersion of the unexplained site effects. What happens in the OPOWER context?

One way to answer this question would be to extrapolate from each of the 14 sites to each of the other 14 sites, controlling for differences in characteristics observed in both sites, and test how well ATEs predicted from Sample data match true ATEs estimated in a Target. However, in many sites the α parameters are imprecisely estimated, making it difficult to control for observable differences across sites. The most precise way to estimate these parameters is to pool data from all 14 sites. Given our data, this pooled approach is a "best-case scenario" in terms of precisely estimating the α's, and it is clearly better than what would be available in the typical scenario with a given number of observations from only one site.

Table 5 presents a series of regressions that pool data across all sites, control for different configurations of observables, and estimate the residual variation in site effects. Indexing sites by r, the estimating equations are:

$$Y_{irt} = \sum_{r=1}^{R} \mu_r T_{ir} P_{irt} + (\alpha X_{irt}) T_{ir} P_{irt} + \sum_{r=1}^{R} (\beta_r X_{irt}) P_{irt} + \sum_{r=1}^{R} \pi_r P_{irt} + \upsilon_{ir} + \varepsilon_{irt} \tag{11}$$

The parameters of interest are the unexplained site effects μ_r, which are equal to γZ in the notation from our model in Section 2. The α parameters capture how observables X moderate the treatment effect; these are assumed to be constant across sites. The regression controls appropriately for lower-order interactions, capturing site-specific post-treatment differences through π_r and interactions of X variables with the post-treatment dummies through the vectors β_r.[9] This

[9] Missing X variables are generated using conditional mean imputation. When a variable is observed at other

equation is comparable to Equation (**??**), except that it allows for heterogeneity on observables and omits month dummies for computational reasons; these dummies have essentially zero impact on the point estimates or standard errors in the site-level regressions.

Column I of Table 5 includes only the site dummies and site-specific post-treatment dummies as right-hand-side variables. As shown at the bottom of the table, the standard deviation of the site-specific effects is 0.608. As with the unconditional estimates, the largest $\widehat{\mu}$ is 2.4 times larger than the smallest. We perform an F test of the joint hypothesis that all site effects μ are equal. The F statistic is 4.24, and the hypothesis is rejected with a p-value of less than 10^{-6}.

Column II controls for the frequency with which the Home Energy Reports are delivered. Because the regression includes site dummies, the frequency controls are identified entirely off of the three sites where frequency was randomly assigned within site. The omitted frequency is quarterly, and monthly treatment has a 0.47 percent larger ATE, with bimonthly having an imprecisely-estimated 0.05 percent larger ATE. These controls shift the site effects μ to their predicted values if all reports were delivered quarterly. This actually slightly increases the standard deviation of the μ's, although it decreases the F statistic because it increases the standard errors on the estimated $\widehat{\mu}$'s.

Column III controls for an indicator variable for whether the experiment is "Immature," or has been running for less than six months. Allcott (2011) shows that the treatment effects tend to increase over the first six months, meaning that an experiment that has been running for one year will mechanically have a smaller ATE than an experiment that has been running for two years. However, the results in Column III show that controlling for this does not substantially change the standard deviation of the site dummies or the F statistic.

In Column IV, we control for the interaction of heating and cooling degree days with the treatment effect. One additional average cooling degree day increases the treatment effect in absolute value by 0.073 percentage points. Heating degree days also appear to increase the treatment effect, but the coefficient is much smaller and is not statistically distinguishable from zero. Although the treatment effect is not larger during colder periods, energy use is of course larger: the $\widehat{\beta}_r$ coefficients, which are omitted from the table to conserve space, show that one additional average heating degree increases energy use by one to four percentage points, depending on the site.

At the bottom of Column IV in Table 5, we see that controlling for weather increases the dispersion of the $\widehat{\mu}$'s. The primary reason is that experiments 10 and 11, which were in relatively hot climates with large average Cooling Degree Days, had relatively small unconditional average treatment effects. After conditioning on weather, the unexplained residual site effects are even smaller relative to the other sites. Of course, it is also possible that the true functional form of the relationship between weather and the treatment effect is not linear. More generally, all columns of Table 5 could be improved if they reflected the true functional forms. We do not have enough data to estimate this relationship between weather and the ATE more non-parametrically, however, and

households within the site, missing values are replaced with the site mean. Otherwise, it is replaced with the mean value across all households in all 14 sites.

it seems unlikely that our qualitative conclusion that unconfounded location does not hold would be affected by using different functional forms.

Column V controls for weather as well as all other observable characteristics from Tables 2 and 3. This increases the standard deviation of the $\widehat{\mu}$'s to 0.872 and inicreases the F-statistic for the test of equal site effects to 7.86. Column VI controls for all X variables from Columns I through V, which further increases the dispersion of the $\widehat{\mu}$'s and the value of the F-statistic.

Even after pooling across all experiments, the α parameters may be imprecisely estimated in finite sample, and extrapolating based on an imprecisely-estimated model can actually worsen the predictions. This is an additional reason why controlling for observables X could increase the dispersion of $\widehat{\mu}$'s. We therefore include Column VII, which controls only for the X variables that are statistically significantly correlated with the treatment effect with 90 percent confidence in Column VI. The dispersion of the site effects and the value of the F-statistic are both higher in Column VII than in Column VI, which suggests that imprecisely-estimated α's are not the primary reason why controlling for X increases the dispersion of the site effects $\widehat{\mu}$. Notice also that the estimated α coefficients are very stable across specifications.

One implication of the failure of unconfounded location is that this makes it difficult to say anything formal about uncertainty over a parameter in a Target population. In the Sample site, the standard error on the estimated ATE properly captures uncertainty. When unconfounded location holds, the standard error on the extrapolated parameter estimate is similarly an appropriate measure of uncertainty. In the presence of unobserved site effects, however, an analyst typically takes (either formally or informally) one of two approaches when generalizing a site-specific result. First, the analyst might rely on informal theoretical arguments about what are unobservables Z and the magnitude of the difference between $\gamma E[Z_{ir}|r]$ in Sample and Target. Second, the analyst might argue that Z is unknown, but the variance of $\gamma E[Z_{ir}|r]$ across sites is small enough that a site-specific estimate is of general interest.

In the OPOWER setting, this first approach is difficult, as it is not obvious what factors Z cause site-level heterogeneity. How large is the variance in site effects? A different way of framing the above F-tests is to compare the uncertainty over a Target population parameter generated by sampling error to the uncertainty generated by unobserved site-level heterogeneity. As shown in Table 5, the standard deviation in site effects is around 0.6 percent, which is twice the simple average of the standard errors on the 14 estimated Sample ATEs. This means that in this context, with these sample sizes and unobserved site effects, unobserved site-level heterogeneity causes twice the parameter uncertainty in a Target population as classical sampling error.

The key empirical result from this section is that in the OPOWER context, there is economically-significant heterogeneity in treatment effects across sites, and this heterogeneity is not explained by individually-varying observable characteristics. This result is despite the fact that the treatment is highly consistent across experiments and we observe a potentially-promising set of observable characteristics that could moderate the treatment effect. In this context, the assumption of unconfounded location is not valid. This means that it is difficult for any one specific potential partner to use past results to predict efficacy in their setting. In the next section, we test whether past results could be used to predict the average efficacy across a number of potential partners.

5 OPOWER Partner Selection Bias

In this section, we test the assumption of expected unconfounded location. The ideal way to test this assumption would be to estimate Average Treatment Effects in the set of partner sites and compare the distribution to the distribution of ATEs in non-partner sites. The problem is that by definition, there have been no experiments in non-partner sites, so these ATEs cannot be estimated.

Instead, we test whether OPOWER partners differ from non-partners on site-level observable characteristics that are correlated with the treatment effect. These site-level characteristics will not vary at the individual level within a site, so it would not be possible to control for them when extrapolating from one site to another. In that sense, they are unobservables in the context of the model in Section 2. To avoid confusion with individual-level characteristics X and Z, we denote these site level observable characteristics by W.

Below, we detail the observable characteristics of OPOWER partners relative to other utilities, estimate selection probabilities, and show that these selection probabilities are robustly correlated with treatment effects within the set of 14 existing partners. This correlation implies that there is a "partner selection bias" that means that the average ATE for existing partners is not an unbiased estimator of the average ATE that would be expected in future implementations.

5.1 Partner Data

In order to examine selection into partnerships with OPOWER, we gathered a dataset of electric utility characteristics. Our sample includes the 939 electric utilities in the U.S. with more than 10,000 residential customers. There are another 2100 utilities that are smaller, most of which are rural cooperatives or small firms in states with competitive retail electricity markets, but we omit these because OPOWER has no partners with fewer than 10,000 residential customers. 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.

We focus on characteristics that could be correlated with selection into treatment and/or the site's treatment effect. These could include the utility's ownership structure (Cooperative, private "Investor-Owned Utility" (IOU), Municipal, or Other Government), average residential energy usage, number of residential consumers, and prices. These variables were all gathered from a regulatory filing called Energy Information Administration (EIA) Form 861. An existing focus on energy conservation and green energy would presumably increase selection probability and be correlated with the population's receptiveness to the treatment. We therefore include the utility's spending and total estimated effects of energy conservation programs and the percent of customers that have voluntarily enrolled in "green pricing programs" that sell renewably-generated energy at a premium price, also from Form 861.

Similarly, utilities with customer populations that are higher-income, better educated, and more liberal might be more interested in and responsive to the treatment. We thus include state-level average income, the percent of residents with a college degree, and the percent of voters that voted for a Democratic candidate for the House of Representatives in elections between 2000 and 2008. These three variables are from the U.S. Census (2010a, 2010b, 2010c). Finally, since state energy regulation has been an important driver of OPOWER's business, we include whether the state has an Energy Efficiency Portfolio Standard (EERS), using data from the Pew Center on Global Climate Change (2010). Using data from the the U.S. Department of Energy (2010), we also include whether the state has a Renewables Portfolio Standard (RPS), which is a policy similar to the EERS that requires utilities to purchase a given percentage of its energy from renewable generation sources such as wind, solar, or geothermal facilities.

Column 1 of Table 6 presents the means and standard deviations of each of these characteristics across the 939 utilities in the sample. Columns 2 and 3 show the same statistics for OPOWER's 45 partners and 894 non-partners, respectively. Column 4 tests whether the characteristics are balanced between the two groups. Eleven 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 site-level observables W_r.

5.2 Tests and Results

Table 7 presents a series of tests for whether observable characteristics that are correlated with selection into partnership are also correlated with the treatment effect. The first three columns focus on each observable characteristic in isolation, while the latter four columns include multivariate selection equations.

5.2.1 Univariate Selection

Columns I through III demonstrate whether each individual observed site-level characteristic W_r suggests positive or negative selection. Column I presents the univariate correlation of the characteristic W_r with $\widehat{\tau}_r$ across the 14 existing experiments. This is estimated with the following regression:

$$\widehat{\tau} = \eta W + \eta_0 + \epsilon \tag{12}$$

We use the $\widehat{\tau}_r$ from Column I of Table 5, although the results that follow are robust to using site effects conditional on different sets of X characteristics given that these X's explain little of the variability in τ. Recall that the treatment reduces energy demand, so "stronger" ATEs are more negative. Column I shows that the ATE is statistically significantly stronger for Cooperatives, less strong for Investor-Owned Utilities and utilities with larger customer bases, stronger in states that are richer and better educated, but less strong in states with more democratic voters. All 14 existing experiments are in states with Energy Efficiency Resource Standards and Renewables Portfolio Standards, so a univariate $\widehat{\eta}$ cannot be calculated for these variables. Standard errors are robust and clustered by utility.

Column II estimates a univariate probit model using the following equation:

$$\Pr(T = 1|W) = \Phi\left(\rho W + \rho_0\right) \tag{13}$$

Column II shows that Cooperatives are statistically significantly less likely to partner with OPOWER, as are utilities with higher mean usage per customer per year. Investor-Owned Utilities and utilities with more customers and higher prices are statistically significantly more likely to partner with OPOWER, as are utilities in weathier, better educated, and more Democratic states.

The product of the signs from Columns I and II tells us which direction each individual observable influences partner selection bias. For example, Investor-Owned Utilities have weaker ATEs in the 14 existing sites, and they are more likely to partner with OPOWER, so this variable in isolation suggests negative selection. A more direct way to see how each observable W influences selection is in Column III, which presents the coefficient of a regression of the ATE on the selection probability:

$$\widehat{\tau} = \theta \widehat{\Pr}(T = 1|W) + \theta_0 + \varepsilon \tag{14}$$

Column III has robust standard errors, clustered by utility, and also accounts for the uncertainty in the estimate of $\widehat{\Pr}(T = 1|W)$ using a procedure adapted from Murphy and Topel (1985). The results of Column III show that two variables, the indicator for Investor-Owned Utility and the log of the number of residential consumers, statistically significantly suggest negative selection. Another nine variables have statistically insignificant indications of negative selection, while two give highly imprecise statistically insignificant indications of positive selection.

5.2.2 Multivariate Selection

The above univariate selection estimations were presented to build intuition around how each individual site-level observable is associated with the treatment effect and the selection probability.

We now present results of multivariate selection equations. Columns IV, V, and VI of Table 7 estimate Equation (13) with different configurations of observables. Column IV estimates the selection probit with all observables W, while Column V includes only the observables that are statistically significant in Column IV, and Column VI chooses a third configuration.

IOUs and Municipal utilities are statistically significantly more likely to partner with OPOWER, relative to Coops and utilities owned by Other Government entities. Utilities with higher mean energy usage per residential customer are less likely to partner with OPOWER. Larger utilities in states with higher median income are more likely to partner. Across the four columns, the magnitudes and statistical significance of these correlations are robust to these different subsets of observable characteristics.

From each of these three selection equations, a selection probability $\widehat{\Pr}(T = 1|W)$ is fitted. The bottom three rows in the table then present the estimated $\widehat{\theta}$ from Equation (13) using the fitted selection probability from each column. In each of Columns IV, V, and VI, the $\widehat{\theta}$ is positive and statistically significantly different from zero. Again recalling that larger treatment effects are more negative, a coefficient of 1.8 means that a ten percentage point increase in selection probability is associated with about a 0.18 percentage point weaker ATE, which is just under one tenth of the mean of the 14 ATEs. This is robust evidence of negative selection on observables: utilities whose observable characteristics make them more likely to have partnered with OPOWER have smaller treatment effects.

Column VII further tests the robustness of the negative selection finding by testing whether it is driven by any one variable. In each row of this column, the selection probit in Equation (13) is estimated using all other observables except the one corresponding to the row. The fitted selection probability is then used to estimate Equation (14), and Column VII presents the corresponding $\widehat{\theta}$. For example, in the first row, the selection probit is estimated using all observables except the $1(Coop)$ indicator variable. Using that subset of observables, $\widehat{\theta} = 1.68$, meaning that the negative selection result is robust to the exclusion of $1(Coop)$. The coefficients in Column VII are highly robust: all but one are between 1.60 and 1.72, while excluding the log of the number of residential customers indicates much stronger but more imprecisely estimated negative selection.

Figure 6 is a graphical presentation of these results. On the horizontal axis is the fitted selection probability for each of the 14 existing experiments, using the selection equation estimated in Column IV of Table 7. On the vertical axis is the ATE. The slope of the best fit line is $\theta = 1.65$, as reported at the bottom of Column IV.

When presented with these statistical results, OPOWER's management suggested that negative selection is driven by a version of the "diminishing returns bias" discussed in Section 2. Utilities that are more likely to partner with OPOWER are also likely to be running other energy conservation programs. These other programs have eliminated many of the lowest-cost opportunities for homeowners to conserve energy. Some of the empirical results suggest this: utilities with lower average usage may have reduced this usage partially through previous energy conservation programs.

These utilities have weaker ATEs in the OPOWER program, yet they are more likely to partner with the company. Similarly, the EIA Form 861 data show that Investor-Owned Utilities spend much more on energy conservation programs, which could render the marginal program offered by OPOWER less effective. IOUs have weaker treatment effects, but they are more likely to partner with OPOWER.

The takeaway from this section is that OPOWER's partners differ on site-level observable characteristics that are correlated with the treatment effect. Unless the relationship between these site-level characteristics and the treatment effect can be precisely estimated and partialled out, the distribution of ATEs in existing sites is unlikely to reflect the distribution of ATEs in future potential sites. Even if there were enough partner sites to precisely estimate how treatment effects vary with the site-level observables W, the fact that there appears to be selection on observables suggests that there could also be selection on unobservables.

6 Partner Selection in Microfinance

We imagine that the potential for partner selection bias is not limited to the OPOWER example. As one further empirical example, we examine what types of microfinancial institutions (MFIs) partner with US-based academic organizations to do randomized field experiments. Unlike with OPOWER, we do not have a set of ATEs for the same treatment across different MFIs, as many microfinance field experiments are tests of more nuanced hypotheses instead of straightforward impact evaluations. Therefore, we do not show, as we had with OPOWER, that treatment effects are correlated with selection probabilities. Instead, we show that MFIs that carry out randomized experiments differ on observables that could influence the results of these experiments.

Aside from being an area of general interest to economists, microfinance is also a convenient area to quantitatively examine partner selection, for two reasons. First, there are many microfinance field experiments with many partners. Third, there is a centralized global database of MFIs that both defines the set of potential partners and contains relevant partner characteristics.

The database we use is called the Microfinance Information Exchange (MIX), which includes information on the characteristics and performance of 1903 MFIs in 115 countries. We consider characteristics that might be correlated with the outcomes of different field experiments, including Non-Profit status, the age of the organization, number of borrowers, percent of borrowers who are women, average loan balance, MFI expenditures per borrower, ratio of borrowers to staff members, and repayment rates. Of course, the characteristics correlated with the treatment effect will vary depending on the treatment, whether it is the presentation of consumer credit offer letters as in Bertrand *et al.* (2010), variation in consumer loan interest rates as in Karlan and Zinman (2009), or the opportunity to take out a microfinance loan as in Banerjee, Duflo, Glennerster, and Kinnan (2009). Table 8 presents descriptive statistics.

For each MFI in the database, we then determined whether it had partnered with major academic groups to carry out a randomized experiment. This was done using the lists of 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 controlled trials.

Column 1 of Table 9 shows the unconditional correlation of each MFI characteristic and whether an MFI is an experiment partner. Columns 2-5 of the same table present probit regressions of the "experiment partner" dummy variable on various sets of characteristics. These columns are analogous to Equation (13) in the previous section.

The signs and magnitudes of the correlations are robust, although the significance levels vary across the five columns. The most robust results are that for-profit, larger, and older MFIs are more likely to carry out randomized trials. This is quite natural: experiments require stable, well-managed partners and large sample sizes. MFIs with a larger share of women borrowers and smaller average loan balances also appear to be more likely to run experiments, and both of these factors could affect baseline repayment rates. There is also suggestive evidence that MFIs with more borrowers per staff member and, relatedly, lower cost per borrower are more likely to be experiment partners. The number of staff per borrower could affect baseline repayment rates through improved monitoring and could also influence the efficacy of interventions that require attention from MFI personnel. A final suggestive correlation, which is also not statistically significant, suggests that partner MFIs have less Portfolio at Risk, which corresponds to better 30-day repayment rates.

The bottom row of the table presents Chi-Squared tests of whether observables predict selection into partnership. In all four probit regressions, this test strongly rejects that partner MFIs do not differ on observables. Of course, whether this implies that a treatment effect or comparative static differs between a Sample and the set of potential Target sites depends on the treatment or theoretical prediction in question. This empirical evidence simply suggests that researchers should continue to exercise caution in extrapolating results from past experiments to the broader population of microfinancial institutions.

7 F-Test of Sub-Site Heterogeneity

7.1 Overview

In Section 4, we showed that unconfounded location does not hold in the OPOWER setting. Documenting this required having data on multiple sites. The difficulty of generalizing, of course, is that we do not know the parameter value in the Target. We must decide whether to assume unconfounded location, without being able to explicitly test this assumption. In this section, we present a suggestive test of whether unconfounded location might hold.

The test is an F-test for whether the treatment effect varies by "sub-sites" within a site. The intuition is that unconfounded location requires that observable characteristics capture all heterogeneity between sites. But a "site" can be arbitarily defined: for example, an OPOWER program could be randomized across an entire state population, a county, a city, or a neighborhood. If the program were implemented at the state level and there is unexplained heterogeneity across cities within the state, this means that there are unobserved factors that affect the treatment effect within different sub-sites within the state. These same unobserved factors could also confound extrapolation from the one Sample state to an adjacent state. Unexplained geographic heterogeneity within the Sample site is *suggestive* of unobserved heterogeneity across sites.

The sub-site is different depending on the context. In the OPOWER experiments, a sub-site can be a county, zip code, or Census tract. In an education experiment carried out across a set of schools, a sub-site could be a city, a school, or a teacher. In a job training program implemented at the state level, a sub-site could be a county, a city, or an individual training center.

To carry out the test, define a vector of sub-site indicator variables M. Then run the following regression, which interacts M and observable characteristics X with the treatment indicator and controls for lower-order interactions:

$$Y_i = [\lambda M_i + \alpha X_i] \cdot T_i + \pi M_i + \beta X_i + \varepsilon_i \tag{15}$$

The F-test of sub-site heterogeneity is a test of the joint hypothesis that all λ's are equal. Notice that Equation (15) is the simplest implementation of this regression, and there are other possible versions. For example, the regression we use below to implement the test in the OPOWER context involves pre- and post-treatment data and household fixed effects.

Of course, the unconfounded location assumption cannot be tested directly, and this test is only a suggestive test. False failures to reject are certainly possible. An important reason for a false failure to reject is that geographic heterogeneity could occur at a higher level than the site. A treatment with homogeneous effects across sub-sites in Kenya could still have different effects in India. Similarly, there could be partner-level effects that affect the entire set of sub-sites. A treatment carefully implemented across many sub-sites by a partner in California might be poorly implemented by another potential partner in New Jersey. Furthermore, if the statistical power of the test is low, perhaps because there are few observations within each sub-site, the test could also falsely fail to reject. Therefore, a failure to reject is not good evidence that unconfounded location holds.

However, the converse is more likely to be true: rejecting equality of λ is stronger evidence that unconfounded location does not hold. Certainly, false rejections are possible: if Sample and Target both have sub-site heterogeneity, it is possible that the distribution of sub-site heterogeneity is identical in the two places. For example, if a treatment effect is a function of teacher quality, and two school districts have the same distribution of teacher quality, the test could reject equal λ's, yet the average treatment effect in each district could be the same. However, rejecting equality puts a burden on the analyst who wants to extrapolate from a Sample to a different Target: the analyst must argue that the distribution of unobserved sub-site effects is the same.

While this test is only suggestive, so are common tests of internal validity. The overidentification test has false rejections, when all instruments are valid but act on different sets of compliers with different Local Average Treatment Effects, as well as false failures to reject, when all instruments are equally biased. In the Regression Discontinuity context, it is common to test whether control variables are discontinuous around the cutoff (Lee and Lemieux 2009). There could be false failures to reject: even if no observable characteristics are discontinuous at the cutoff, there could be unobservables that are. In principle, there could also be false rejections: if there are discontinuities in observable variables but not in unobservables, in some contexts these might be controlled for, allowing a consistent estimate of the treatment effect at the cutoff.

7.2 Sub-Site Heterogeneity in OPOWER Experiments

We now present the results of the F-test for sub-site heterogeneity in the context of the 14 OPOWER experiments. In this context, the empirical implementation of Equation (15) also includes household fixed effects, time controls, and interactions of X and M with the post-treatment indicator P_{it}:

$$Y_{it} = [\lambda M_i + \alpha X_i] \cdot T_i P_{it} + [\pi M_i + \beta X_i] \cdot P_{it} + \mu_t + \upsilon_i + \varepsilon_{it} \tag{16}$$

In separate tests, we define sub-sites at two different levels: Census tract and zip code.[10] In each experiment, we control for the set of observed X variables detailed in Tables 2 and 3. Standard errors are robust and clustered by household.

As a visual illustration of the test, Figure 6 presents the the distribution of sub-site heterogeneity in Experiment 3, when a "sub-site" is defined to be each of the 81 Census tracts within the site. In other words, Figure 6 plots the elements of the $\widehat{\lambda}$ vector, normalized to have mean zero. After this normalization, the sub-site point estimates range from -7.5 to 5.2 percent, with a standard deviation of 1.9 percent. Of course, this figure is illustrative only: it is not a statistical test of whether $\lambda = 0$. This is because the distribution of estimated sub-site ATEs depends both on the true underlying distribution and the precision with which they are estimated.

Table 10 presents the formal results of the F-tests of sub-site heterogeneity for each of the 14 OPOWER sites. Columns 3 and 4 present the results for sub-sites defined by zip codes, while Columns 5 and 6 present the results for sub-sites defined by Census tracts. At each site, the results tend to be fairly consistent between the two levels of geographical aggregation. At sites 6 and 13, the F-tests reject that λ's are equal with greater than 90 percent confidence at both levels of aggregation, and at sites 5 and 14, the F-tests reject equality at one at one level of aggregation

[10] In several cases, we aggregate sub-sites when the total number of households in a sub-site was less than 100. This is an ad-hoc approach to address concerns over non-normality of the error distribution in finite sample.

and nearly reject (with 89 percent confidence) at the other level. At site 3, the test rejects equality at the tract level but not at the zip code level. At all other sites, however, the tests fail to reject equality.

Why does the F-test reject at these five sites and not others? One likely explanation is that these experiments have fewer X variables which can control for observable treatment effect heterogeneity. As shown in Tables 2 and 3, these sites have slightly poorer coverage; in particular, the First Comparison variable is not available in sites 13 and 14, and this variable is strongly correlated with the treatment effect. Including additional X variables tends to decrease the precision with which the sub-site effects are estimated, making it less likely that the F-test will reject that the λ's are equal.

The results illustrate the possibility of false failures to reject: as we showed in Section 4, unconfounded location does not hold for these experiments, yet in many cases the F-test fails to reject equality of the λ's. These failures to reject likely result either from insufficient power, meaning that there are are too few observations to precisely estimate the sub-site heterogeneity, or from the fact that there are unobservables Z with more variation between sites than within sites. However, in the several sites where the F-test does reject equality, this result would correctly force the analyst to proceed with caution in extrapolating results to other sites.

8 Conclusion

While external validity has long been of concern to empiricists in economics and other fields, there have been few opportunities to quantitatively assess the ability to generalize parameter estimates across settings. This paper exploits a remarkable opportunity by analyzing a series of nearly-identical energy conservation field experiments run by OPOWER in a number of different sites across the U.S. We document significant heterogeneity in treatment effects across sites and show that observable population characteristics explain very little of this variation. Furthermore, we show that the electric utilities that partner with OPOWER differ from those that do not on observable characteristics that correlate with the treatment effect, suggesting a negative selection into the experiments. The data in this setting reject location unconfoundedness and expected location unconfoundedness, the assumptions required for our two notions of external validity.

While our quantitative results are specific to this set of energy conservation experiments, we argue that two messages are of general interest. First, unobservables at the individual level that affect internal validity have received more rigorous attention than unobservables at the population level or in the economic environment that affect external validity. In some settings, these unobservables can substantially affect policy conclusions. Second, randomized experiments can suffer from a partner selection bias mathematically analogous to biases from selection of units into treatment. This can systematically bias experimentally-estimated effects away from the effects that would be realized by partners that do not select into treatment.

What are the implications? First, these results by no means argue against internally-valid estimators. Although the OPOWER experiments are not externally valid in our two senses, Allcott (2011) shows that non-experimental estimates perform extremely poorly in the OPOWER context. Internally-valid estimators from other Sample sites predict true ATEs at a Target site far better than non-experimental estimators from the same Target. Second, when treatment effects are difficult to generalize, these results suggest the importance of internally valid estimators in the Target population of policy interest. For example, each of the OPOWER experiments is at a site where the ATE is of some interest *per se*, as the partner utility decides whether to continue running the program. If the effects of large-scale social programs are of interest, this means that it is especially important to implement the program with a randomized trial, as has been done with conditional cash transfer programs in several countries.

Third, in presenting results of completed projects, researchers can clearly define whether there is a different or larger Target population of policy interest and provide quantitative descriptive statistics and qualitative discussion of how it might differ from the Sample in ways that moderate the treatment effect. Similarly, researchers can discuss how the economic environment and the treatment itself might vary across settings or be different in a setting of particular policy interest. Fourth, researchers can attempt to carry out similar field experiments in multiple locations. The locations would ideally be chosen to lie in different parts of the distribution of factors that moderate the treatment effect. Finally, some have argued that "mechanisms" can in some circumstances be more easily generalized across sites and domains than average treatment effects for specific projects (Deaton 2010, Ludwig, Kling, and Mullainathan 2011). If this were the case, researchers can focus on identifying mechanisms upon which policy decisions hinge and designing empirical studies to tease them out.

References

[1] Aigner, Dennis (1984). "The Welfare Econometrics of Peak-Load Pricing for Electricity." *Journal of Econometrics*, Vol. 26, No. 1-2, pages 1-15.

[2] Allcott, Hunt (2011). "Social Norms and Energy Conservation." *Journal of Public Economics*, in press.

[3] Allcott, Hunt, and Sendhil Mullainathan (2010). "Behavior and Energy Policy." *Science*, Vol. 327, No. 5970 (March 5th).

[4] Altonji, Joseph, Todd Elder, and Christopher Taber (2005). "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy*, Vol. 113, No. 1 (February), pages 151-184.

[5] 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.

[6] 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.

[7] 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).

[8] Attanasio, Orazio, Costas Meghir, and Miguel Szekely (2004). "Using Randomized Experiments and Structural Models for 'Scaling Up': Evidence from the PROGRESA Evaluation." Centre for the Evaluation of Development Policies Working Paper EWP04/03 (May).

[9] 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).

[10] 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.

[11] 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.

[12] Banerjee, Abhijit, Esther Duflo, Rachel Glennerster, and Cynthia Kinnan (2009). "The Miracle of Microfinance? Evidence from a Randomized Evaluation." Working Paper, MIT (May).

[13] 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.

[14] 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*, forthcoming.

[15] 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.

[16] 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.

[17] Campbell, Donald, and Julian Stanley (1966). Experimental and Quasi-Experimental Designs for Research. Chicago, Illinois: Rand McNally.

[18] Card, David, Jochen Kluve, and Andrea Weber (2009). "Active Labor Market Policy Evaluations: A Meta-Analysis." IZA Discussion Paper No. 4002 (February).

[19] Cartwright, Nancy (2007a), "Are RCTs the Gold Standard?" *Biosocieties*, Vol. 2, No. 2 pages 11–20.

[20] Cartwright, Nancy (2007b). Hunting Causes and Using Them: Approaches in Philosophy and Economics. Cambridge: Cambridge University Press.

[21] Cartwright, Nancy (2010). "What are randomized trials good for?" *Philosophical Studies*, Vol. 147, 59–70.

[22] 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.

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

[24] Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik (2009). "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika*, Vol. 96, pages 187–99.

[25] Davis, Matthew (2011). "Behavior and Energy Savings." Working Paper, Environmental Defense Fund (May). http://blogs.edf.org/energyexchange/files/2011/05/BehaviorAndEnergySavings.pdf

[26] Davis, Lucas (2008). "Durable Goods and Residential Demand for Energy and Water: Evidence from a Field Trial." *RAND Journal of Economics*, Vol. 39, No. 2 (Summer), pages 530-546.

[27] Deaton, Angus (2010). "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature*, Vol. 48, No. 2 (June), pages 424–455.

[28] 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.

[29] Duflo, Esther (2004). "Scaling Up and Evaluation." Conference Paper, Annual World Bank Conference on Development Economics.

[30] 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).

[31] Duflo, Esther, Rema Hanna, and Stephen Ryan (2007). "Monitoring Works: Getting Teachers to Come to School." Working Paper, MIT.

[32] Feynman, Richard (1964). "The Great Conservation Principles." Lecture at Cornell University. Accessible from http://research.microsoft.com/apps/tools/tuva/.

[33] Greenberg, David, and Mark Schroder (2004). The Digest of Social Experiments; Third Edition. Washington, DC: Urban Institute Press.

[34] 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).

[35] Heckman, James (1980). "Varieties of Selection Bias." *American Economic Review*, Vol. 80, No. 2, pages 313-318.

[36] 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.

[37] 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.

[38] Heckman, James, and Jeffrey Smith (1995). "Assessing the Case for Social Experiments." *Journal of Economic Perspectives*, Vol. 9, No. 2 (Spring), pages 85-110.

[39] Heckman, James, and Jeffrey Smith (1997). "The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study," NBER Working Paper No. 6105 (July).

[40] Heckman, James, and Sergio Urzua (2010). "Comparing IV with Structural Models: What Simple IV Can and Cannot Identify." *Journal of Econometrics*, Vol. 156, No. 1, pages 27-37.

[41] 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.

[42] Heckman, James, and Edward Vytlacil (2001). "Policy-Relevant Treatment Effects." *American Economic Review*, Vol. 91, No. 2 (May), pages 107–111.

[43] Heckman, James, and Edward Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation." *Econometrica*, Vol. 73, No. 3 (May), pages 669–738.

[44] 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.

[45] 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.

[46] Hirano, Keisuke, Guido W. Imbens, and Geert Ridder (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score." *Econometrica*, Vol. 71, No. 4, pages 1161–89.

[47] Holla, Alaka, and Michael Kremer (2009). "Pricing and Access: Lessons from Randomized Evaluations in Education and Health." Center for Global Development Working Paper Number 158 (January).

[48] Horvitz, D. G., and D. J. Thompson (1952). “A Generalization of Sampling without Replacement from a Finite Universe.” *Journal of the American Statistical Association*, Vol. 47, No. 260, pages 663–85.

[49] 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).

[50] 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.

[51] 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.

[52] Imbens, Guido (2010). "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, Vol. 48 (June), pages 399-423.

[53] Imbens, Guido, and Joshua Angrist (1994) “Identification and Estimation of Local Average Treatment Effects.” *Econometrica*, Vol. 62, No. 2, pages 467-475.

[54] Imbens, Guido, and Jeffrey Wooldridge (2009). "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature*, Vol. 47, No. 1, pages 5-86.

[55] 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).

[56] Lalonde, Robert (1986). "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review*, Vol. 76, No. 4, pages 604-620.

[57] Lee, David, and Thomas Lemieux (2009). "Regression Discontinuity Designs in Economics." NBER Working Paper 14723 (February).

[58] Levitt, Steven, and John List (2007). "What Do Field Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives*, Vol. 21, No. 2 (Spring), pages 153-174.

[59] 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.

[60] Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan (2011). "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*, forthcoming.

[61] Manski, Charles, and Irwin Garfinkel (1992). "Introduction." In Charles Manski and Irwin Garfinkel (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press: Cambridge, MA, pages 1-24.

[62] Manski, Charles (2011). "Policy Analysis with Incredible Certitude." *The Economic Journal*, forthcoming.

[63] Meyer, Bruce (1995). "Lessons from U.S. Unemployment Insurance Experiments." *Journal of Economic Literature*, Vol. 33, No. 1 (March), pages 91-131.

[64] 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.

[65] 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.

[66] Nevo, Aviv, and Michael Whinston (2010). "Taking the Dogma out of Econometrics. Structural Modeling and Credible Inference." *Journal of Economic Perspectives*, Vol. 24, No. 2 (Spring), pages 69-82.

[67] Nolan, Jessica, Wesley Schultz, Robert Cialdini, Noah Goldstein, and Vladas Griskevicius (2008). “Normative Influence is Underdetected.” *Personality and Social Psychology Bulletin*, Vol. 34, pages 913-923.

[68] Parker, Ian (2010). "The Poverty Lab." *The New Yorker*, May 17th, page 79.

[69] Pew Center on Global Climate Change (2010). "Energy Efficiency Standards and Targets." http://www.pewclimate.org/what_s_being_done/in_the_states/efficiency_resource.cfm

[70] Pritchett, Lant (2002). "It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation." Working Paper, Kennedy School of Government (April).

[71] Reiss, Peter, and Matthew White (2008). "What Changes Energy Consumption? Prices and Public Pressure." *RAND Journal of Economics*, Vol. 39, No. 3 (Autumn), pages 636-663.

[72] 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.

[73] 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.

[74] 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.

[75] Rubin, Donald (1990). “Formal Mode of Statistical Inference for Causal Effects.” *Journal of Statistical Planning and Inference*, Vol. 25, No. 3, pages 279–292.

[76] Schultz, Wesley, Jessica Nolan, Robert Cialdini, Noah Goldstein, and Vladas Griskevicius (2007). "The Constructive, Destructive, and Reconstructive Power of Social Norms." *Psychological Science*, Vol. 18, pages 429 –434.

[77] Todd, Pettra, and Kenneth Wolpin (2006). "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review*, Vol. 96, No. 5 (December), pages 1384-1417.

[78] 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.

[79] 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

[80] 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

[81] U.S. Department of Energy (2010). "States with Renewable Portfolio Standards." http://apps1.eere.energy.gov/states/maps/renewable_portfolio_states.cfm.

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

[83] Worrall, John (2007). "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass*, Vol. 2, No. 6, pages 981-1022.

Tables

Table 1: Overview of Experiments

Experiment			**N**		
Number	*Region*	*Start Date*	*Households*	*Treated*	*Obs*
1	Urban Midwest	July, 2009	37,484	18,790	1,264,375
2	Urban Midwest	July, 2009	56,187	28,027	1,873,482
3	Rural Midwest	January, 2009	78,273	39,024	3,421,306
4	Suburban Mountain	October, 2009	11,612	7,254	394,525
5	Suburban Mountain	October, 2009	27,237	16,947	914,344
6	West Coast	October, 2009	24,940	23,906	570,386
7	Rural Midwest	April, 2009	17,889	9,861	794,457
8	Urban Northeast	September, 2009	49,671	24,808	1,712,530
9	West Coast	October, 2008	79,229	34,893	3,121,879
10	West Coast	January, 2009	25,211	5,570	985,148
11	West Coast	January, 2009	17,849	3,852	672,629
12	West Coast	September, 2009	39,336	19,663	671,990
13	West Coast	March, 2008	59,666	24,761	2,543,372
14	West Coast	April, 2008	24,293	9,903	1,036,768
Combined		March, 2008	548,877	267,259	19,977,191

Table 2: Weather and Occupant Descriptive Statistics

Expr	*First Comparison*	*Mean Age*	*Median Income ($000s)*	*Pct White*	*HDDs*	*CDDs*
1	-1.28 (1.88)	50.5 (5.3)	84.5 (30.3)	0.85 (0.21)	13.5 (13.4)	2.7 (3.9)
2	-0.47 (1.17)	49.4 (5.7)	68.2 (25.9)	0.76 (0.3)	13.6 (13.4)	2.7 (3.9)
3	-0.05 (1.41)	45.1 (2.4)	62.2 (9.4)	0.96 (0.02)	16.6 (16.4)	1.7 (2.5)
4	-1.08 (1.55)	43.9 (4.2)	56.2 (16.5)	0.91 (0.04)	17.1 (14.8)	1.7 (2.7)
5	0.3 (0.93)	43.3 (4.4)	50.5 (16.6)	0.9 (0.05)	17.5 (14.8)	1.7 (2.7)
6	-0.3 (1.23)	50.3 (3)	49.5 (21.6)	0.66 (0.1)	4.1 (3.7)	2.3 (2.6)
7	0.09 (1.11)	52.9 (2.1)	38.8 (6.5)	0.95 (0.07)	18.8 (17.7)	0.7 (1.1)
8	-0.28 (1.26)	51 (2.7)	65.9 (23)	0.93 (0.1)	13.6 (12.4)	2.4 (3.6)
9	-0.26 (1.18)	47.2 (3.5)	71.5 (19.8)	0.8 (0.09)	13.3 (8.4)	0.5 (1.2)
10	0.05 (1.35)	55.3 (8)	43.8 (12.8)	0.79 (0.14)	2.7 (4.1)	10 (10.3)
11	0.27 (1.24)	57 (7.5)	42.1 (16.6)	0.81 (0.15)	2.8 (4.2)	10.3 (10.5)
12	-0.41 (1.63)	48.1 (3.5)	54.1 (13.8)	0.72 (0.22)	12.5 (6.2)	0.4 (0.6)
13	-	49.7 (4.4)	60.7 (14)	0.77 (0.15)	5.9 (6.6)	2.9 (3.1)
14	-	49.9 (4.2)	56.9 (12.7)	0.77 (0.15)	6 (6.64)	3 (3.2)

Experiment-level means. Standard deviations in parenthesis.

Table 3: House Descriptive Statistics

Expr	*1(Elec Heat)*	*House Age*	*Value ($000's)*	*1(Pool)*	*1(Rent)*	*1(Single Fam)*	*Sq Feet (000s)*
1	-	44.6 (20.6)	-	-	0.05 (0.2)	0.89 (0.32)	2.76 (1.27)
2	-	48.6 (18.2)	-	-	0.1 (0.27)	0.75 (0.44)	-
3	-	31.6 (28)	394 (139)	-	-	-	1.66 (0.45)
4	0.2 (0.35)	23.4 (15.3)	-	-	0.23 (0.42)	0.87 (0.34)	2.25 (0.76)
5	0.11 (0.22)	26.5 (17.9)	-	-	0.36 (0.48)	0.7 (0.46)	1.91 (0.53)
6	-	59.2 (16.7)	-	0.1 (0.31)	0.35 (0.4)	0.5 (0.5)	1.69 (0.52)
7	0.31 (0.46)	-	-	-	0.05 (0.21)	-	-
8	-	58.7 (36.5)	-	0.02 (0.15)	0.06 (0.22)	-	2.03 (0.72)
9	0.07 (0.25)	31.2 (15.6)	361 (176)	-	0.03 (0.16)	-	2.2 (0.67)
10	-	27.1 (15.3)	-	0.37 (0.48)	0.01 (0.08)	1 (0)	1.95 (0.67)
11	-	28.6 (6.6)	-	0.06 (0.23)	-	0.08 (0.27)	1.74 (0.51)
12	0.17 (0.38)	65.1 (25.4)	437 (293)	-	0.06 (0.22)	-	1.83 (0.77)
13	0.32 (0.47)	34.1 (16.4)	229 (149)	0.28 (0.45)	0.01 (0.1)	-	1.84 (0.6)
14	0.12 (0.32)	43.1 (20.9)	174 (117)	0.04 (0.19)	0.01 (0.09)	-	1.49 (0.43)

Experiment-level means. Standard deviations in parenthesis.

Table 4: Experimental ATEs and Cost Effectiveness

Experiment	**ATEs (%)**			**CE (c/kWh)**		
Number	*Monthly*	*BiMonthly*	*Quarterly*	*Monthly*	*BiMonthly*	*Quarterly*
1	-1.83 (0.2)	-	-	2.02 (-0.22) -	-	
2	-	-1.4 (0.19)	-1.37 (0.19)	-	4.09 (-0.56)	3.8 (-0.53)
3	-2.72 (0.18)	-	-2.26 (0.21)	2.64 (-0.17)	-	2.04 (-0.19)
4	-	-2.7 (0.44)	-	-	1.82 (-0.3)	-
5	-	-	-1.64 (0.33)	-	-	5.36 (-1.08)
6	-	-2.48 (0.25)	-	-	3.65 (-0.37)	-
7	-	-3.32 (0.54)	-	-	1.28 (-0.21)	-
8	-	-1.63 (0.15)	-	-	3.67 (-0.34)	-
9	-1.96 (0.14)	-	-1.49 (0.2)	4.02 (-0.29)	-	2.99 (-0.4)
10	-1.39 (0.34)	-	-	4.47 (-1.09)	-	-
11	-	-	-1.44 (0.51)	-	-	5.33 (-1.89)
12	-	-1.89 (0.21)	-	-	2.27 (-0.25)	-
13	-3.14 (0.37)	-	-	2.12 (-0.25)	-	-
14	-	-	-1.84 (0.43)	-	-	4.7 (-1.1)
Mean	-2.21	-2.24	-1.67	3.05	2.80	4.04

Standard errors in parenthesis.

Table 5: Tests for Site Effects

	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
TxPostx(Experiment 1)	-1.84	-1.37	-1.97	-1.54	3.82	4.12	4.19
	(0.20)	(0.26)	(0.21)	(0.25)	(0.92)	(0.97)	(0.94)
TxPostx(Experiment 2)	-1.38	-1.35	-1.53	-1.11	2.85	2.82	2.76
	(0.16)	(0.19)	(0.17)	(0.21)	(0.86)	(0.90)	(0.87)
TxPostx(Experiment 3)	-2.54	-2.25	-2.65	-2.33	0.88	1.21	1.08
	(0.15)	(0.18)	(0.16)	(0.23)	(0.76)	(0.83)	(0.77)
TxPostx(Experiment 4)	-2.70	-2.66	-2.86	-2.45	2.13	2.00	2.07
	(0.44)	(0.49)	(0.44)	(0.47)	(0.90)	(0.98)	(0.93)
TxPostx(Experiment 5)	-1.63	-1.63	-1.82	-1.41	1.13	0.89	0.95
	(0.33)	(0.33)	(0.34)	(0.37)	(0.80)	(0.87)	(0.81)
TxPostx(Experiment 6)	-2.49	-2.45	-2.71	-2.30	1.12	0.93	1.02
	(0.25)	(0.33)	(0.25)	(0.26)	(0.85)	(0.93)	(0.89)
TxPostx(Experiment 7)	-3.32	-3.28	-3.47	-3.21	1.00	0.89	0.95
	(0.54)	(0.58)	(0.54)	(0.60)	(0.96)	(1.03)	(0.98)
TxPostx(Experiment 8)	-1.62	-1.58	-1.79	-1.38	2.29	2.27	2.22
	(0.15)	(0.26)	(0.15)	(0.20)	(0.86)	(0.94)	(0.90)
TxPostx(Experiment 9)	-1.82	-1.48	-1.91	-1.71	1.95	2.19	2.22
	(0.13)	(0.17)	(0.13)	(0.17)	(0.79)	(0.83)	(0.80)
TxPostx(Experiment 10)	-1.48	-1.01	-1.62	-0.66	3.35	3.50	3.68
	(0.34)	(0.38)	(0.34)	(0.46)	(1.00)	(1.05)	(1.03)
TxPostx(Experiment 11)	-1.54	-1.54	-1.69	-0.68	1.91	1.81	1.76
	(0.52)	(0.52)	(0.52)	(0.59)	(1.06)	(1.10)	(1.08)
TxPostx(Experiment 12)	-1.86	-1.82	-2.05	-1.80	2.00	2.06	1.93
	(0.21)	(0.31)	(0.21)	(0.23)	(0.82)	(0.89)	(0.85)
TxPostx(Experiment 13)	-2.70	-2.22	-2.80	-2.47	1.63	1.83	2.05
	(0.42)	(0.45)	(0.41)	(0.40)	(0.85)	(0.89)	(0.86)
TxPostx(Experiment 14)	-1.43	-1.43	-1.53	-1.28	2.19	1.98	2.11
	(0.42)	(0.42)	(0.42)	(0.43)	(0.91)	(0.94)	(0.92)

Table 5 continues on the next page.

Table 5 (Continued): Tests for Site Effects

	I	*II*	*III*	*IV*	*V*	*VI*	*VII*
Tx(Monthly)xPost		-0.47				-0.46	-0.50
		(0.16)				(0.15)	(0.15)
Tx(BiMonthly)xPost		-0.05				-0.08	-0.07
		(0.22)				(0.21)	(0.21)
Tx(Immature)xPost			0.51			0.48	0.46
			(0.13)			(0.15)	(0.14)
TxPostxCDD				-0.073	-0.074	-0.075	-0.060
				(0.030)	(0.030)	(0.030)	(0.031)
TxPostxHDD				-0.005	-0.006	-0.009	
				(0.008)	(0.008)	(0.008)	
TxPostxFirstComp					1.09	1.09	1.10
					(0.09)	(0.09)	(0.09)
TxPostxCMeanAge					-0.028	-0.028	-0.037
					(0.016)	(0.016)	(0.015)
TxPostxCMedianIncome					-0.0031	-0.0031	
					(0.0038)	(0.0038)	
TxPostxCPctWhite					-0.030	-0.041	
					(0.469)	(0.469)	
TxPostxElecHeat					-0.68	-0.69	-0.63
					(0.36)	(0.36)	(0.37)
TxPostxHouseAge					-0.0022	-0.0022	
					(0.0028)	(0.0028)	
TxPostxHouseValue					-0.0010	-0.0010	
					(0.0007)	(0.0007)	
TxPostxPool					-1.23	-1.22	-1.25
					(0.33)	(0.33)	(0.33)
TxPostxRent					0.34	0.32	
					(0.35)	(0.35)	
TxPostxSingleFam					-0.78	-0.79	-0.95
					(0.31)	(0.31)	(0.28)
TxPostxSqFt					-0.26	-0.25	-0.36
					(0.13)	(0.13)	(0.11)
N (millions)	18.40	18.40	18.40	18.40	18.40	18.40	18.40
R^2	0.005	0.005	0.016	0.094	0.099	0.101	0.080
F Stat (Regression)	1221	1140	4227	13189	4339	4081	3693
F Stat (Experiment Dummies)	4.24	3.91	4.18	4.64	7.86	8.12	8.76
Site Effects F-test p-Value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
SD (Experiment Dummies)	0.608	0.625	0.608	0.735	0.872	0.953	0.974

All regressions include lower-order interactions of Post with all included X variables and with Experiment dummies. Household fixed effects. Standard errors are robust, clustered by household.

Table 6: Partner Characteristics

	All	*Partners*	*Non-Partners*	*Difference*
Mean: 1(Coop)	0.48	0.07	0.50	-0.43
SD (and SE): 1(Coop)	(0.50)	(0.25)	(0.50)	(0.04)***
1(Investor-Owned)	0.18	0.62	0.16	0.46
	(0.39)	(0.49)	(0.37)	(0.07)***
1(Municipal)	0.25	0.27	0.25	0.02
	(0.43)	(0.45)	(0.43)	(0.07)
1(Other Government)	0.04	0.02	0.04	-0.02
	(0.19)	(0.15)	(0.20)	(0.02)
1(State Has EERS)	0.57	0.93	0.55	0.39
	(0.50)	(0.25)	(0.50)	(0.04)***
1(State Has RPS)	0.52	0.84	0.50	0.34
	(0.50)	(0.37)	(0.50)	(0.06)***
Conservation (/cust.-yr)	25.30	48.33	24.14	24.19
	(178.14)	(69.87)	(181.84)	(11.97)**
EE Spending (/cust.-yr)	15.38	25.10	14.89	10.20
	(138.16)	(29.68)	(141.43)	(6.45)
Mean Usage (MWh/year)	12.41	9.43	12.56	-3.13
	(3.41)	(2.39)	(3.39)	(0.37)***
Pct Green Pricing	0.72	1.58	0.67	0.90
	(4.98)	(3.89)	(5.02)	(0.60)
Price (cents/kWh)	10.59	12.21	10.51	1.70
	(3.20)	(4.08)	(3.13)	(0.61)***
State Med. Income	49.31	55.87	48.98	6.89
	(6.84)	(5.51)	(6.73)	(0.84)***
State Pct College	25.99	29.49	25.82	3.68
	(4.43)	(3.82)	(4.38)	(0.58)***
State Pct Democrat	49.09	55.46	48.77	6.69
	(9.36)	(8.61)	(9.28)	(1.31)***
log(Res. Customers)	3.66	5.69	3.56	2.13
	(1.26)	(1.63)	(1.15)	(0.24)***
N	939	45	894	
F Test p-Value				0.000 ***

Table 7: Selection and Associations with ATE

Column:	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*
Specification:	*Corr w/ ATE*	*Probit*	*Selection*	*Probit 1*	*Probit 2*	*Probit 3*	*Leave Out*
Dependent Variable:	$\widehat{\tau}$	$1(T_r = 1)$	$\widehat{\tau}$	$1(T_r = 1)$	$1(T_r = 1)$	$1(T_r = 1)$	$\widehat{\tau}$
1(Coop)	-1.06	-1.11	13.3	0.65			1.68
	(0.33)***	(0.24)***	(16.5)	(0.59)			(0.49)***
1(Investor-Owned)	0.72	1.04	5.1	0.63		0.41	1.65
	(0.19)***	(0.19)***	(3.0)*	(0.49)		(0.20)**	(0.49)***
1(Municipal)	-0.21	0.05	-40.8	1.20	0.68	0.92	1.70
	(0.24)	(0.20)	(659.4)	(0.56)**	(0.20)***	(0.23)***	(0.52)***
1(Other Government)	-0.04	-0.27	1.9	0.12			1.62
	(0.19)	(0.38)	(16.0)	(0.65)			(0.47)***
Mean Usage (MWh/year)	-0.12	-0.13	5.0	-0.05	-0.07		1.61
	(0.13)	(0.03)***	(6.9)	(0.03)*	(0.03)**		(0.47)***
EE Spending (/cust.-yr)	0.010	0.000	479	0.000			1.65
	(0.007)	(0.000)	(2356)	(0.000)			(0.48)***
Conservation (/cust.-yr)	0.003	0.000	105	0.000			1.65
	(0.002)	(0.000)	(271)	(0.000)			(0.48)***
Pct Green Pricing	-0.016	0.011	-13.5	0.012			1.62
	(0.037)	(0.008)	(50.1)	(0.010)			(0.47)***
Price (cents/kWh)	0.035	0.055	6.2	-0.008			1.63
	(0.038)	(0.021)***	(11.1)	(0.024)			(0.46)***
log(Res. Customers)	0.25	0.462	1.9	0.519	0.538	0.49	3.60
	(0.07)***	(0.070)***	(0.6)***	(0.077)***	(0.082)***	(0.09)***	(1.77)**
1(State Has EERS)		1.03		0.42			1.67
		(0.25)***		(0.28)			(0.48)***
1(State Has RPS)		0.74		0.26			1.66
		(0.23)***		(0.26)			(0.48)***
State Med. Income	-0.065	0.066	-5.3	0.060	0.055		1.61
	(0.024)***	(0.012)***	(3.4)	(0.029)**	(0.016)***		(0.45)***
State Pct College	-0.009	0.081	0.1	-0.044			1.68
	(0.046)	(0.018)***	(2.8)	(0.038)			(0.47)***
State Pct Democrat	0.018	0.030	2.6	0.013			1.72
	(0.005)***	(0.011)***	(1.9)	(0.009)			(0.50)***
N	14		14	938	938	939	
F Test p-Value				0.000	0.000	0.000	
Psuedo R^2				0.397	0.374	0.289	
Coeff: $\widehat{\tau}$ on $\widehat{Pr}(T = 1)$				1.65	1.86	1.87	
SE (Murphy-Topel)				(0.48)***	(0.52)***	(0.58)***	
SE (Robust)				(0.37)***	(0.39)***	(0.43)***	

Robust standard errors in parenthesis, clustered by utility. *, **, ***: Different from zero with 90%, 95%, and 99% confidence, respectively. Columns III, VIII: Standard errors also account for uncertainty in estimated selection probability (Murphy and Topel 1985).

Table 8: MFI Summary Statistics≤

Variable	Mean	Std. Dev.	Min.	Max.	N
1(Non-Profit)	0.63	0.48	0	1	1804
MFI Age (Years)	13.99	10.43	0	115	1824
Borrowers (10^6)	0.06	0.4	0	7.54	1597
Pct Women Borrowers	0.62	0.27	0	2.12	1516
Av Loan Balance ($000's)	1.42	3.07	0	64.09	1593
Cost per Borrower ($000's)	0.18	0.19	0	1	1352
Borrowers/Staff Ratio (10^3)	0.13	0.21	0	5.07	1589
Pct Portfolio at Risk	0.08	0.12	0	1	1551
1(JPAL, IPA, or FAI Partner)	0.02	0.13	0	1	1903

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 9: Microfinance Partner Selection

	Individual Correlation	Probit 1	Probit 2	Probit 3	Probit 4
	(1)	(2)	(3)	(4)	(5)
1(Non-Profit)	-.022	-.45	-.54	-.49	-.60
	(0.077)***	(0.16)***	(0.17)***	(0.16)***	(0.18)***
MFI Age (Years)	0.0014	0.02	0.02	0.02	0.03
	(0.0004)***	(0.007)***	(0.007)***	(0.007)***	(0.006)***
Borrowers (10^6)	0.11	0.39	0.37	0.37	0.4
	(0.03)***	(0.1)***	(0.1)***	(0.1)***	(0.11)***
Pct Women Borrowers	0.021		0.54		0.32
	(0.014)		(0.28)*		(0.34)
Av Loan Balance ($000's)	-.0019			-.20	-.12
	(0.00058)***			(0.07)***	(0.09)
Cost per Borrower ($000's)	-.050				-.33
	(0.012)***				(0.54)
Borrowers/Staff Ratio (10^3)	0.044				0.13
	(0.028)				(0.26)
Pct Portfolio at Risk	-.023				-.57
	(0.017)				(0.66)
Const.		-2.15	-2.45	-1.97	-2.23
		(0.13)***	(0.23)***	(0.14)***	(0.3)***
Obs.		1557	1477	1553	1269
Psuedo R2		0.15	0.16	0.17	0.2
Prob>Chi2		0.000019	0.0000107	1.99e-06	1.89e-09

Dependent variable: 1(JPAL, IPA, or FAI partner). *, **, ***: Different from zero with 90%, 95%, and 99% confidence, respectively. Standard errors in parentheses.

Table 10: F-Test Results

Experiment		**Zip**		**Tract**	
Number	*N*	*p*	*DOF*	*p*	*DOF*
1	36,565	0.56	39	0.79	112
2	54,427	0.38	40	0.14	179
3	78,124	0.34	30	0.04 **	80
4	11,591	0.75	4	0.93	31
5	27,115	0.10	4	0.02 **	35
6	33,486	0.08 *	9	0.01 ***	35
7	17,677	0.19	44	1.00	32
8	49,510	0.55	25	0.73	87
9	78,841	0.99	39	0.41	167
10	25,145	0.49	4	0.23	32
11	17,665	0.41	3	0.17	30
12	39,178	0.48	20	0.94	112
13	42,129	0.07 *	28	0.00 ***	82
14	17,099	0.04 **	27	0.10	70
Mean	37,754	0.39	23	0.39	77

N is the number of households at the Site. p and DOF are the p-value and degrees of freedom of the F test. *, **, ***: λ different from zero with 90%, 95%, and 99% confidence, respectively.

Figures

Figure 1: Home Energy Reports: Social Comparison Module

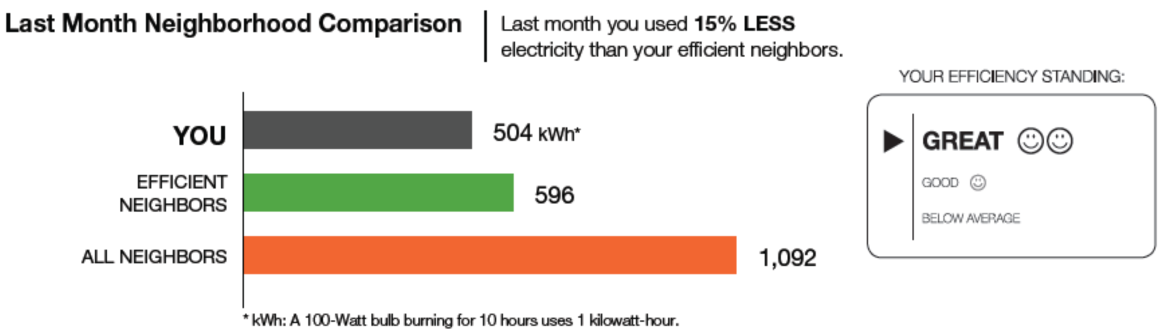
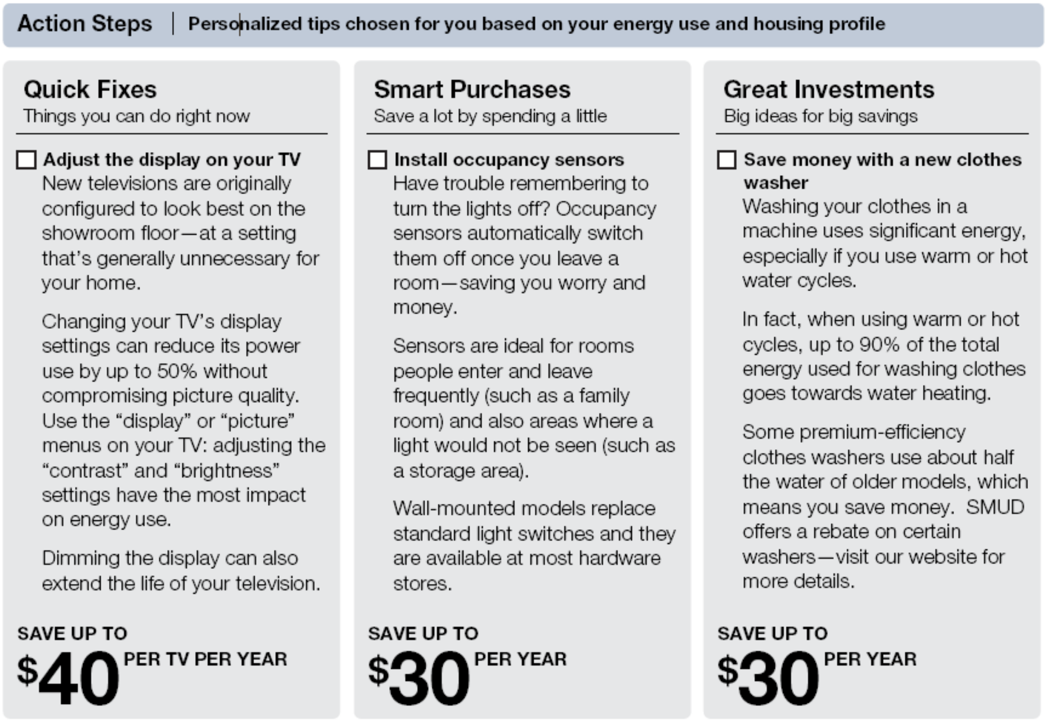


Figure 2: Home Energy Reports: Action Steps Module



Action Steps | **Personalized tips chosen for you based on your energy use and housing profile**

Quick Fixes
Things you can do right now

□ **Adjust the display on your TV**
New televisions are originally configured to look best on the showroom floor—at a setting that's generally unnecessary for your home.

Changing your TV's display settings can reduce its power use by up to 50% without compromising picture quality. Use the "display" or "picture" menus on your TV: adjusting the "contrast" and "brightness" settings have the most impact on energy use.

Dimming the display can also extend the life of your television.

SAVE UP TO
$40 PER TV PER YEAR

Smart Purchases
Save a lot by spending a little

□ **Install occupancy sensors**
Have trouble remembering to turn the lights off? Occupancy sensors automatically switch them off once you leave a room—saving you worry and money.

Sensors are ideal for rooms people enter and leave frequently (such as a family room) and also areas where a light would not be seen (such as a storage area).

Wall-mounted models replace standard light switches and they are available at most hardware stores.

SAVE UP TO
$30 PER YEAR

Great Investments
Big ideas for big savings

□ **Save money with a new clothes washer**
Washing your clothes in a machine uses significant energy, especially if you use warm or hot water cycles.

In fact, when using warm or hot cycles, up to 90% of the total energy used for washing clothes goes towards water heating.

Some premium-efficiency clothes washers use about half the water of older models, which means you save money. SMUD offers a rebate on certain washers—visit our website for more details.

SAVE UP TO
$30 PER YEAR

Figure 3: Map of OPOWER Locations

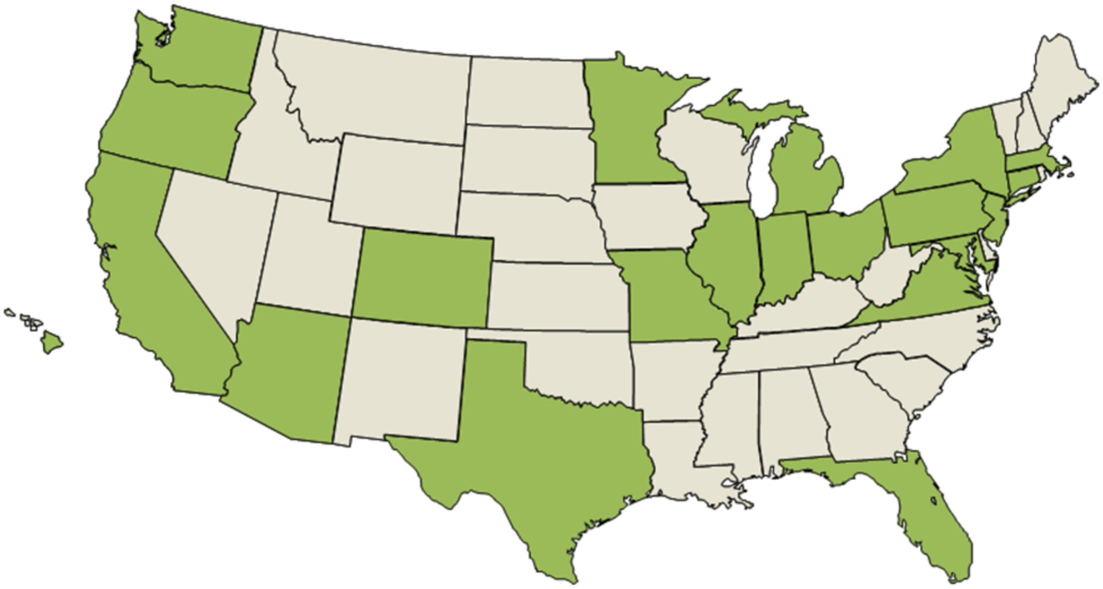


Figure 4: Predicted National Annual Savings

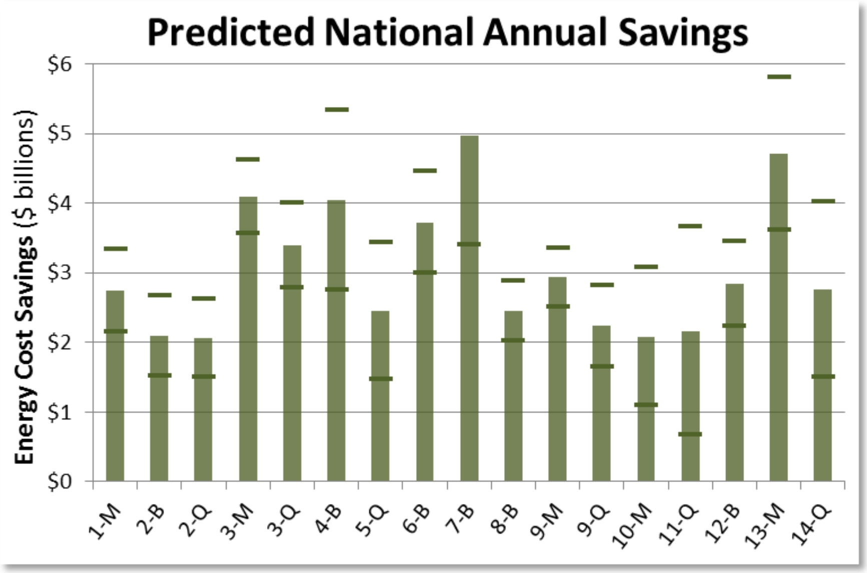


Figure 5: Selection Probability and ATE

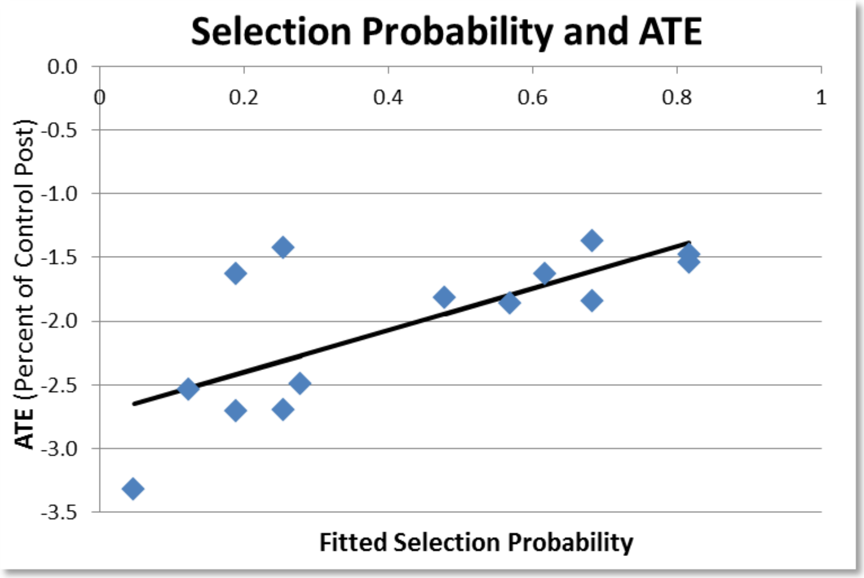
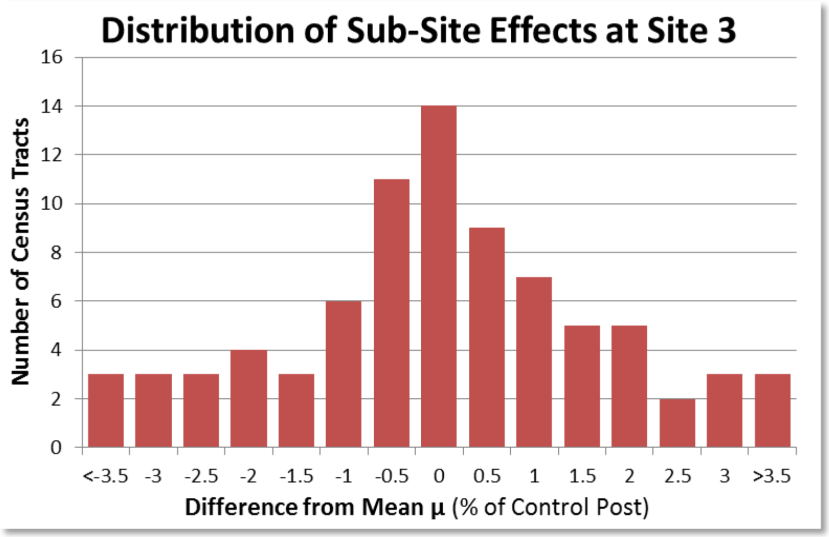


Figure 6: Distribution of Sub-Site Effects at Site 3


Distribution of Sub-Site Effects at Site 3
<-3.5
>3.5
Number of Census Tracts
Difference from Mean μ (% of Control Post)