External Validity and Partner Selection Bias

Hunt Allcott and Sendhil Mullainathan*

January 31, 2011

Abstract

External validity is an integral part of empirical work and has become a central theme in recent debates over the usefulness of randomized field experiments and the merits of structural vs. reduced form approaches. While there is substantial evidence of the value of experimental relative to non-experimental estimates (e.g. Lalonde 1986), there is little quantitative evidence on the external validity of experimental estimates because experiments are rarely implemented in multiple settings. This paper exploits a remarkable series of 14 nearly-identical energy conservation field experiments run by a company called OPOWER, involving 550,000 households across the U.S. Despite the availability of potentially-promising individual-level covariates, we show that differences in observable characteristics explain very little of the variation in treatment effects across populations. 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 present suggestive but quantitative evidence that partner selection bias may exist in other domains.

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

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

We thank Josh Angrist, Guido Imbens, and Todd Rogers for their insights. Thanks also to Tyler Curtis, Alex Laskey, and many 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 of this research.

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

1 Introduction

Since Lalonde (1986) and related work, economists have focused efforts on experimental and quasi-experimental research designs to estimate treatment effects that are "internally valid" to the experimental setting. Treatment effects can vary within and across settings, however, and the experimental setting is typically not the only setting of interest. The ability to infer treatment effects out sample has become a central question in recent debates over the merits of structural vs. reduced form empirical work (Angrist and Pischke 2010, Nevo and Whinston 2010) and of randomized controlled trials (Deaton 2010, Heckman 2010, Imbens 2010, Rodrik 2009). Although "external validity" is central to this discussion of when and how to use field experiments, there is little quantitative evidence of its importance in any domain.

We focus on one particular aspect of external validity: how well treatment effects can be generalized across heterogeneous populations and economic environments. Under a strong definition of external validity, treatment effects could be extrapolated from one setting to another if observable differences are taken into account. Just as individual-level outcomes vary on unobservables, however, treatment effects may vary on unobservables. Under a weaker definition of external validity, unobservables may vary across settings, but one could infer the distribution of possible treatment effects by running many experiments in a random sample of potential settings. Just as individual units select into treatment, however, experimental sites may themselves be selected.

As in Lalonde (1986), testing internal validity requires one setting where a consistent estimator is available, and involves comparing a set of proposed estimators to that consistent estimator. Testing external validity in our two senses requires a set of internally-valid estimators of the same parameter in different settings. To test the stronger version of external validity, one extrapolates between settings and tests for whether unobservables generate excess variance between predicted and actual treatment effects. Explicitly testing the weaker version is not possible because we do not observe the distribution of treatment effects across non-experimental settings. However, just as individual-level selection into treatment on observables suggests selection on unobservables (Altonji, Elder, and Taber 2005), one can test whether experimental and non-experimental settings vary on observable factors that correlate with the treatment effect.

We take advantage of a remarkable opportunity to quantify the importance of external validity in one domain: a series of 14 nearly-identical randomized field 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 compare their energy use to that of their neighbors and provide energy conservation tips. Although we use these programs to make a broader scientific point, the generalizability of this program's effects to potential future sites is of great specific interest to environmental economists and policymakers. This is due to a proliferation of new regulations mandating energy conservation, spurred partially by concern over climate change and high energy prices, which will force additional utilities across the country to decide whether to adopt the program.

We begin by introducing a formal model of external validity. The model begins with the standard individual-level framework of potential outcomes and treatment effects, as in Rubin (1974), and captures selection into treatment using a simple generalized Roy framework that draws on Heckman and Vytlacil's (2005, 2007) Marginal Treatment Effect approach. We then aggregate to the site level, where potential partner sites have different potential outcomes depending on whether they adopt and evaluate a program. The site-level decision of whether to adopt the program closely resembles the individual-level decision of whether to select into treatment.

The model formalizes our two notions of external validity. Under the stronger notion, results from one Sample site generalize to any other Target site after controlling for differences in the distributions of observables. This requires an assumption called "unconfounded location" (Hotz, Imbens, and Mortimer 2005). Under the weaker notion of external validity, there is no bias on average when extrapolating from many Sample sites that have adopted and evaluated the treatment to many Target sites that have not. This weaker notion requires a weaker assumption, which we call "expectation unconfounded location," which allows unobservables to vary across sites but requires that selection be independent of these unobservables.

We then turn to the OPOWER programs as one empirical test of these assumptions. Average Treatment Effects vary by 240 percent across the 14 existing experiments, an amount which is both statistically and economically significant. Compared to other settings of interest to economists, however, one might expect relative success at using observable population characteristics to explain this variation. This is because the treatment is almost exactly the same across experiments and because we observe a large set of covariates which theory predicts could moderate treatment effects.

We use "Sample" data from each of the 14 experiments and adjust for observable differences in population characteristics in an effort to predict average treatment effects in each of the 13 other "Target" populations. Our first adjustment procedure is to estimate the treatment effect in the Sample as a linear function of observed covariates, then predict the treatment effect given the different covariate means in the Target. Our second procedure uses Inverse Probability Weights to reweight the Sample population to resemble the Target population on observables. Using these two procedures, we compare the ATEs predicted using Sample data to the experimentally-estimated ATEs in the Target populations. The unconfounded location assumption is rejected: the predicted and true Target ATEs differ with 90 percent confidence about 50 percent of the time.

We then test the assumption of "expected unconfounded location." Including the 14 initial experiments and future contracted programs, the company has 47 partner electric utilities in 21 states. These partners were drawn from the larger population of nearly 1000 utilities in the U.S. We show that there is selection on observables: OPOWER's partners tend to have different ownership structure, are larger, and tend to be in wealthier states with stronger environmental regulation. Furthermore, within the 14 experiments where results are available, there is statistical evidence of "partner selection bias": selection probabilities conditional on observables are systematically correlated with treatment 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.

As with Lalonde's (1986) analysis of experimental vs. non-experimental estimators, the OPOWER field experiments are only one example in one setting. The precise magnitudes of the difficulties in generalizing OPOWER's results surely cannot be generalized to other settings. 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 OPOWER.

Our analysis cannot be used to argue that randomized field experiments are not useful and important in this context. We show that the 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. Perhaps more interestingly, 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. Thus, while one would not argue against internal validity, our analysis does suggest that external validity merits attention and rigorous treatment. In the conclusion, we discuss several forms that this might take.

1.1 Related Literature

The literature that most directly motivates our analysis are recent debates of the merits of "structural" vs. "reduced form" approaches and the role of randomized evaluations in generating useful new knowledge. One factor precipitating this debate is the increasing popularity of randomized field experiments in economics, perhaps most prominently in development economics, but also in behavioral economics, labor, education, environmental economics, and other fields.

While the debates have many facets, external validity has been a central issue. Imbens (2010) writes, "Much of the debate between structural and causal approaches ultimately centers on the weight researchers put on external validity versus internal validity of a study." Rodrik's (2009) main criticism of development microeconomics is that its practitioners "have to recognize that the utility of randomized evaluations is restricted by the narrow and limited scope of their application." Similarly, a fundamental part of Deaton's (2010) argument is that randomized trials are "too narrow and too local to tell us "what works" in development."

Duflo (2004), Duflo, Glennerster, and Kremer (2007), and others discuss several 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. Features of the "economic environment" may vary across settings, such as the unemployment rate in job training programs or the weather in energy conservation programs. 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.

The OPOWER setting isolates the impact of two of these factors: differences in population and differences in economic environment. This is because OPOWER's programs are implemented as natural field experiments, meaning that most subjects do not know they are being studied, entire populations are treated, eliminating selection into the experiment, and the treatment is implemented at large scale and is almost completely identical across locations, reducing concerns about treatment fidelity.

The opportunity afforded by OPOWER to study the generalizability of experimental results across sites is unusual, but not unique: our analysis connects to a small literature that compares impacts of programs implemented at multiple sites. In development economics, there are several published multi-site randomized evaluations. Banerjee, Cole, Duflo, and Linden (2007) evaluate a remedial education program carried out in two Indian cities, Vadodara and Mumbai. The effects are similar for math across the two cities but stronger for language in Vadodara, which the authors attribute to the fact that more students in Mumbai already had basic reading skills. Chattopadhyay and Duflo (2004) examine an Indian policy, which was implemented as a randomized trial, that reserved additional spaces on village councils for women. They compare effects from two districts in different parts of the country, showing that women invest more in public goods more closely linked to women's interests in both locations, although those interests vary between the two locations. Miguel and Kremer (2004) and Bobonis, Miguel, and Sharma (2006) carry out similar health interventions in Kenya and India and find qualitatively similar results.

As one weakens the definition of what a "similar" treatment is, there are increasingly large literatures that compare "similar" treatments in different settings. For example, Aigner (1984) compares demand parameters from a number of related electricity pricing randomized experiments. Card, Kluve, and Weber (2009) carry out a meta-evaluation of 97 studies of labor market policies. Holla and Kremer (2009) survey evidence from a number randomized trials on the effects of price on access to education and health care in developing countries. Meyer (1995) draws lessons from unemployment insurance experiments carried out in different parts of the U.S. There are many other examples, including studies of class size reductions, returns to education, and the effects of military service (Angrist and Pischke 2010).

Quasi-experimental parameter estimates can also be compared across locations or across groups to whom different instruments are "local." For example, Angrist, Lavy, and Schlosser (2010) compare the effects of multiple different instruments that act on different populations to estimate the causal link between the quantity and quality of children.

The papers most closely related to ours analyze job training programs implemented across multiple sites in the U.S. during the 1980s. The Job Training Partnership Act (JTPA) of 1982 implemented job training programs at 600 sites, of which 16 were implemented as randomized trials.[1] Although they were claimed to "represent the diversity of the program," the experimental sites were those that agreed to participate out of more than 200 that were originally approached (Hotz 1992). 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.

Hotz, Imbens, and Mortimer (2005) analyze the Work INcentive (WIN) job training program implemented at four separate locations in the 1980s. They are relatively successful at using observable characteristics to extrapolate treatment and control outcomes for most subgroups that they examine. 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. Both papers use observables to control for differences in economic factors and population characteristics between sites in order to determine whether the Riverside County site, which emphasized a particular method of job training called "Labor Force Attachment," was more effective.

Within this job training literature, there is thus an apparent disagreement over the generalizability of results of randomized evaluations. The JTPA analyses would lead one to believe that even 16 randomized trials are insufficient to know a program's expected impact. The WIN and GAIN analyses, on the other hand, lead one to believe that unexplained differences between any pair of sites might be attributed to a specific element of program design.

The three WIN and GAIN studies, as well as others that formally or informally use observables to generalize or extrapolate treatment effects, involve a subtle conceptual contradiction. The studies evaluate randomized trials that were run because unobservable factors confound non-experimental comparisons of treatment and control group outcomes within a population. However, the studies assume that unobservable factors do not confound comparisons of treatment effects across populations. While these three studies are explicit about these assumptions and present suggestive tests that they are valid, these studies are unusual in the rigor they apply to extrapolation.

[1] The official Department of Labor evaluation (Bloom *et al.* 1993) illustrates a common tension that empirical researchers face between being cautious about external validity and claiming that the study has produced generalizable knowledge. The authors write (page 73) that the sites and participants "resemble [sites] and their participants nationally and also include much of their diversity." In an appendix, however, they caution that "we make no claims about the national representativeness of the impact estimates" (page 295). Despite this word of caution, the impact evaluation's results were an important justification for Congress in cutting national JTPA program funding by 80 percent (Heckman and Smith 1997).

Deaton (2010), citing Cartwright (2007a), highlights this issue. He draws a contrast "between the rigor applied to establish internal validity - to establish the gold standard status of RCTs - and the much looser arguments that are used to defend the transplantation of the experimental results to policy. For example, running RCTs to find out whether a project works is often defended on the grounds that the experimental project is like the policy that it might support. But the "like" is typically argued by an appeal to similar circumstances, or a similar environment, arguments that depend entirely on observable variables. Yet controlling for observables is the key to the matching estimators that are one of the main competitors for RCTs and that are typically rejected by the advocates of RCTs on the grounds that RCTs control not only for the things that we observe but things that we cannot."

The next section uses a formal model to bring this issue to life, showing that unobservables and selection problems that have received significant attention in discussions of internal validity have direct analogues at the population level that affect external validity.

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 similarly to Heckman and Vytlacil (2005, 2007). 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:

$$Y_{0i} = \beta X_i + \zeta Z_i \tag{1a}$$
$$Y_{1i} = (\alpha + \beta) X_i + (\gamma + \zeta) Z_i \tag{1b}$$

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:

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

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

We now present an analogous model of potential partners at different sites that decide whether to take up a program. In the OPOWER context, this is intended to describe a set of sites with decisionmakers who choose whether to enroll their populations in the program. In other contexts, this could reflect an NGO that decides whether to evaluate an existing program with a randomized trial. 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 partner adopts 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:

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

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.

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 set of treated and untreated sites. If equation (6a) determines selection of sites into treatment, expected location unconfoundedness holds under one of two conditions. First, it holds if τ is constant across sites. In most situations, including with OPOWER, this not the case. Second, it holds if selection is driven only by C and C is independent of unobservables that influence the treatment effect: $\omega = 0$ and $Z \perp C$. When the assumption is violated, we call this "partner selection bias." We test for partner selection bias in the OPOWER context in Section 5.

The covariance between C and τ determines whether partner selection is positive or negative. Potential partners that are capable of running effective programs and have receptive populations might also be most capable of and interested in randomized trials. This "cherry-picking" of experimental partners would generate positive selection. Alternatively, potential partners that might be most capable and interested could also be running other substitute programs or could have already treated the parts of their population with the highest τ_i. This form of "diminishing returns" would generate negative selection.

Consider the class of situations where expected location unconfoundedness holds but location unconfoundedness does not. There are two interpretations. We could say that we have a biased estimator when we extrapolate from one Sample to one Target. Alternatively, we could imagine taking the limit in the sense of extrapolating between many pairs of Sample and Target sites. The average estimand over many extrapolations would be unbiased, but the usual standard errors would not reflect the true variance in the extrapolation errors because they do not reflect variance in $f(Z)$ across sites. We explore these issues empirically in Section 4.

3 Experiment Overview

Our empirical example is 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 conserve energy. A substantial body of work studies or comments on these experiments, including Allcott and Mullainathan (2010), Ayres, Raseman, and Shih (2009), Costa and Kahn (2010), Davis and Wagner (2011), and Violette, Provencher, and Klos (2009). These programs have also been extensively covered in the popular press and are

at the forefront of a shift in the energy industry toward "behavior-based" (instead of "technology-based") energy conservation programs that are evaluated using randomized controlled trials. Allcott (2010) provides a comprehensive program evaluation, including a more detailed background, interpretation of the effects, and a 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. This presentation was directly inspired by academic work by Nolan *et al.* (2008) and Schultz *et al.* (2007) showing that providing information on social norms caused people to reduce energy use. 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 presentational differences over time, within an experiment, and across experiments. Because these differences are so small, it is likely that the bulk of the unexplained variation in treatment effects across experiments 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.

When interpreting differences in ATEs across experiments, it is worth keeping in mind that this treatment is quite consistent compared to other experiments of interest in economics. For example, job training programs take very different forms at different sites (Dehejia 2003), Hotz, Imbens, and Klerman 2006). Similarly, variation in teacher quality complicates the ability to generalize the results of programs to reduce class size or provide additional teachers.

As of the end of 2010, OPOWER has contracts to work with 47 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.

As of October 2009, experiments had begun at 10 of these sites, 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 sites, 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 separate "experiments" at these ten "sites."

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

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 control group average post-treatment consumption.

As documented by Allcott (2010), the estimated ATEs are not very sensitive to different specifications of control variables and fixed effects. The estimated ATEs are shown in Table 2. Notice that in experiments 2, 3, and 9, the population or subpopulation was randomly assigned between monthly, bimonthly, and/or quarterly frequencies, while all other experiments involved only one frequency.[2] The unweighted mean ATE across experiments is a 2.06 percent reduction in electricity use. The ATEs vary by a factor of 2.4, from 1.37 percent to 3.32 percent.

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. Using these pricing data and each experiment's percent average treatment effect and average electricity consumption per year, we calculate cost effectiveness using the same formula as Allcott (2010). The unweighted mean cost effectiveness is 3.30 cents/kilowatt-hour. As shown in Table 2, cost effectiveness varies across experiments by a factor of 4.2, from 1.27 to 5.36 cents/kilowatt-hour.

A key motivating stylized fact is that the variation in ATEs and cost effectiveness across experiments is both statistically and economically significant. ATEs are statistically different with 90 percent confidence in over half of the experiment pairs. While some variation in treatment effects is associated with the frequency of receiving Reports, Table 2 shows that there is still substantial variation within frequency across experiments.

Perhaps the most fundamental measure of the economic significance of variation in treatment effects is whether it influences whether the treatment will be adopted or not. Each electric utility has

[2] 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.

an energy conservation target, a budget constraint or profit maximization objective, and a portfolio of different energy conservation programs that it can fund. These programs been estimated to cost between 5.5 and 6.4 cents per kilowatt-hour (Arimura, Newell, and Palmer 2009) or between 1.6 and 3.3 cents per kilowatt-hour (Friedrich *et al.* 2009).[3] Whether an HER program has cost effectiveness at the lower or upper end of the range for existing experiments therefore has has important implications for whether a utility would partner with OPOWER and for the size of the project they would implement.

In this case, external validity is more than an academic question: electric utilities around the country are grappling with this very issue. If a utility partners with OPOWER and the savings are less than expected, the utility risks falling out of compliance with regulatory requirements. If realized savings are more than expected, the utility could have reduced expenditures on other energy conservation programs.

4 Extrapolation on Observables

In this section, we test the ability to extrapolate from each of the 14 OPOWER experiments to each of the other 13 experiments, controlling for observable characteristics that vary across populations and are correlated with the treatment effect. We describe the individual-level characteristics available, estimate their correlations with the treatment effect, describe the two methodological approaches to extrapolation that we employ, and present results. We will conclude that individual-level observables explain very little of the variation in treatment effects across experiments. Instead, this variation is driven by unobserved factors.

4.1 Data

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 are more likely to be associated with the treatment effect.

Table 3 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.

[3] 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.

Zero corresponds to the mean of the neighbor distribution, and households with higher values used less energy. As detailed in Allcott (2010), 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 stronger treatment effects. The next three variables in Table 3 are tract-level average characteristics which we hypothesized could be associated with these sorts of "cultural" differences that moderate the treatment effect.

Table 4 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.

4.2 Empirical Approach

The goal of this section is to take data from a Sample population s and predict a treatment effect τ_g in Target population g, controlling for differences in distributions of observable characteristics between the two populations. We use two standard approaches, parametric extrapolation and reweighting using Inverse Probability Weights.

4.2.1 Parametric Extrapolation

The parametric extrapolation approach is straightforward: we estimate heterogeneous treatment effects in the Sample as parametric functions of observable characteristics, then fit predicted treatment effects onto the Target population given its different characteristics. The estimating equation is:

$$Y_{it} = (\alpha X_i) \cdot T_i P_{it} + (\beta X_i) \cdot P_{it} + \mu_{my} + \upsilon_i + \varepsilon_{it} \tag{11}$$

In this equation, α and β are row vectors and X_i is a column vector that includes household characteristics and a constant. Standard errors are robust and clustered by household i. This gives an estimated parameter matrix $\widehat{\alpha}$, with variance-covariance matrix $\widehat{V}_\alpha$. Assuming linearity and unconfounded location, the treatment effect in the target population g is predicted by adjusting

the estimated treatment effect in Sample $\widehat{\tau}_s$ for the difference in mean characteristics between Target and Sample:

$$\widehat{\tau}_g = \widehat{\tau}_s + \widehat{\alpha}\left(\overline{X}_g - \overline{X}_s\right) \tag{12}$$

In this equation, $\overline{X}_g$ and $\overline{X}_s$ are column vectors of mean characteristics in the Target and Sample populations.

The Sample population ATE $\widehat{\tau}_s$ and standard error $\widehat{\sigma}_{\tau_s}$ are estimated using equation (10) above. The standard error of the Target ATE can be computed using the Delta Method:

$$\widehat{\sigma}^2_{\tau_g} = \widehat{\sigma}^2_{\tau_s} + \left(\overline{X}_g - \overline{X}_s\right)' \cdot \widehat{V}_a \cdot \left(\overline{X}_g - \overline{X}_s\right) \tag{13}$$

4.2.2 Inverse Probability Weighting

The Inverse Probability Weighting approach is fundamentally due to Horvitz and Thompson (1952). Intuitively, we household weights w_i that, when applied to the Sample Population, make the Sample resemble the Target on observable characteristics. We then estimate the treatment effect using Sample data in Weighted Least Squares using weights w_i. The appeal of the IPW approach relative to parametric extrapolation is that it does not require that the treatment effect be a linear function of observables.

The Inverse Probability Weights w_i are calculated by first pooling all Sample and Target households into one dataset and estimating the probability that a household is in the Target conditional on observables: $\Pr(G_i = 1|X_i)$. In a slight abuse of language, we call this a "propensity score," although it is a probability of being in Target instead of Sample, not a probability of being in Treatment instead of Control. We estimate the propensity score using three approaches: a linear probit model and two Series Logit Estimators, as suggested by the work of Hirano, Imbens, and Ridder (2003).[4] In the first Series Logit specification, we index characteristics by j or k and include each characteristic X_{ij} and interactions with each of the other characteristics, $X_{ij} \cdot X_{ik}$. In the second specification, we include the same characteristics and interactions plus squares of each characteristic.

Denoting the normal and logistic cumulative density functions by Φ and Λ, respectively, our three estimators are:

[4]The Series Logit Estimators are inspired of Hirano, Imbens, and Ridder's (2003) discussion of estimating Population Average Treatment Effects under unconfoundedness when part of the population is treated and propensity scores are strictly between zero and 1. They derive an estimator of the treatment effect and show that it achieves the semiparametric efficiency bound when the propensity score is estimated using a Series Logit Estimator of appropriate order. Our setting is different, as we have a Sample population that is treated with known probability and a Target where none have been treated and the outcome variable may not be observed. We are not aware of analogous results about the efficiency of the Series Logit Estimator in our setting.

$$\text{Probit:} \qquad \Pr(G_i = 1|X_i) = \Phi\left[\sum_{jj} \theta X_{ij}\right] \tag{14a}$$

$$\text{Series Logit 1:} \qquad \Pr(G_i = 1|X_i) = \Lambda\left[\sum_{j}\left(\theta_j X_{ij} + \sum_{k\setminus j} \theta_{jk} X_{ij} X_{ik}\right)\right] \tag{14b}$$

$$\text{Series Logit 2:} \qquad \Pr(G_i = 1|X_i) = \Lambda\left[\sum_{j}\left(\theta_j X_{ij} + \theta_{jj} X_{ij}^2 + \sum_{k\setminus j} \theta_{jk} X_{ij} X_{ik}\right)\right] \tag{14c}$$

As discussed by Crump *et al.* (2009) and others, estimators of Average Treatment Effects perform poorly in finite sample when the propensity scores are close to zero or one. This will also be true when we extrapolate from Sample to Target. We address this using two trimming approaches. The first approach is to eliminate observations i such that the propensity score distributions in Sample and Target have common support. The second approach is to keep only observations with $0.1 < \widehat{\Pr}(G_i = 1|X_i) < 0.9$.[5]

The fitted Inverse Probability Weights $w_i|X_i$ are:

$$w_i|X_i = \frac{\widehat{\Pr}(G_i = 1|X_i)}{1 - \widehat{\Pr}(G_i = 1|X_i)} \tag{15}$$

The Target ATE is then estimated with Sample data and weights $\widehat{w}_i$, using estimating equation (10) above.

4.3 Conditioning Variables

We have not yet discussed which observable characteristics X we condition on in extrapolating from Sample to Target. Asymptotically, one would use all observed covariates and possible subgroups. In finite sample, however, one would not.

For intuition, consider first the parametric extrapolation case. Asymptotically, one would condition the estimated Sample treatment effect on all observed covariates. In finite sample, however, the correlation of some variables with the treatment effect may be imprecisely estimated, and extrapolating based on an imprecisely-estimated model can actually worsen the predictions. In our application, for example, parametric extrapolation using all observed covariates produces wildly inaccurate out-of-sample predictions.

[5] This "90/10" trimming rule is suggested by the results of Crump *et al.* (2009), although we recognize that this result was derived in the context of traditional propensity score matching.

Extrapolating with Invese Probability Weights in finite sample produces a similar problem. The observable characteristics can be used to generate large number of subgroups, each with different CATEs and IPWs. In finite sample, the more subgroups, the smaller the group size, which makes the IPWs and CATEs more and more imprecisely estimated. In addition, there are some variables that are only weakly correlated with the treatment effect, meaning that their predictive value is limited, whose distributions in Target and Sample have little overlap. In terms of eliminating bias, one gains little from using these variable to calculate the IPWs, and the efficiency costs are high, as the IPWs give low weight to the large number of observations where there is low overlap, and very high weight to the small number of observations that do overlap. One can imagine a limiting case where a variable has zero overlap between Sample and Target, but one might still consider it sensible to try to extrapolate using IPWs if that variable has sufficiently small correlation with the treatment effect in the Sample.

One approach to this problem would be to develop a procedure using cross-validation to determine the optimal set of variables on which to weight and extrapolate. Instead, we use an approach which is more ad-hoc but more closely approximates what is often done in practice: we extrapolate only on variables that theory predicts could moderate the treatment effect, where there is empirical evidence across multiple experiments that these variables are associated with the treatment effect, when that variable has a statistically-significant correlation with the treatment effect in the Sample population, and if there is sufficient overlap between Sample and Target such that the estimates do not become unduly noisy.

To implement this approach, we began by limiting to the set of occupant and house characteristics that one might theoretically expect to be correlated with the treatment effect, as described in Section 4.1 above. We then estimate equation (11) above for all characteristics available in each of the experiments. We proceed only with the set of variables that are statistically significantly positively correlated with the treatment effect in three or more experiments, with a ratio of at least three-to-one of significant positive correlations to significant negative correlations, or statistically significantly negatively correlated with the treatment effect in three or more experiments, with a ratio of at least three-to-one of significant negative correlations to significant positive correlations. Our threshold for statistical significance is 90 percent confidence in a two-sided test.

Tables 5 and 6 present these estimates for experiments 1 to 14. First Comparison, the indicator for Electric Heat, House Age, and the indicator for whether the house has a pool are the four variables that meet the condition above. First Comparison has the clearest correlation with the treatment effect: households that use more energy than their neighbors conserve more after receiving the reports. Electric Heat, House Age, and Pool are associated with larger treatment effects in absolute value, consistent with a theoretical prediction that these households have lower-cost opportunities to reduce electricity use.

In each pair of experiments where we extrapolate from a Sample to a Target, the set of variables used for extrapolation is further restricted from those four. First, the variable must be observed

in both experiments. Second, the variable must be statistically significantly associated with the treatment effect in the Sample population in the same direction as the association typically observed. Third, in order to avoid excessive imbalance in the IPWs, we eliminate any variables on which there is little overlap between Sample and Target. Specifically, this is done by calculating the propensity that an observation drawn from the Sample and Target population is in the Target, based only one individual covariate. We use that covariate for weighting only if the mean propensity score in the Sample is greater than 0.43.

4.4 Results

Table 7 presents the results of our extrapolation procedures. We use a total of 16 procedures. Seven are "Linear" parametric procedures. The first does not trim the populations. In the next six, we estimate the propensity scores using the three different estimators (Probit, Series Logit 1, and Series Logit 2) and trim using the two trimming approaches (Support and 90/10). The next nine procedures use IPWs, combining each of the three propensity score estimators with no trimming and with the two trimming approaches.

We extrapolate from each of the 14 experiments to each of the other 13 experiments, for a total of 182 observations. For each extrapolation, we find the difference between $\widehat{\tau}_g|(X_s, X_g, Y_s)$ and $\widehat{\tau}_g|(X_g, Y_g)$. This difference is summarized in the "Error" columns. Because the trimming procedures do not substantially change the mean characteristics $\overline{X}_g$ and $\overline{X}_s$, the linear extrapolation results are similar without trimming and under the different trimming procedures. Trimming does reduce Root Mean-Squared Error (RMSE) slightly, however, for the linear estimators, and the best-performing estimator is the linear extrapolation with 90/10 trimming. The IPW estimators have comparably low MSE if and only if 90/10 trimming is used.

The error reflects both experiment-specific unobservables and sampling error in the Sample and Target estimators of τ_g. Under the assumption of unconfounded location, the error is due only to sampling error, and the test statistic for equality of the two estimated ATEs is:

$$\frac{\widehat{\tau}_g|(X_s, X_g, Y_s) - \widehat{\tau}_g|(X_g, Y_g)}{\sqrt{\widehat{\sigma}^2_{\tau_g} + \widehat{\sigma}^2_{\tau_s}}} \tag{16}$$

Under unconfounded location, these test statistics take the standard normal distribution, and equality of the two estimated ATEs should be rejected with 90 percent confidence ten percent of the time. As shown in Figure 4, these statistics have variance visibly wider than the standard normal. The standard deviation of these statistics is presented in the next to last column of Table 7; for the best-performing procedures it is over 2. The last column of table 7 shows that even for the best-performing prcedures, equality of the ATEs is rejected with 90 percent confidence in 47 percent of cases. Figure 5 shows the magnitude of the prediction errors by comparing them to

the distribution of Target treatment effects. This illustrates that there is economically-significant variation when generalizing across experiments which is unexplained by observables.

The last row of Table 7 shows the performance of non-experimental estimators. In these specifications, we estimate the treatment effect in using only Treatment group data from the Target population, using a difference estimator that conditions on differences in weather between the pre- and post-treatment periods. This estimator's RMSE is substantially larger than the RMSEs from the best-performing procedures that extrapolate internally-valid estimators from other Samples. This result underscores why our analysis cannot be used to argue against the importance of randomized trials in this context.

The key takeaway from this section is that 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, there are statistically and economically significant unexplained differences in treatment effects. In this application, unconfounded location does not hold.

5 Partner Selection

Unconfounded location is a strong assumption which might hold in few applications. In this section, we test the weaker assumption of expected location unconfoundedness by examining whether average treatment effects are correlated with selection probabilities. 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. We will conclude that there is a "partner selection bias" that biases the results of existing projects relative to those that could 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 properly model how state and local policies affect utilities' decisions, we actually define our observations at the level of a state-by-utility.

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," 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 8 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. There is clearly selection on observables.

5.2 Correlations with Treatment Effects

For there to be selection bias, observables associated with the selection probability must also be associated with the treatment effect. Column 1 of Table 9 shows the correlation of each utility-level characteristic with the average treatment effect in the 14 experiments where we have an estimated ATE. Recall that the treatment reduces energy demand, so "stronger" ATEs are more negative. Column 1 shows that the ATE is statistically significantly stronger for Cooperatives, less strong for IOUs and utilities with larger customer bases, stronger in states that are richer and better educated, but less strong in states with more democratic voters. The standard errors in this column, as well as the standard errors in the last row of the table to be discussed momentarily, are heteroskedasticity-robust, clustered by utility, and account for the standard error in the estimated ATE (Murphy and Topel 2002).[6]

The second through fifth columns of Table 9 present probit regressions that estimate partner

[6] More precisely, our regression is:

selection probabilities as linear functions of observables. 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 lower average energy usage per residential customer are less likely to partner with OPOWER. Larger utilities in states with EERS regulations and higher median income are more likely to partner. Across the four columns, the magnitudes and statistical significance of these correlations are highly robust to including different subsets of observable characteristics.

The bottom rows of the table are the key results. These present the coefficient of a regression of the average treatment effect on the fitted selection probability for the 14 experiments where we have ATEs. This coefficient is statistically-significant in three of the four columns, and consistently positive with magnitude of around 2. Again recalling that larger treatment effects are more negative, this means that a ten percentage point increase in selection probability is associated with about a 0.2 percentage point smaller ATE. There is robust evidence of *negative* selection: utilities whose observable characteristics make them more likely to partner with OPOWER actually have smaller treatment effects.

Figure 6 plots the graph of average treatment effects vs. fitted selection probability, using the fit from Column 2 of Table 9.

5.3 Related Example: Microfinance Experiments

We suspect that the potential bias caused by partner selection is not limited to the OPOWER example. As one further example, we consider what types of microfinancial institutions (MFIs) partner with US-based academic organizations to do randomized field experiments. Unlike with OPOWER, here we do not have average treatment effects for different MFIs, and indeed many microfinance field experiments test more nuanced hypotheses instead of straightforward impact evaluations.

Microfinance is a convenient area to examine partner selection for three reasons. First, it is a domain of interest to a number of mainstream economists. Second, there are many microfinance field experiments with many partners. Third, there is a centralized global database of microfinan-

$$\widehat{\tau}_r = \beta_1 X_r + \beta_0 + \varepsilon_r \tag{17}$$

Denote by $\widehat{\sigma}^2_\beta$ the heteroskedasticity-robust standard error, clustered by utility. X is the 14-by-2 matrix formed by combining the characteristic vector with elements X_r and a vector of ones. Denote by $\widehat{V}_\tau$ a 14-by-14 diagonal matrix with element $\{r, r\}$ being the standard error of the estimated treatment effect in experiment r. Our standard error estimator is:

$$\widetilde{\sigma}^2_\beta = \widehat{\sigma}^2_\beta + \begin{Bmatrix} 1 & 0 \end{Bmatrix} \cdot (X'X)^{-1} X' \widehat{V}_\tau X (X'X)^{-1} \cdot \begin{Bmatrix} 1 \\ 0 \end{Bmatrix} \tag{18}$$

cial institutions (MFIs) that both contains relevant partner characteristics and defines the set of potential partners.

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 10 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.

The first column of Table 11 shows the unconditional correlation of each MFI characteristic and whether an MFI is an experiment partner. Columns 2-5 of the same table present various probit regressions of the "experiment partner" dummy variable on various sets of characteristics. The correlations are robust, although the significance levels vary across columns 1-5: experimental partners are more likely to be for-profit, larger, and older. They seem to have more women borrowers and smaller average loan balances, both of which could affect repayment rates. Partner MFIs also have more borrowers per staff member and, relatedly, lower cost per borrower, although these are marginally significant. A final correlation, which is also marginally statistically significant, suggests that partner MFIs have less Portfolio at Risk, which corresponds to better 30-day repayment rates. Across set of included characteristics, a Chi-Squared test clearly shows that partner MFIs are selected on observables.

6 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 effects of similar treatments 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 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. For example, although the OPOWER experiments are not externally valid in our two senses, 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 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 was done with Mexico's Progresa conditional cash transfer program.

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" are more easily generalized across sites and domains than average treatment effects for specific projects (Deaton 2010, Ludwig, Kling, and Mullainathan 2010). 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 (2010). "Social Norms and Energy Conservation." Working Paper, Massachusetts Institute of Technology (June).

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

[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, 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.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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

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

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

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

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

[20] 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.

[21] 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).

[22] 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.

[23] Davis, Matthew, and Gernot Wagner (2011). "Heterogeneous Treatment Effects and the Behavioral "Wedge": The Carbon Impacts of Peer Energy Comparisons." Working Paper, Environmental Defense Fund (January).

[24] 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.

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

[26] 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.

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

[28] 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).

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

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

[31] 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).

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

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

[34] 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.

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

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

[37] Heckman, James, and Edward Vytlacil (2007). "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 Handbook of Econometrics, Vol. 6B.

[38] 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.

[39] 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).

[40] 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.

[41] 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).

[42] 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.

[43] 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.

[44] 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.

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

[46] 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.

[47] 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).

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

[49] 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.

[50] Levitt, Steven D. and John A. List (2009). "Field Experiments in Economics: The Past, the Present, and the Future." European Economic Review, forthcoming.

[51] Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan (2010). "Policy Field Experiments." *Journal of Economic Perspectives*, forthcoming.

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

[53] 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.

[54] Murphy, Kevin M., and Robert Topel (2002). "Estimation and Inference in Two-Step Econometric Models." *Journal of Business and Economic Statistics*, Vol. 20, No. 1 (January), pages 88-97.

[55] 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.

[56] 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.

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

[58] 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

[59] 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.

[60] 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.

[61] 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.

[62] 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.

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

[64] 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.

[65] 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.

[66] 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.

[67] 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

[68] 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

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

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

[71] 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: Experimental ATEs and Cost Effectiveness

Experiment	**ATEs (%)**			**CE (c/kWh)**		
Number	*Monthly*	*BiMonthly*	*Quarterly*	*Monthly*	*BiMonthly*	*Quarterly*
1	-1.83 (0.2)	-	-	2.03 (-0.22) -	-	
2	-	-1.4 (0.19)	-1.37 (0.19)	-	4.09 (-0.56)	3.8 (-0.53)
3	-2.87 (0.18)	-2.48 (0.3)	-2.23 (0.21)	3.11 (-0.2)	2.5 (-0.3)	2.25 (-0.21)
4	-	-2.7 (0.44)	-	-	1.82 (-0.3)	-
5	-	-	-1.64 (0.33)	-	-	5.36 (-1.08)
6	-	-2.47 (0.25)	-	-	3.67 (-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	-	-1.9 (0.21)	-	-	2.26 (-0.25)	-
13	-3.14 (0.37)	-	-	2.12 (-0.25)	-	-
14	-	-	-1.84 (0.43)	-	-	4.7 (-1.1)
Mean	-2.24	-2.27	-1.67	3.15	2.76	4.07

Standard errors in parenthesis.

Table 3: Occupant Descriptive Statistics by Experiment

Experiment	*First Comparison*	*Tract Mean Age*	*Tract Median Income ($000s)*	*Tract Pct White*
1	-1.28 (1.88)	50.5 (5.3)	84.5 (30.3)	0.85 (0.21)
2	-0.47 (1.17)	49.4 (5.7)	68.2 (25.9)	0.76 (0.3)
3	-0.05 (1.41)	45.1 (2.4)	62.2 (9.4)	0.96 (0.02)
4	-1.08 (1.55)	43.9 (4.2)	56.2 (16.5)	0.91 (0.04)
5	0.3 (0.93)	43.3 (4.4)	50.5 (16.6)	0.9 (0.05)
6	-0.3 (1.23)	50.3 (3)	49.5 (21.6)	0.66 (0.1)
7	0.09 (1.11)	52.9 (2.1)	38.8 (6.5)	0.95 (0.07)
8	-0.28 (1.26)	51 (2.7)	65.9 (23)	0.93 (0.1)
9	-0.26 (1.18)	47.2 (3.5)	71.5 (19.8)	0.8 (0.09)
10	0.05 (1.35)	55.3 (8)	43.8 (12.8)	0.79 (0.14)
11	0.27 (1.24)	57 (7.5)	42.1 (16.6)	0.81 (0.15)
12	-0.41 (1.63)	48.1 (3.5)	54.1 (13.8)	0.72 (0.22)
13	-	48.5 (3.2)	51.5 (11.2)	0.73 (0.16)
14	-	48.6 (3.4)	48.4 (10.8)	0.71 (0.17)

Experiment-level means. Standard deviations in parenthesis.

Table 4: House Descriptive Statistics by Experiment

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 5: Heterogeneous Treatment Effects 1

	Expr 1	Expr 2	Expr 3	Expr 4	Expr 5	Expr 6	Expr 7
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
TxPost	0.4	-.85	19.52	3.34	1.00	1.99	1.28
	(2.67)	(1.90)	(6.45)***	(13.31)	(8.28)	(5.95)	(16.52)
TxPostxFirstComp	0.92	0.3	1.14	-.36	1.59	1.71	1.96
	(0.23)***	(0.21)	(0.22)***	(0.58)	(0.66)**	(0.45)***	(0.54)***
TxPostxCMeanAge	0.06	-.02	-.24	-.17	0.08	-.04	-.18
	(0.04)	(0.03)	(0.06)***	(0.14)	(0.08)	(0.15)	(0.28)
TxPostxCMedianIncome	0.005	-.0004	-.03	-.04	0.004	-.004	0.03
	(0.01)	(0.008)	(0.02)*	(0.04)	(0.03)	(0.02)	(0.1)
TxPostxCPctWhite	-2.33	-.88	-8.85	5.50	-6.32	-2.01	4.57
	(1.11)**	(0.69)	(6.23)	(17.57)	(10.01)	(4.06)	(6.39)
TxPostxElecHeat				-.26	-2.96		-1.60
				(1.55)	(1.96)		(1.29)
TxPostxHouseAge	-.02	-.0003	-.004	-.06	0.03	0.008	
	(0.01)*	(0.01)	(0.006)	(0.04)	(0.02)	(0.02)	
TxPostxHouseValue			-.002				
			(0.002)				
TxPostxPool						-2.22	
						(1.14)*	
TxPostxRent	1.02	0.07		0.74	-.45	0.79	3.82
	(1.42)	(0.75)		(1.41)	(0.98)	(0.69)	(2.32)*
TxPostxSingleFam	0.48	-.15		-.51	-1.90	-1.22	
	(0.8)	(0.47)		(1.51)	(1.02)*	(0.82)	
TxPostxSqFt	-.68	0.53	0.29	0.06	-.20	-.21	
	(0.34)**	(0.31)*	(0.51)	(0.59)	(0.63)	(0.78)	
Obs.	1055639	1574206	3262622	394525	914344	570375	750187
R^2	0.2	0.29	0.18	0.07	0.07	0.23	0.24
F statistic	2348.78	5372.74	3674.49	549.61	827.23	1071.49	340.51

*, **, ***: Different from zero with 90%, 95%, and 99% confidence, respectively. Standard errors in parentheses, robust and clustered by household. Dependent variable is electricity consumption, normalized by average control group consumption in the Post period. All specifications include month-by-year dummies, household fixed effects, and interactions of the X variables with the Post-treatment indicator.

Table 6: Heterogeneous Treatment Effects 2

	Expr 8	Expr 9	Expr 10	Expr 11	Expr 12	Expr 13	Expr 14
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
TxPost	-7.32	-.87	-.52	5.20	4.53	-8.40	-2.54
	(3.03)**	(2.16)	(2.94)	(5.56)	(3.51)	(2.61)***	(4.80)
TxPostxFirstComp	0.68	1.39	0.13	1.11	1.03		
	(0.22)***	(0.21)***	(0.39)	(0.8)	(0.37)***		
TxPostxCMeanAge	0.13	-.03	-.15	-.24	-.06	0.17	0.03
	(0.06)**	(0.05)	(0.08)*	(0.16)	(0.07)	(0.06)***	(0.11)
TxPostxCMedianIncome	-.006	0.02	-.02	0.03	-.04	-.01	-.005
	(0.008)	(0.009)*	(0.04)	(0.04)	(0.02)**	(0.02)	(0.04)
TxPostxCPctWhite	-.60	0.71	8.93	11.40	1.67	1.32	4.17
	(2.02)	(1.73)	(4.89)*	(8.24)	(1.24)	(1.19)	(2.25)*
TxPostxElecHeat		-.76			-1.11	-1.96	-2.38
		(0.58)			(0.65)*	(0.4)***	(1.21)**
TxPostxHouseAge	0.006	0.007	0.02	-.05	-.01	-.04	-.05
	(0.004)	(0.01)	(0.03)	(0.08)	(0.01)	(0.01)***	(0.02)***
TxPostxHouseValue		-.002			0.0005	0.0007	-.006
		(0.002)			(0.001)	(0.001)	(0.003)*
TxPostxPool	0.15		-.89	-6.78		-1.33	0.5
	(0.97)		(0.85)	(4.07)*		(0.39)***	(1.93)
TxPostxRent	1.22	-.90	-.77		-.27	-1.82	1.76
	(0.77)	(0.96)	(3.36)		(1.13)	(1.53)	(3.02)
TxPostxSingleFam				-.39			
				(2.04)			
TxPostxSqFt	0.01	-.26	0.37	-.94	-.56	-.46	-.15
	(0.28)	(0.38)	(1.16)	(1.41)	(0.43)	(0.39)	(0.85)
Obs.	1712495	3121879	985148	672629	640133	2543372	1036784
R^2	0.14	0.24	0.54	0.3	0.35	0.24	0.25
F statistic	1808.33	2450.86	1365.09	404.43	779.55	3889.49	1466.32

*, **, ***: Different from zero with 90%, 95%, and 99% confidence, respectively. Standard errors in parentheses, robust and clustered by household. Dependent variable is electricity consumption, normalized by average control group consumption in the Post period. All specifications include month-by-year dummies, household fixed effects, and interactions of the X variables with the Post-treatment indicator.

Table 7: Extrapolation Results

			$\widehat{\tau}_g$					%Trim		Error				t-Stats	
Approach	*Matching*	*Trim*	*Mean*	*SD*	*Min*	*Max*	*Mean(SE)*	*Sample*	*Target*	*Bias*	*Av Abs*	*RMSE*	*90-10*	*SD*	*% Reject*
Linear	N/A	None	-2.10	0.75	-4.35	-0.75	0.33	0.0	0.0	-0.03	0.77	0.98	2.63	2.12	49
Linear	Probit	Support	-2.14	0.90	-9.04	-0.81	0.35	0.7	0.6	-0.06	0.78	1.00	2.64	2.12	49
Linear	Probit	90/10	-2.09	0.72	-4.29	-1.03	0.33	0.5	0.3	-0.03	0.75	0.95	2.59	2.12	48
Linear	SL 1	Support	-2.14	0.91	-9.04	-0.66	0.35	2.2	4.1	-0.05	0.79	1.01	2.64	2.12	48
Linear	SL 1	90/10	-2.09	0.73	-4.25	-0.78	0.33	1.7	3.6	-0.02	0.76	0.97	2.60	2.12	47
Linear	SL 2	Support	-2.13	0.91	-9.04	-0.65	0.35	2.2	4.1	-0.04	0.79	1.01	2.63	2.12	48
Linear	SL 2	90/10	-2.10	0.73	-4.34	-0.89	0.33	2.0	3.9	-0.03	0.77	0.97	2.62	2.13	50
IPW	Probit	None	-3.45	20.59	-253.10	78.26	1.34	0.0	0.0	-1.39	3.75	20.60	3.03	2.46	50
IPW	Probit	Support	-3.25	20.59	-253.10	78.26	1.29	0.7	0.6	-1.16	3.75	20.59	3.02	2.46	51
IPW	Probit	90/10	-2.10	0.74	-4.52	-1.14	0.31	0.5	0.3	-0.05	0.76	0.97	2.57	2.23	48
IPW	SL 1	None	-1.73	10.64	-63.47	60.33	1.29	0.0	0.0	0.33	3.13	10.59	2.99	2.34	50
IPW	SL 1	Support	-1.66	9.39	-63.47	60.33	0.96	2.2	4.1	0.43	2.52	9.36	2.91	2.31	49
IPW	SL 1	90/10	-2.10	0.74	-4.49	-1.09	0.32	1.7	3.6	-0.04	0.76	0.98	2.62	2.21	47
IPW	SL 2	None	-0.38	29.35	-276.64	169.57	1.59	0.0	0.0	1.69	7.12	29.33	3.33	11.09	52
IPW	SL 2	Support	-0.24	28.49	-276.64	169.57	1.33	2.2	4.1	1.85	6.06	28.49	3.00	10.95	49
IPW	SL 2	90/10	-2.10	0.76	-4.58	-0.96	0.32	2.0	3.9	-0.04	0.78	0.99	2.63	2.23	50
Non-Exper.	N/A	None	-3.15	2.21	-6.27	1.97	0.22	0.0	0.0	-1.08	2.09	2.36	5.03	7.27	100

Table 8: 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 Utility)	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)
Average Usage (MWh/year)	12.41	9.43	12.56	-3.13
	(3.41)	(2.39)	(3.39)	(0.37)***
EE Spending (/cust.-yr)	15.38	25.10	14.89	10.20
	(138.16)	(29.68)	(141.43)	(6.45)
Energy Conservation (kWh/cust.-yr)	25.30	48.33	24.14	24.19
	(178.14)	(69.87)	(181.84)	(11.97)**
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 Has EERS	0.57	0.93	0.55	0.39
	(0.50)	(0.25)	(0.50)	(0.04)***
State Has RPS	0.52	0.84	0.50	0.34
	(0.50)	(0.37)	(0.50)	(0.06)***
State Median Income (1000's)	49.31	55.87	48.98	6.89
	(6.84)	(5.51)	(6.73)	(0.84)***
State Pct College Degree	25.99	29.49	25.82	3.68
	(4.43)	(3.82)	(4.38)	(0.58)***
State Pct Democrat Voters	49.09	55.46	48.77	6.69
	(9.36)	(8.61)	(9.28)	(1.31)***
log(Residential 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 9: Selection and Associations with ATE

	Corr with ATE	*Probit 1*	*Probit 2*	*Probit 3*	*Probit 4*
1(Coop)	-0.86	0.65			
	(0.33)**	(0.67)			
1(Investor-Owned Utility)	0.90	0.63	0.47		
	(0.20)***	(0.57)	(0.28)*		
1(Municipal)	-0.15	1.20	0.88		
	(0.31)	(0.58)**	(0.26)***		
1(Other Government)	-0.48	0.12			
	(0.35)	(0.74)			
Average Usage (MWh/year)	-0.11	-0.05		-0.08	-0.07
	(0.13)	(0.03)*		(0.02)***	(0.02)***
EE Spending (/cust.-yr)	0.008	0.000			
	(0.008)	(0.000)			
Energy Conservation (kWh/cust.-yr)	0.003	0.000			
	(0.002)	(0.000)			
Pct Green Pricing	-0.063	0.012			
	(0.052)	(0.009)			
Price (cents/kWh)	0.046	-0.008			
	(0.049)	(0.026)			
log(Residential Customers)	0.27	0.519	0.447		0.448
	(0.08)***	(0.086)***	(0.071)***		(0.056)***
State Has EERS		0.42	0.90	0.61	
		(0.35)	(0.28)***	(0.26)**	
State Has RPS		0.26			
		(0.28)			
State Median Income (1000's)	-0.085	0.060		0.039	0.051
	(0.042)**	(0.022)***		(0.011)***	(0.011)***
State Pct College Degree	-0.011	-0.044			
	(0.063)	(0.039)			
State Pct Democrat Voters	0.020	0.013		0.005	0.007
	(0.009)**	(0.010)		(0.006)	(0.006)
N	10	938	939	938	938
F Test p-Value		0.000	0.000	0.000	0.000
Psuedo R2		0.397	0.333	0.176	0.352
Coefficient: ATE on Fitted Probability		1.89	2.09	2.133	1.868
SE (Murphy-Topel)		(0.42)***	(0.45)***	(4.818)	(0.468)***

Standard errors in Column 1 are robust and account for the standard error in the estimated ATE (Murphy and Topel 2002). Standard errors in Columns 2-5 are robust.

Table 10: 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

Notes: Currencies are in US dollars at market exchange rates. Percent of Portfolio at Risk is the percent of gross loan portfolio that is renegotiated or overdue by more than 30 days.

Table 11: 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.

Figures

Figure 1: Home Energy Reports: Social Comparison Module

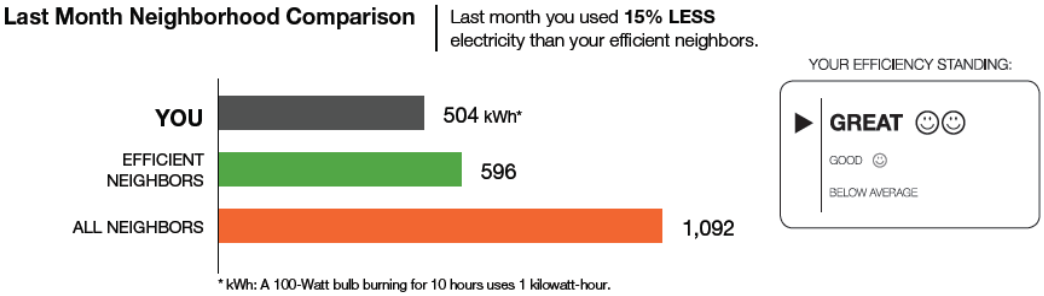


Figure 2: Home Energy Reports: Action Steps Module

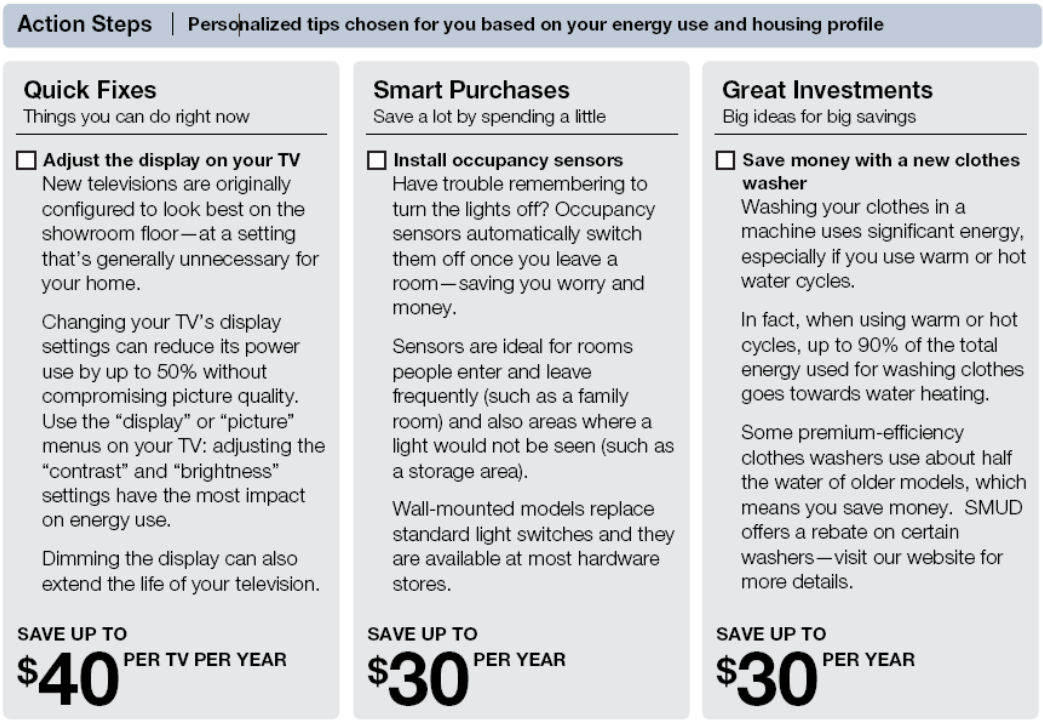


Figure 3: Map of OPOWER Locations

Figure 4: t-Statistics vs. Standard Normal

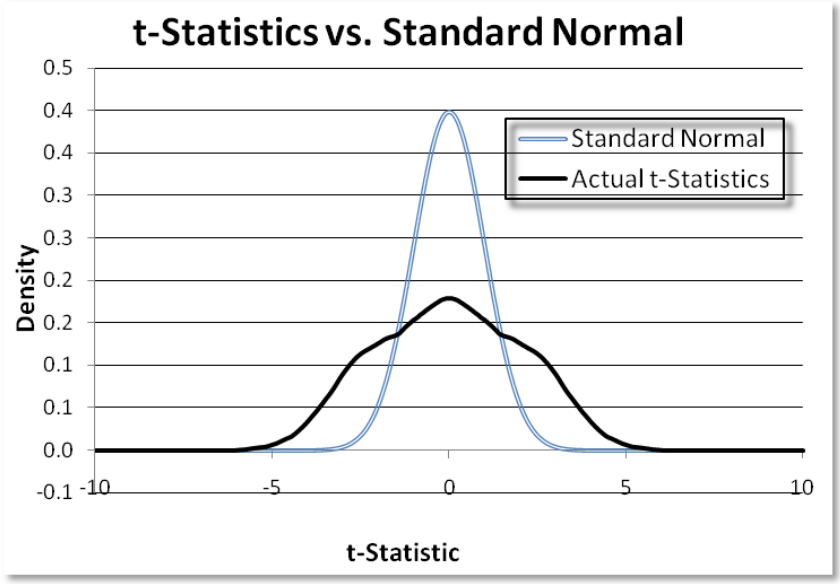


Figure 5: ATEs and Extrapolation Errors

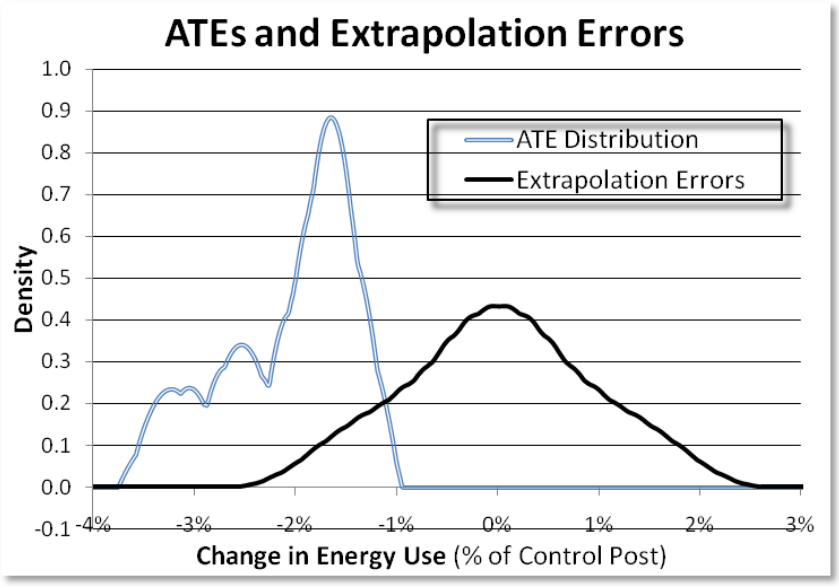


Figure 6: Selection Probability and ATE

