NBER WORKING PAPER SERIES

EXTERNAL VALIDITY AND PARTNER SELECTION BIAS

Hunt Allcott
Sendhil Mullainathan

Working Paper 18373
http://www.nber.org/papers/w18373

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2012

We thank Josh Angrist, Amitabh Chandra, Lucas Davis, Angus Deaton, Meredith Fowlie, Xavier Gine, Chuck Goldman, Matt Harding, Joe Hotz, Guido Imbens, Larry Katz, Dan Levy, Jens Ludwig, Konrad Menzel, Rohini Pande, Todd Rogers, Ed Vytlacil, and seminar participants at the ASSA meetings, Berkeley, Harvard, MIT, NBER, NEUDC, the UCSB/UCLA Conference on Field Experiments, and the World Bank for insights and helpful comments. Thanks also to Tyler Curtis, Marc Laitin, Alex Laskey, Dan Yates, and others at Opower for fruitful discussions. Piyush Tantia and Scott Gaul provided useful help in accessing the Microfinance Information Exchange data. We are grateful to the Sloan Foundation for financial support for this paper and our related research on the economics of energy efficiency. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

External Validity and Partner Selection Bias
Hunt Allcott and Sendhil Mullainathan
NBER Working Paper No. 18373
September 2012
JEL No. C93,D12,L94,O12,Q41

ABSTRACT

Program evaluation often involves generalizing internally-valid site-specific estimates to a different target population. While many analyses have tested the assumptions required for internal validity (e.g. LaLonde 1986), there has been little empirical assessment of external validity in any context, because identical treatments are rarely evaluated in multiple sites. This paper examines a remarkable series of 14 energy conservation field experiments run by a company called Opower, involving 550,000 households in different cities across the U.S. We first show that the site-specific treatment effect heterogeneity is both statistically and economically significant. We then show that Opower partners are selected on partner-level characteristics that are correlated with the treatment effect. This "partner selection bias" implies that replications with additional partners have not given an unbiased estimate of the distribution of treatment effects in non-partner sites. We augment these results in a different context by showing that partner microfinance institutions (MFIs) that carry out randomized experiments are selected on observable characteristics from the global pool of MFIs. Finally, we propose two simple suggestive tests of external validity that can be used in the absence of data from many sites: comparison of observable sample and target site characteristics and an F-test of heterogeneous treatment effects across "sub-sites" within a site.

Hunt Allcott
Department of Economics
New York University
19 W. 4th Street, 6th Floor
New York, NY 10012
and NBER
hunt.allcott@nyu.edu

Sendhil Mullainathan
Department of Economics
Littauer M-18
Harvard University
Cambridge, MA 02138
and Consumer Financial Protection Bureau
and also NBER
mullain@fas.harvard.edu

1 Introduction

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

When generalizing empirical results, we often implicitly or explicitly make one of two assumptions. First, if extrapolating from one sample site to one different target, we might assume that the site-specific treatment effect heterogeneity is small enough that the results can be meaningfully generalized. However, in some contexts, this assumption is unrealistically strong, meaning that it is important to replicate in additional sites. After enough replications, we might make a second assumption: that the distribution of treatment effects in the sample sites is a reasonable predictor of the distribution of effects in other target sites. This assumption would hold if the sample sites had been selected randomly from the population of target sites.

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

While there is a substantial theoretical discussion of external validity[1] and the importance of the problem is broadly recognized[2], we know very little about the nature of partner selection bias and other external validity problems in practice. The reason is simple: to explicitly test for heterogeneous site-level effects, one needs to compare results from multiple internally valid studies in multiple sites. However, it is unusual for an identical treatment to be experimentally or quasi-

[1] Formal theoretical analyses of external validity are included in Angrist (2004), Heckman (1992), Heckman and Vytlacil (2007a, 2007b), Hotz, Imbens, and Mortimer (2005), Imbens (2010), and others.

[2] Other recent articles that contain discussions of the importance of external validity include Angrist and Pischke (2010), Banerjee (2009), Cartwright (2007, 2010), Deaton (2010a, 2010b), Duflo (2004), Duflo, Glennerster, and Kremer (2007), Greenberg and Shroder (2004), Heckman and Smith (1995), Ludwig, Kling, and Mullainathan (2011), Manski and Garfinkel (1992), Manski (2011), Rodrik (2009), Rothwell (2005), Worrall (2007), and many others. See also Campbell (1957) for an early discussion.

experimentally evaluated multiple times, because randomized field experiments are costly and useful natural experiments are rare.[3] By contrast, many papers provide evidence on individual selection bias and the "internal validity" of an estimator, as this requires a comparison of an internally valid estimate to the non-experimental results in only one setting. Explicitly testing for partner selection bias is even more difficult: one must define a population of potential partner sites and somehow infer treatment effects in sites where studies have not yet been carried out.

In this paper, we empirically analyze a series of 14 randomized experiments involving more than one-half million households in different sites across the United States. The experiments are run by a company called Opower, which mails Home Energy Reports to residential electricity consumers that provide energy conservation tips and compare their energy use to that of their neighbors. Because these Reports are effectively the same in each site and because there is effectively no non-compliance with treatment assignment, we have the unusual opportunity to focus on one particular aspect of external validity: how well the effects of an identical treatment can be generalized across heterogeneous populations and economic environments.[4] The quantitative results are of course context-specific. However, just as LaLonde (1986) and other context-specific studies of individual-level selection bias[5] have been broadly informative about internal validity, some of the qualitative findings from this type of analysis may similarly be informative about aspects of external validity.

The generalizability of the Opower program's effects to potential future sites is also of great interest *per se*. This is because a proliferation of new regulations mandating energy conservation, spurred partially by concern over climate change and high energy prices, is causing many utilities across the country to decide whether to adopt the program. Opower is also of special interest to us because we have extrapolated the results from one early experiment, implicitly assuming strong external unconfoundedness. We carried out this extrapolation in a short article in Science magazine, where we argued that the treatment effects from one Opower experiment in Minnesota suggested that a nationwide rollout of the program would be cost effective relative to other energy

[3] There is some literature that compares impacts of programs implemented at multiple sites. In the development field, this includes Banerjee, Cole, Duflo, and Linden (2007), Chattopadhyay and Duflo (2004), a pair of related papers by Miguel and Kremer (2004) and Bobonis, Miguel, and Sharma (2006), and some more recent ongoing experiments. Also, the YouthBuild program being run by the U.S. Department of Labor requires that all sites receiving funding participate in randomized evaluations, and the French job training program studied by Crepon *et al.* (2012) randomly assigns participants at 235 different sites.

Quasi-experimental estimates can also be compared across locations or across groups to whom different instruments are "local," as in Angrist and Fernandez-Val (2010) and Angrist, Lavy, and Schlosser (2010). Of course, as one weakens the definition of what a "similar" treatment is, there are increasingly large literatures of meta-analyses that compare the effects of "similar" treatments in different settings, including Abdulkadiroglu, Angrist, Dynarski, Kane, and Pathak (2009) and Angrist, Pathak, and Walters (2011) on pilot and charter schools, Aigner (1984) on electricity pricing, Card, Kluve, and Weber (2009) on labor market policies, and Meyer (1995) on unemployment insurance.

[4] It is well-understood that there are other classes of threats to a study's external validity. Randomized trials may suffer from Hawthorne effects, in which the subjects behave differently because they know they are being studied. Subjects who choose or are allowed to select into randomized trials may differ from the population of interest. 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.

[5] Other closely-related studies include Dehejia and Wahba (1999), Heckman, Ichimura, and Todd (1997), Heckman Ichimura, Smith, and Todd (1998), and Smith and Todd (2004).

conservation programs and would generate billions of dollars in energy cost savings each year (Allcott and Mullainathan 2010).

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

We also use the Opower example to provide evidence on partner selection bias. Our test exploits the fact that we observe the characteristics of the population of Opower's potential partner sites: the set of electric utilities in the United States. We show that Opower's current partners are selected on site-level observables: partner utilities 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 on observables: selection probabilities conditional on observables are systematically correlated with treatment effects. This suggests that even relatively extensive replication has not solved the external validity problem in the Opower context: ATEs in partner sites are unlikely to be an unbiased measure of ATEs in non-partner sites.

As with LaLonde (1986) and related analyses of experimental vs. non-experimental estimators, the Opower field experiments are only one example in one setting. To provide one additional data point on the conceptual issue of partner selection bias, we turn to microfinance. We examine the characteristics of microfinance 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 average loan size, staff per borrower, for-profit status, years of experience since opening, and size. Because microfinance field experiments study a variety of different "treatments," we cannot correlate selection probabilities with treatment effects as we do for the Opower experiments. However, this basic evidence of selection suggests that partner selection bias may not be unique to the Opower energy conservation programs.

Indeed, analyses of the Job Training Partnership Act of 1982 (JTPA) also provide closely-related existing evidence. The JTPA initiated job training programs at 600 sites, of which 16 were evaluated with randomized trials. These 16 experimental sites were those that agreed to participate out of more than 200 that were originally approached (Hotz 1992). Heckman (1992) discusses how "randomization bias" may have affected the selection of experimental sites and populations within the sites. Even if the 16 sites were representative of the broader population of sites, Heckman and Smith (1997) simulate that because of the substantial variability in effects across sites, the aggregate experimental impact estimates would have differed substantially depending on which set of sites were evaluated. Our paper complements this work by providing large-sample evidence of site effects and partner selection bias in a different context and by formalizing a simple model of the partner selection process.[6]

Of course, one rarely has the luxury of a multi-site program evaluation. We propose a simple set of four concrete steps that analysts can take when generalizing empirical results. First, we can clearly define the target site or population of interest. Second, just as it is common to provide evidence on internal validity by comparing observable characteristics of treatment and control groups, we can provide suggestive evidence on external validity by comparing the observable characteristics of the sample population and the target population of policy interest. Similarly, we can compare the observable characteristics of the experimental partner to the observable characteristics of other organizations that might implement a scaled program. Third, observable characteristics can be combined with a theoretical discussion of the partner selection process and how the experimental population and partner might differ on unobserved characteristics that moderate the treatment effect.

The fourth potential step is an empirical test that provides suggestive evidence on strong external unconfoundedness: an F-test of treatment effect heterogeneity across sub-sites within the sample site. The idea is very simple: when individuals are categorized into sites by some factor such as geographic location, that same factor can be used to categorize at a more disaggregated level. Put differently, within each site, there are a set of "sub-sites," such as zip codes within a city, schools within a district, or job training centers within a county. If we reject the null of no unexplained heterogeneity across sub-sites within the sample, this suggests unexplained heterogeneity between sample and target sites unless the distribution of sub-site heterogeneity somehow happens to be identical in sample and target. For example, in many contexts, one might extrapolate from one geographic location to another and assume that the geographic heterogeneity is sufficiently small. The assumption of no geographic heterogeneity can be explicitly tested on a more disaggregated level using geographic sub-sites within the sample. In the body of the paper, we develop this idea more formally and discuss the possibilities for Type I and Type II errors.[7]

We emphasize from the outset that our analysis cannot be used to argue that randomized

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

[7] Ours is not the only suggestive test of strong external unconfoundedness. Hotz, Imbens, and Mortimer (2005) and Hotz, Imbens, and Klerman (2006) test whether control group data from a sample can predict outcomes in untreated target sites. The assumption underlying this test is that if untreated outcomes can be predicted in the target, then is more likely that treatment effects can be predicted. Stuart, Cole, Bradshaw, and Leaf (2011) propose a version of this same test using propensity score methods. Angrist and Fernandez-Val (2010) test whether differences in Local Average Treatment Effects (LATEs) from different instruments can be explained by observable characteristics of compliers. The assumption underlying this test is that if different instruments give the same conditional LATEs in samples of compliers, then it is more likely that treatment effects can be predicted for target populations comprised of always-takers and never-takers.

control trials are not useful and important in this context. As shown in Allcott (2011), non-experimental approaches to evaluating the Opower programs that would necessarily be used in the absence of experimental data perform dramatically worse than experimental estimators in the same population. In fact, non-experimental estimates from the correct Target population also perform substantially worse than treatment effects predicted for the Target using experimental data from different Sample populations. Furthermore, while partner selection bias largely pertains to RCTs, the rest of our discussion of the generalizability of site-specific parameter estimates is relevant to both "structural" and "reduced form" parameters estimated using either randomized experiments or natural experiments.

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

2 Model

We begin the model by setting up the basic Rubin (1974) Causal Model with selection into treatment from a generalized Roy (1951) model. We then aggregate to the site level and draw direct analogies between the assumptions for internal and external validity and the selection processes that compromise these assumptions.

2.1 Individual-Level Model

2.1.1 Setup

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, $Y_i(1)$ if exposed to treatment and $Y_i(0)$ if not. For expositional simplicity, we assume that Y_i is a linear and additively-separable function of observed and unobserved characteristics X_i and Z_i:

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

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

The linear functional form is not central to our argument and could certainly be relaxed. What is central is that there are individual-level unobservables Z that influence the treatment effect.

Individual i's treatment effect is the difference in Y_i between the treated and untreated states:

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

In the context of program evaluation, an estimator is "internally valid" if it can be used to consistently estimate the Average Treatment Effect for some subpopulation. Denote $T_i \in \{1, 0\}$ as the indicator variable for individual i's actual treatment assignment. Comparing the mean outcomes of treated vs. untreated units gives:

$$\begin{aligned} E[Y_i(1)|T_i = 1] - E[Y_i(0)|T_i = 0] &= E[\tau_i|T_i = 1] \\ &+ \beta(E[X_i|T_i = 1] - E[X_i|T_i = 0]) + \zeta(E[Z_i|T_i = 1] - E[Z_i|T_i = 0]) \end{aligned} \tag{3}$$

The right hand side of the first line is the Average Treatment Effect on the Treated (ATT). The second line is selection bias. The first term in the second line is a function of observables X and can be estimated empirically. The second term is a function of unobservables Z.

2.1.2 Unconfoundedness

The second term in the second line of Equation (3) above equals zero under the assumption of *unconfoundedness* (Rosenbaum and Rubin 1983):

$$T_i \perp (Y_i(1), Y_i(0))\,|X_i \tag{4}$$

If unconfoundedness does not hold, then the ATT cannot be consistently estimated, meaning that there is a failure of internal validity.

2.1.3 Assignment Mechanisms

Imbens and Wooldridge (2009) specify three classes of mechanisms through which individuals are assigned to treatment or control. The first is random assignment, under which unconfoundedness should hold by construction. The second is non-experimental mechanisms under which unconfoundedness holds by assumption. The third is all other assignment mechanisms under which unconfoundedness does not hold.

2.1.4 The Economics of Selection

In the absence of experimental assignment, there are economic processes that drive selection into treatment. One natural process is that individuals decide whether to participate in a program if the private benefits outweigh the private costs. To model this, assume that individuals incur some positive or negative net cost C_i if they select into the program, and they weight their outcome Y_i by ω. An individual selects into treatment if the net benefits are positive:

$$\begin{aligned} T_i &= 1\left[\omega\tau_i - C_i > 0\right] && \text{(5a)} \\ &= 1\left[\omega(\alpha X_i + \gamma Z_i) - C_i > 0\right] && \text{(5b)} \end{aligned}$$

If individuals self-select into treatment in this way, unconfoundedness holds if and only if it happens to be the case that $(\omega(\alpha X + \gamma Z) - C) \perp Z$.[8] Otherwise, there is positive or negative selection: referring to Equation (3), the conditional difference in mean outcomes between treatment and control may be larger or smaller than the ATT. Of course, the meaning of C_i and ω, and thus the economics of the selection process, vary by application.

2.2 Site-Level Model

2.2.1 Setup

Imagine now that the population of individual units is divided mutually exclusively and exhaustively into "sites." Informally, we think of sites as sets of individual units that often are geographically clustered, where one program evaluation might be carried out. In applied work, this might be a school or school district, a job training center, or a microfinance institution. Index sites by r, and define an integer variable R_i that indicates the site of which individual i is a member. Denote the population of sites by script $\mathcal{R}$. Within each site is also a "site-level population" of individual units.

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

$$\tau_r = \alpha X_r + \gamma Z_r \tag{6}$$

[8] Here and in analogous future statements, we formally mean that selection must be independent of $Z|X$. For expositional simplicity, we refer to the "unobservables" instead of "the unobservables conditional on the observables."

In the context of program evaluation, we refer to an estimator as being "externally valid" if it can use data from one or more Sample sites to consistently estimate treatment effects in other Target sites. We consider two different senses of external validity. First, suppose that there is one Sample site, indexed $r = s$. Imagine that the Sample ATE τ_s can be consistently estimated, and the analyst wishes to generalize to one Target site, indexed $r = g$. The ATE in the Target is:

$$\begin{aligned} \tau_g &= \tau_s \\ &+ \alpha(X_g - X_s) + \gamma(Z_g - Z_s) \end{aligned} \tag{7}$$

Second, suppose that the Target sites are the entire population of sites $\mathcal{R}$, and imagine that a program evaluation has been replicated in many Sample sites drawn from $\mathcal{R}$. Denoting $D_r \in \{1, 0\}$ as an indicator for whether site r is a Sample site, the expectation of the Target treatment effects is:

$$\begin{aligned} E[\tau_r] &= E[\tau_r | D_r = 1] \\ &+ \alpha\left(E[X_r] - E[X_r | D_r = 1]\right) + \gamma\left(E[Z_r] - E[Z_r | D_r = 1]\right) \end{aligned} \tag{8}$$

These two equations are comparable to each other and to Equation (3) above. The right hand side of the first line is the ATE in the Sample, or the expectation across many Samples. The second line is the bias from extrapolation: the difference between Target and Sample ATEs. The first term in the second line is a function of observables X and can be estimated empirically. The second term is a function of unobservables Z.

2.2.2 External Unconfoundedness

Denote $D_i \in \{1, 0\}$ as an indicator variable for whether individual i is a member of a Sample site. Hotz, Imbens, and Mortimer (2005) introduce the assumption of unconfounded location, which we also call *external unconfoundedness*:

$$D_i \perp (Y_i(1), Y_i(0)) | X_i \tag{9}$$

In words, external unconfoundedness is the assumption that whether an individual unit is a member of a Sample or Target site-level population is independent of potential outcomes conditional on observables X.[9] Here, we propose that this assumption has two different interpretations. The first is relevant when there is a pair of Sample and Target sites, as in Equation (7). The second is

[9] In our context, we could instead formalize weaker assumptions about independence of the *difference* in potential outcomes: $D_i \perp (Y_i(1) - Y_i(0)) | X_i$.

relevant in the context of replication, when there are many Sample and Target sites, as in Equation (8).

Consider first the case when the analyst has data from one Sample site-level population and wishes to extrapolate to one Target site-level population. This extrapolation from one Sample to one Target requires an assumption we call *strong external unconfoundedness.*

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

In words, this is that external unconfoundedness holds in a *pair* of sites, the Sample and the Target. This means that the distribution of unobservables is identical in Sample and Target. Therefore, the Conditional Average Treatment Effects (CATEs), conditional on X, are also asymptotically equal. If strong external unconfoundedness does not hold, the researcher cannot infer the Target treatment effect. An estimator based on Equation (7) above would not satisfy what we call *strong external validity.*

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

In many contexts, one expects unobservables to vary across sites, and strong external unconfoundedness is unrealistically restrictive. As a result, the analyst may wish to replicate an experiment in additional sites, or perform a meta-analysis. Suppose that the Target sites are the entire population of sites $\mathcal{R}$, and imagine that the researcher could draw a random sample of sites from $\mathcal{R}$. As the number of randomly-selected Sample sites increases, the distribution of treatment effects in the set of Sample sites would asymptotically equal the distribution of treatment effects in the Target sites. This motivates an assumption we call *external unconfoundedness in distribution.*

Definition 2 *External Unconfoundedness in Distribution:* $D_i \perp (Y_i(1), Y_i(0)) | (X_i, R_i \in \mathcal{R})$

In words, this is that external unconfoundedness holds in a *population* of sites. When there is exactly one Sample site and one Target site, this assumption is identical to strong external unconfoundedness. However, when there are replications in many sites, external unconfoundedness in distribution is a weaker assumption. Under this assumption, distributions of unobservables may differ between any pair of Sample and Target sites, as long as the unobservables in the sets of Sample and Target sites converge in distribution as the number of sites grows large. As a result, the CATEs from any one Sample may not equal the CATEs from any one Target. However, the mean CATE from Sample sites is a consistent estimator of the expected CATE in Target sites as the number of sites increases. Put simply, external unconfoundedness in distribution means that once a program is replicated in enough sites, the distribution of Target treatment effects is known. If this assumption does not hold, however, an estimator based on Equation (8) would not satisfy what we call *external validity in distribution.*

2.2.3 Assignment Mechanisms

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

We specify three classes of mechanisms that assign a potential partner site to being an actual partner for a randomized control trial. These parallel the individual-level assignment mechanisms. The first is random assignment: Sample sites are randomly selected from $\mathcal{R}$, the population of Target sites. As the number of sites grows large, external unconfoundedness in distribution holds. Of course, it is rare that the number of experimental sites would be large enough for asymptotics to be valid. With a small number of sites such as the 14 in the Opower example, unobservables may not be balanced between Sample and Target even if the Sample sites were randomly selected from the population of sites. In finite sample, just as stratified randomization can improve balance between treatment and control groups, stratified partner sampling can improve balance between Sample and Target sites. For example, the JTPA evaluation initially hoped to randomly select sites for evaluations within 20 strata defined by size, region, and a measure of program quality (Hotz 1992).

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

The third class of partner assignment mechanisms includes all other assignment mechanisms under which external unconfoundedness in distribution does not hold. In the absence of random assignment or other processes intentionally designed to maximize external validity, there are economic processes that drive selection into partnership. One natural process is that decisionmakers at each potential partner site decide whether to adopt or evaluate a program based on whether the costs outweigh the benefits. As in the individual-level model, we assume that the decisionmaker knows the treatment effect; we would obtain analogous results under imperfect information as long as the decisionmaker has some informative signal that the analyst does not observe. Potential partners incur some positive or negative net cost C_r of adopting or evaluating the treatment and weight average outcomes Y_r by ω. The potential partner becomes an actual partner if its net benefits are positive:

$$\begin{aligned} D_r &= 1\left[\omega\tau_r - C_r > 0\right] && \text{(10a)} \\ &= 1\left[\omega(\alpha X_r + \gamma Z_r) - C_r > 0\right] && \text{(10b)} \end{aligned}$$

If this process determines selection into partnership, external unconfoundedness in distribution only holds only if it happens to be the case that $(\omega(\alpha X + \gamma Z) - C) \perp Z$. Otherwise, there is positive or negative *partner selection bias*: referring to Equations (7) and (8), the Sample ATEs may be larger or smaller than the Target ATEs.[10]

2.2.4 The Economics of Partner Selection

So far we have been very general about the practical meaning of ω and C and how different real-world factors might generate positive or negative partner selection bias. These factors will vary across contexts, and many different models might apply. Here we flesh out an example set of partner selection mechanisms that might be relatively general and discuss the sign of $corr((\omega(\alpha X + \gamma Z) - C), \gamma Z)$.

A first category of mechanisms is driven by the mechanical correlation between $\omega\gamma Z$ and γZ: if potential partners care about outcomes when they decide whether to evaluate a program, and if they have some private information about outcomes that the analyst does not have, then actual partners will be selected on unobservables. One mechanism within this category results from the fact the "It Pays to Be Ignorant" (Pritchett 2002). Because rigorous evaluations are publicized and affect funding from foundations, governments, and other sources, potential partners that believe they are running effective programs are willing to have them evaluated, while those that believe they are running ineffective programs strategically choose to remain ignorant by avoiding randomized evaluations. Other sources of imperfect information about program quality keep this equilibrium from unraveling into an equilibrium in which RCTs are run at all sites. A second mechanism within this category is a simple form of "ability bias": even if there is no agency problem between potential partners and their funders, potential partners considering whether to adopt and evaluate a new program only want to do so if they believe it will work well at their site. Both of these two example

[10] Partner selection bias is related to the discussion of "randomization bias" that originates in Heckman (1992) and continues in later work (e.g. Heckman and Smith 1995, Heckman and Vytlacil 2007b), although that discussion is more directly concerned with how randomized experiments affect the selection of individuals into programs at the partner sites.

mechanisms generate positive partner selection bias: potential partners with unobservably higher returns are more likely to adopt a program.

A second category of mechanisms is driven by the potential correlation between C and γZ: if the net costs of running an RCT are positively or negatively correlated with unobservable moderators of the treatment effect, then actual partners will differ from non-partners on unobservables. One mechanism within this category results from the fact that implementing randomized trials requires managerial ability and operational efficacy. The potential partners that are best equipped to run RCTs, and thus have low C, may also run the most effective programs. This form of positive partner selection bias is related to the idea of gold plating (Duflo, Glennerster, and Kremer 2008): in order to cleanly measure efficacy, treatments are often implemented with much greater precision and quality in experimental settings than they would be elsewhere. A second mechanism within this category is "diminishing returns bias." Potential partners that might be most capable and interested in experimenting with new programs could also be running many other effective programs or could have already treated the parts of their population that have the largest treatment effects. This would generate negative partner selection bias.

Of course, partner selection bias does not mean that the estimated Sample ATEs are biased away from the true Sample ATEs. At its core, our model simply illustrates heterogeneous Conditional Average Treatment Effects that could vary across sites. The reason why we use the phrase "partner selection bias" is to emphasize that these CATEs in the set of partner sites may be *systematically* different from the effects in the set of non-partner sites. Furthermore, these systematic differences arise from a selection process that can be theoretically understood and observed in practice.

2.2.5 The Magnitude of Partner Selection Bias

What is the magnitude of partner selection bias? More precisely, how much does the expected Sample ATE differ on unobservables compared to the expected ATE in the population of Target sites $\mathcal{R}$? We can see this mathematically through an analogy to Heckman's (1979) exposition of individual-level selection bias. For simplicity, assume that γZ_r and $\omega\tau_r - C_r$ are jointly normally distributed in the population of sites, with standard deviations $\sigma_{\gamma Z}$ and $\sigma_{\omega\tau-C}$, respectively, and correlation coefficient ρ. We define $\psi = E[\omega\tau_r - C_r]$ and, without loss of generality, impose that $E[\gamma Z_r] = 0$, because X can include a constant. If selection is governed by Equation (10a), then the expected ATE in the Sample sites is:

$$E[\tau_r|\,(X_r, D_r = 1)] = \alpha E[X_r] + E\left[\gamma Z_r|\omega\tau_r - C_r > 0\right] = E[\tau_g] + \sigma_{\gamma Z}\cdot\rho\cdot\frac{\phi\left(\frac{-\psi}{\sigma_{\omega\tau-C}}\right)}{\Phi\left(\frac{\psi}{\sigma_{\omega\tau-C}}\right)} \qquad (11)$$

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

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

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

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

There are two basic takeaways from this section. First, just as individual units may endogenously select into treatment, there is an analogous self-selection process for partners in randomized controlled trials. Second, unconfoundedness and the external unconfoundedness assumptions are mathematically similar and have similar statistical and economic implications. Despite this, external unconfoundedness often receives much less attention in applied work. In the following sections, we test the two versions of external unconfoundedness in one particular context.

3 Opower Experiment Overview

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

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

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

Aside from treatment fidelity, there are two other useful features of the Opower experiments. First, in the taxonomy of Levitt and List (2009), these are "natural field experiments," meaning that people are in general not aware that they are being studied. Therefore, there are no "Hawthorne Effects." Second, because opting out of the letters requires active effort, there is effectively no non-compliance. This means that there is no need to model essential heterogeneity or the individual-level selection into the experimental treatment (Heckman, Urzua, and Vytlacil 2006), and the treatment effect is a Policy-Relevant Treatment Effect in the sense of Heckman and Vytlacil (2001).[11]

As of the end of 2010, Opower had contracts to work with 45 utilities in the 21 shaded states in Figure 3. While the partners are spread throughout the country, they tend to be concentrated along

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

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 for-profit 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 that are supposed to act in the public interest, with goals that often include environmental conservation and saving money for their customers. In Section 5, we quantitatively analyze the characteristics of Opower's partners.

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

3.1 Data

Table 1 provides an overview of the start date and size for each experiment. In total, we observe 19 million monthly electricity bills from 553,798 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. Site pairs 1 and 2, 4 and 5, 10 and 11, and 13 and 14 are the four involving different customer subpopulations at the same utility.

This study benefits from exceptionally good household-level data, which improves the likelihood that we might be able to use these data to explain differences in treatment effects across locations. Opower, and the utilities they work with, gather demographic data for each customer from surveys, public records, and private-sector marketing data providers. In addition, we have augmented the household-level data with Census Tract-level information from the 2000 U.S. Census. From the outset, we focused our analysis on the set of covariates that theory predicts might moderate the treatment effect.

Table 2 details the means and standard deviations of the observed individual-level characteristics X for each of the 14 sites. There are four categories of variables: weather, energy use, Census Tract-level demographics, and house characteristics. The first two columns are heating and cooling degree-days, which measure how far temperatures deviate from 65 degrees during each month, and thus how much electricity might be required to heat or cool a house to a comfortable temperature.[12]

[12] More precisely, the average Cooling Degree-Days for an observation is the mean, over all of the days in the billing period, of the maximum of zero and the difference between the day's average temperature and 65 degrees. A day with average temperature 75 has 10 CDDs, while a day with average temperature 30 has zero CDDs. Average Heating Degree-Days is the mean, over all the days in the billing period, of the maximum of zero and the difference between

These vary over time within a site, but they do not vary across households within a site on any given day. Sites 10 and 11 are in an especially warm climate, with low average heating degrees and high cooling degrees, while sites 13 and 14 are in a moderate climate, and many other sites are relatively cold. The third column is "Baseline Comparison," a normalized measure of the household's baseline energy usage compared to its neighbors, as presented to them on the first Home Energy Report they receive. Zero corresponds to the mean of the neighbor distribution, and households with lower values used relatively more energy. As detailed in Allcott (2011), theory predicts that responses to these social comparisons depend on how individuals compare to their neighbors, and the treatment effects vary substantially with baseline energy usage.

As documented in Costa and Kahn (2010), households that vote Democratic, donate to environmental groups, or voluntarily purchase renewable energy have different treatment effects. The next three variables in Table 2 are Census tract-level average characteristics which we hypothesized could be associated with these sorts of "cultural" differences that moderate the treatment effect. The final seven variables in Table 2 are house characteristics. 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 or information regarding 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 Opower Site-Level Heterogeneity

In this section, we test the assumption of strong external unconfoundedness: how much unexplained heterogeneity is there across Opower sites? We first present the empirical specification, then results, then discuss economic significance.

4.1 Empirical Specification

The Average Treatment Effects (ATEs) for each site can be estimated simply by comparing post-treatment energy usage Y_{it} for treatment and control groups, controlling for pre-treatment average usage Y_{i0} and site-specific constant π_r. Denoting $1(R_i = r)$ as an indicator function for whether household i is in site r, the unconditional estimating equation is:

65 degrees and the day's average temperature. A day with average temperature 75 has zero HDDs, while a day with average temperature 30 has 35 HDDs.

$$Y_{it} = \sum_r \left[\tau_r T_i + \phi_r Y_{i0} + \pi_r\right] \cdot 1(R_i = r) + \varepsilon_{it} \tag{12}$$

In this equation, Y_{it} is household i's average daily consumption on the electricity bill for period t, normalized by the control group average post-treatment consumption. This normalization is different for each site, so reducing energy use by two percent in a site with high consumption entails a larger level of kilowatt-hour reduction than a reduction of two percent in a site with low consumption. If we do not normalize in this way, the unconditional dispersion of site-level heterogeneity would appear to be larger.

Because these samples are so large and because these are randomized experiments, the estimated ATEs are similar between this specification and differences-in-differences models with and without household fixed effects. They are also not sensitive to different configurations of control variables such as month-by-year indicators. The coefficients and standard errors are also very similar if we collapse the data over time and use average post-treatment energy usage over all periods as the outcome variable. However, we do not collapse over time because this temporal variation will momentarily be needed to identify interactions of the treatment effect with time-varying factors such as weather. Standard errors are robust and clustered by household to account for serially autocorrelated errors.

To estimate how much site-level heterogeneity is explained by observable differences across sites, we add controls for X to Equation (12):

$$Y_{it} = \alpha X_{it} T_i + \sum_r \left[\mu_r T_i + \beta_r X_{it} + \phi_r Y_{i0} + \pi_r\right] \cdot 1(R_i = r) + \varepsilon_{it} \tag{13}$$

In this equation, the parameters of interest are the unexplained site effects μ_r, which are determined by γZ_r in the notation from our model in Section 2. The α parameters capture how observables X moderate the treatment effect; these are assumed to be constant across sites. The regression includes controls for the main effects of the X variables, which can differ across sites through coefficients β_r, as well as a vector of site-level constants π_r.[13] Notice that controlling for X could either increase or decrease the remaining variation in site effects. Put differently, the variance in τ's from Equation (12) could be larger or smaller than the variance in μ's from (13). Mathematically, if the site-level means of αX and γZ are negatively (positively) correlated, then controlling for αX increases (decreases) the dispersion of the residual site effects.

One other way to test strong external unconfoundedness would be to extrapolate from each of the 14 sites to each of the other 14 sites, controlling for differences in characteristics observed in both sites, and test how well ATEs predicted from each Sample match true ATEs estimated in each Target. However, the α parameters are often imprecisely estimated when using data from only one

[13] Missing X variables are imputed using mean imputation. When a variable is observed at other households within the site, missing values are replaced with the site mean. Otherwise, it is replaced with the mean value across all households in all 14 sites. There are certainly other ways of doing this, but they are unlikely to make much difference and are not central to our argument.

site, making it difficult to control for observable differences across sites. The most precise way to estimate these parameters is to pool data from all 14 sites. Notice that this pooled approach is a "best-case scenario" in terms of explaining variability across sites: typically, the analyst only has data from one site, and the reduced precision of the α's increases the variance of the residual site effects. In earlier drafts, we have also experimented with a number of re-weighting procedures to balance each Sample with each Target on observables, and these procedures do not perform any better.

4.2 Results

Table 3 presents the estimated unconditional ATEs from Equation (12). In sites 2, 3, and 9, the site population was randomly assigned between monthly, bimonthly, and/or quarterly frequencies, while all other sites involved only one frequency.[14] We therefore separately present the ATEs for each frequency. The ATEs vary by a factor of 2.0, from -1.43 percent to -2.84 percent. (Recall that the treatment induces a reduction in energy use, so the ATEs are negative.) While some variation in treatment effects is associated with the frequency of receiving Reports, Table 3 shows that there is still significant variation within frequency across sites.

Table 4 presents the estimates of Equation (13). The specification in Column I includes only the site dummies and site-specific post-treatment dummies as right-hand-side variables. At the bottom of the table, we perform an F test of the joint hypothesis that all site effects μ are equal. The F statistic is 4.24, and the hypothesis is rejected with a p-value of 0.38×10^{-6}. The estimated $\widehat{\mu}$ coefficients are relegated to Online Appendix Table A1.

The standard deviation of the $\widehat{\mu}$ point estimates is 0.48 percent of control group energy use. This variation results both from sampling error and from true underlying variation. From the F-test, we know that this true underlying variation is statistically significant. To estimate the magnitude of this true underlying variation, we first report the sampling standard deviation of the site indicators: the standard deviation of the point estimates that we would expect if there were zero underlying variation in the true μ parameters. The implied true underlying variance in the site indicators is the difference between the variance in the estimated site indicators and the sampling variance. As reported at the bottom of Column I, this is 0.39 percent of energy use.

Column II controls for the frequency with which the Home Energy Reports are delivered, with quarterly frequency as the omitted category. Because the regression includes site dummies, the frequency controls are identified entirely off of the three sites where frequency was randomly assigned within site. The coefficients reported in Table 4 show that monthly treatment frequency causes a 0.54 percent larger ATE than the quarterly frequency, with bimonthly causing an imprecisely-estimated 0.02 percent smaller ATE. Although controlling for treatment frequency decreases the F-statistic because it increases the standard errors on the estimated $\widehat{\mu}$'s, it does not change the

[14] 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.

implied true underlying standard deviation of the μ's. The primary reason for this is that although the ATEs do vary systematically by treatment frequency, site 10 has monthly frequency and also has one of the smallest ATEs. After controlling for frequency, the residual site effect in site 10 is small enough to slightly increase the standard deviation of the set of residual site effects.

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

In Column IV, we control for the interaction of heating and cooling degrees with the treatment effect. The coefficients reported in Table 4 show that one additional average cooling degree increases the treatment effect in absolute value by 0.074 percentage points. Heating degrees also appear to increase the treatment effect, but the coefficient is much smaller and is not statistically different from zero. Although the treatment effect is not statistically larger during colder periods, energy use is of course larger: the $\widehat{\beta}_r$ coefficients, which are omitted to conserve space, show that one additional average heating degree increases energy use by one to four percentage points, depending on the site. At the bottom of Column IV in Table 4, we see that controlling for weather increases the dispersion of the estimated $\widehat{\mu}$'s and also of the implied true underlying variation of the μ's. The primary reason is that sites 10 and 11, which were in relatively hot climates with large average Cooling Degree Days, had relatively small unconditional ATEs. After conditioning on weather, their unexplained residual site effects $\widehat{\mu}$ are even smaller relative to the other sites, can be seen in Online Appendix Table A1.

Of course, it is also possible that the true functional form of the relationship between weather and the treatment effect is not linear. We do not have enough data to estimate this relationship between weather and the treatment effect more flexibly. However, the non-parametric relationship between degree days and electricity use is close to linear over a wide range of degree days. More generally, while all columns of Table 4 could be improved if they reflected the true functional forms, it seems unlikely that the rejection of strong external unconfoundedness would be overturned by using different functional forms.

Column V controls for all time-invariant observable characteristics from Table 2. The coefficients in Table 4 show that the treatment effect is significantly stronger for households with low Baseline Comparison, i.e. households informed that they use more energy than their neighbors. The treatment effect is also stronger in Census tracts with larger mean age, well as for single family homes and houses with electric instead of natural gas heat, pools, and more square footage. These latter three results are consistent with the idea that conservation is less costly when more margins of adjustment are available. Controlling for these variables increases the F-statistic to 9.51 and also increases the implied true underlying standard deviation of the μ parameters to 0.85.

Column VI controls for all X variables from Columns I through V, which further increases the F-statistic and the implied true underlying variation in the μ's. The fact that controlling for observable characteristics increases the unexplained variation in the site effects suggests that observables are negatively correlated with unobservables in this setting.

However, there is an additional reason why controlling for observables could increase the dispersion of $\widehat{\mu}$'s: even after pooling across all experiments, the α parameters may be imprecisely estimated in finite sample, and extrapolating based on an imprecisely-estimated model can worsen the predictions. We therefore include Column VII, which controls only for the X variables that are statistically significantly correlated with the treatment effect with at least 90 percent confidence in Column VI. The implied true underlying dispersion of the site effects and the value of the F-statistic are actually both higher in Column VII than in Column VI, which suggests that imprecisely-estimated α's are not the primary reason why controlling for X increases the dispersion of the site effects $\widehat{\mu}$. Furthermore, Table 4 also shows that the estimated α coefficients are very stable across specifications.

When generalizing a site-specific result in situations like this when there are unobserved site effects, an analyst will typically either formally or informally adopt one of two arguments. First, the analyst might rely on informal theoretical arguments about what are unobservables Z and the magnitude of the difference between γZ_r in Sample and Target. In the Opower setting, the first argument is difficult: because the program is new and unusual, it is still not fully clear what factors Z cause site-level heterogeneity. Second, the analyst might argue that Z is unknown, but the variance of γZ_r across sites is small enough that a site-specific estimate is of general interest. We now examine this second potential argument by asking whether the variance in Opower site effects is economically significant.

4.3 Economic Significance

In an economic sense, how inaccurate is it to assume strong external unconfoundedness and extrapolate results across sites? We consider two measures of economic significance: variation in predicted effects at scale and variation in cost-effectiveness. This is a conservative approach to evaluating economic significance, because as we have seen, conditioning on observables increases the variation in site effects.

Variation in predicted effects at scale is particularly relevant in the case of Opower because policy analyses such as Allcott and Mullainathan (2010) and Davis (2011) have predicted the potential impacts of scaling up the Opower program to utility customers nationwide. If this prediction were done with results from one initial site, how much would the results vary depending on which experiment had been evaluated first? Figure 4 presents the total energy cost savings predicted by multiplying each site's average treatment effect by nationwide annual electricity costs. In this and the subsequent figure, sites are ordered by frequency and then by increasing ATE. As Figure 4 illustrates, the predicted savings would differ by several billion dollars per year depending on which site's ATE is used for the prediction. These values are mechanically connected to the dispersion in the ATEs, so the smallest and largest again differ by a factor of 2.0.

As shown in Figure 5, there is also large variation in cost effectiveness, which we report in cents of program cost per kilowatt-hour of electricity conserved. To calculate this, we divide annual costs, which were provided on a confidential basis by Opower, by the product of the site's ATE and its average electricity consumption per year. The unweighted mean is 3.30 cents per kilowatt-hour. Cost effectiveness varies across sites by a factor of 3.5, from 1.66 to 5.82 cents per kilowatt-hour.

This variation is economically significant in the sense that it can cause program adoption errors: program managers at a Target site might make the wrong decision if they extrapolate cost effectiveness from another site to that Target in order to decide whether to implement the program. Alternative energy conservation programs have been estimated to cost approximately five cents per kilowatt-hour (Arimura, Li, Newell, and Palmer 2011) or between 1.6 and 3.3 cents per kilowatt-hour (Friedrich *et al.* 2009). Whether an Opower program at a new site has cost effectiveness at the lower end (1.66 cents per kilowatt-hour) or upper end (5.82 cents per kilowatt-hour) of the range illustrated in Figure 5 therefore could change whether a manager would or would not want to adopt. Extrapolating cost effectiveness from other Sample sites could lead a Target to implement when it is in fact not cost effective, or fail to implement when it would be cost effective. As a concrete example, we note that in sites 10 and 11, which have two of the smallest ATEs and the worst cost effectiveness, the partner utility has cancelled the programs.

The basic takeaway from this section is that in the Opower context, there is statistically and economically significant heterogeneity in treatment effects across sites, and this heterogeneity is not explained by individually-varying observable characteristics. This is despite the fact that the treatment is highly consistent across sites and we observe a potentially-promising set of observables that could moderate the treatment effect. In this context, the assumption of strong external unconfoundedness does not hold: just as there are often unobservable differences between treatment and control in the absence of a randomized experiment, there are unobservable differences between Sample and Target. Of course, strong external unconfoundedness might be an unrealistically strong assumption for most settings. In the next section, we examine the weaker assumption of external unconfoundedness in distribution.

5 Opower Partner Selection on Observables

A potential response to the failure of strong external unconfoundedness is to replicate a program in multiple sites. If the experimental sites are drawn randomly, then external unconfoundedness in distribution holds. If not, then there is partner selection bias, meaning that even extensive replication does not provide an unbiased estimate of the treatment effect if the program were scaled to the entire population of Target sites.

The ideal way to test external unconfoundedness in distribution would be to estimate Average

Treatment Effects in the set of partner sites and compare the distribution to the distribution of ATEs in non-partner sites. The problem is that by definition, there have been no experiments in non-partner sites, so these ATEs cannot be estimated. Instead, we test for partner selection *on observables*. Specifically, we first gather characteristics of the population of potential Opower partners and estimate a selection equation in the population of partners. We then test whether predicted selection probabilities are correlated with the treatment effect in the set of 14 sites where we have data to estimate treatment effects.

The site-level observed characteristics used to predict partner selection do not vary at the individual level within a site, so it would not be possible to control for them in the typical case when an analyst is extrapolating from one Sample site. Despite being observed at the site level, these characteristics are therefore unobservables in the context of the model in Section 2. We denote these site-level observable characteristics by W.

5.1 Partner Data

We define our population of sites $\mathcal{R}$ to include the 939 electric utilities in the U.S. with more than 10,000 residential customers. There are another 2100 utilities that are smaller, most of which are rural cooperatives or small firms in states with competitive retail electricity markets, but we omit these because Opower has no partners with fewer than 10,000 residential customers. About five percent of utilities operate in multiple states. In order to model how state and local policies affect utilities' decisions, a utility is defined as a separate observation for each state in which it operates.

As detailed in Table 5, we gathered a set of utility-level characteristics that could be correlated with selection into treatment and/or the treatment effect. The first ten are from the Energy Information Administration (EIA) Form 861 for calendar year 2008. These include the utility's ownership structure (Cooperative, private "Investor-Owned Utility" (IOU), Municipal, or Other Government), log of the number of residential consumers, average residential electricity price and usage, reported spending on and energy conserved from energy efficiency programs, and the share of consumers that have voluntarily enrolled in "green pricing programs" that sell renewably-generated energy at a premium price. Next, we include whether the state in which the utility operates has a Renewables Portfolio Standard, which requires utilities to procure a certain proportion of electricity from renewable sources, or an Energy Efficiency Portfolio Standard. Finally, we 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 from the U.S. Census (2010a, 2010b, 2010c).

Table 5 is structured similarly to the commonly-included table that provides suggestive evidence on internal validity by comparing observable characteristics of treatment and control groups. The first column presents the means and standard deviations of each of these characteristics across the 939 utilities in the sample. The second and third columns show the same statistics for Opower's

45 partners and 894 non-partners, respectively. The right-most column tests whether the characteristics are balanced between the two groups. Eleven out of 15 are unbalanced with more than 90 percent confidence, and an F-test easily rejects the hypothesis that the observables are jointly uncorrelated with partner status. Opower's partners clearly differ on site-level observables W_r.

5.2 Empirical Specification

Our test of partner selection on observables has two steps. First, we estimate a probit model of utility selection into partnership with Opower:

$$\Pr(D_r = 1|W_r) = \Phi\left(\rho W_r\right) \tag{14}$$

In this equation, Φ is the CDF of the standard normal distribution, and W_r is a vector of utility-level characteristics. Standard errors are robust and clustered by state.

We then regress the ATEs from the 14 initial sites on the fitted selection probabilities:

$$\widehat{\tau}_r = \theta\widehat{\Pr}(D_r = 1|W_r) + \eta + \varepsilon_r \tag{15}$$

In this regression, we use the $\widehat{\tau}_r$ from the unconditional specification in Column I of Table 4, although the results that follow are comparable when using site effects conditional on different sets of X characteristics. This regression will have 14 observations corresponding to the 14 $\widehat{\tau}_r$'s. We use robust standard errors and cluster by utility. These standard errors are further adjusted to account for the uncertainty in the first-step estimate of $\widehat{\Pr}(D_r = 1|W_r)$, per Murphy and Topel (1985). Because of the small sample, we also report OLS standard errors.

Recall that the treatment reduces electricity demand, so $\widehat{\tau}_r < 0$, and "stronger" ATEs are more negative. If $\theta < 0$, the utilities with higher selection probabilities have more negative ATEs. This implies positive partner selection on observables: utilities that have partnered with Opower have characteristics associated with stronger ATEs compared to non-partners. If $\theta > 0$, this implies negative partner selection on observables: partners have characteristics associated with weaker ATEs than non-partners.

5.3 Results

Table 6 presents the empirical results of the test for partner selection on observables. The top part of the table presents the first-step estimates of Equation (14), while the bottom part presents the second-step estimates of Equation (15). Column I is the specification with the entire set of site-level characteristics. Investor-Owned and Municipality-Owned utilities are statistically significantly more likely to partner with Opower relative to utilities with other ownership structures. Utilities with higher mean energy usage per residential customer are less likely to partner with Opower. Larger utilities and utilities in states with higher median income are more likely to partner.

As reported in the bottom part of Column I, the estimated $\widehat{\theta}$ is 1.26. This means that a ten percentage point increase in selection probability is associated with an ATE that is weaker by 0.126 percent of energy use. The positive $\widehat{\theta}$ implies negative partner selection on observables: utilities whose observable characteristics make them more likely to have partnered with Opower have weaker treatment effects. In this and the other columns, the statistical significance levels are unaffected by which standard errors are used for inference.

Figure 6 illustrates this regression. On the horizontal axis is the fitted selection probability for each of the 14 existing experiments, using the selection equation estimated in Column I of Table 6. On the vertical axis is the ATE. The slope of the best fit line is $\widehat{\theta} = 1.26$.

Of course, the sign and magnitude of selection on observed characteristics will depend on what characteristics are observed. To provide evidence of the robustness of the empirical result of negative selection on observables, we can estimate the selection equation with different sets of observable characteristics. This also allows us to test two potential mechanisms of negative partner selection, both of which were suggested in discussions with Opower's managers.

The first mechanism is factors related to utility size and ownership structure. For example, large investor-owned utilities might be more likely to partner with Opower but are often thought to have less trust from their customers compared to smaller utilities run by government agencies that operate in the "public interest." The Opower Home Energy Reports are co-branded between Opower and the utility, and consumers who distrust the information provider might rationally be less responsive to information. Column II of Table 6 fits the selection probabilities using only the ownership type indicator variables and the log of the number of residential customers. The results show that Investor-Owned Utilities and large utilities are more likely to partner with Opower and tend to have weaker ATEs. The resulting partner selection coefficient $\widehat{\theta}$ does not change significantly from the estimation with all covariates in Column I, suggesting that mechanisms related to size and ownership drive the negative partner selection result.

The second potential mechanism of negative partner selection is "diminishing returns bias," which we introduced earlier in the more general discussion selection mechanisms. In this context, the utilities that were sufficiently innovative and interested in energy efficiency to be one of Opower's early partners might already be running a number of other energy efficiency programs. These previous programs may have already eliminated the low-cost energy conservation opportunities, and the marginal opportunities may be more costly. For example, as discussed in Allcott (2011), a common way that households respond to the Opower treatment is to be more assiduous about turning off lights when not in use. A common energy efficiency program that many Opower partners run is to encourage households to replace standard incandescent lightbulbs with energy efficient Compact Fluorescent Lightbulbs (CFLs). Because CFLs use one-fourth the electricity of an incandescent, a household that has participated in one of these programs and then responds to the Opower treatment by turning off the lights more would save one-fourth the electricity of household that still had incandescents.

To test for diminishing returns bias, Column III of Table 6 fits the selection probability with only the Mean Electricity Usage, Energy Efficiency Spending, and Estimated Energy Conserved variables. While the former variable also largely reflects variation in weather across utilities, the latter two are more direct measures of the extent of previous energy conservation efforts. Mean Electricity Usage and Estimated Energy Conserved are both statistically significantly negatively associated with being an Opower partner. The estimated selection coefficient $\widehat{\theta}$ is positive but very imprecisely estimated.

We implemented four tests of the robustness of negative partner selection on observables. First, we present Column IV of Table 6, which fits the selection equation with the remaining utility-level characteristics not included in Columns II or III. The estimated selection coefficient $\widehat{\theta}$ is again positive but also imprecisely estimated. Second, we repeated the second-step estimation of Equation (15) in Column I 14 additional times, each time leaving out one of the sites. All of the $\widehat{\theta}$ coefficients were still positive and statistically significantly different than zero, and none were statistically different from the $\widehat{\theta}$ in Column I.

Third, we estimate the selection equation an additional 15 times, each time using only one of the W_r variables. These results are presented in Table 7. Columns I and II, respectively, present the correlation of each individual W_r variable with the ATE and the ρ parameter from a univariate estimate of Equation (14). Column III presents the $\widehat{\theta}$ from the second-step estimation of Equation (15). Two of the $\widehat{\theta}$ coefficients are statistically significant and positive: when using either the Investor-Owned Utility indicator variable or the log of the number of residential consumers. Seven other coefficients were positive and not statistically different than zero, and three coefficients were negative and insignificant. Only one, State Median Income, is negative and significant. The coefficients for Renewables Portfolio Standards and Energy Efficiency Portfolio Standards cannot be identified because all of the initial 14 sites are in states with these policies in place.

Our fourth robustness check is to carry out the full selection estimation in Column I of Table 6 15 additional times, each time leaving out one of the W_r variables. These results are in Column IV of Table 7, with each row presenting the $\widehat{\theta}$ when the corresponding variable is omitted. The resulting $\widehat{\theta}$ coefficients are highly robust: all are still positive and statistically significantly different than zero, none are statistically different from the $\widehat{\theta}$ in Column I of Table 6, and the point estimates change very little.

In practical terms, how wrong is it to assume external unconfoundedness in distribution? Imagine a policymaker who wants to know the distribution of ATEs across the 939 potential partner sites in the United States. Under external unconfoundedness in distribution, the policymaker would assume that these first 14 replications are random draws from the population of potential partner sites. The solid black line in Figure 7 presents the resulting predicted distribution of site ATEs; this is just the distribution of the 14 ATEs also illustrated in Figure 6. The mean ATE is -2.02 percent.

The dotted blue line is the predicted distribution of ATEs after adjusting for partner selection on observables. This distribution is calculated by giving each of the Target potential partner sites a random draw of the 14 Sample ATEs and then adjusting for the difference in predicted ATE between the Sample and Target using the estimates of Equation (15) in Column I of Table 4. Because there is negative partner selection on observables, this distribution lies to the left of the unadjusted distribution drawn in black. The mean predicted ATE is now -2.55 percent. Of course, this does not mean that if Opower were to expand nationwide, the mean ATE would be exactly -2.55 percent. This would be the case only if there are no other unobserved factors associated with the initial partner selection mechanism.

The basic empirical result from this section is thus that Opower's partners are negatively selected on observables. This is driven by the fact that large Investor-Owned Utilities are more likely to partner with Opower and, within the first 14 sites, have weaker average treatment effects. If this association holds more generally, an operational takeaway might be that more small publicly-run utilities should consider partnering with Opower. This interpretation is analogous to the way one would interpret meta-analyses that identify factors that moderate treatment effects across sites, such as tests of what school management approaches are associated with larger test score gains.

However, behind this operational takeaway is a deeper point, which is our core argument. In many situations, conditioning on observables is insufficient to solve selection problems, and evidence of selection on observables generates concern about selection on unobservables. In the Opower context, we do not know whether there is positive or negative partner selection on unobservables. Opower is close to a best-case scenario for program evaluation: a nearly-identical treatment evaluated with randomized control trials and replicated 14 times. But unless we can use either econometric analysis or economic theory to understand partner selection on both observables and unobservables, we still do not know what the effects would be if the program were scaled up.

6 Partner Selection in Microfinance

Of course, Opower is only one context. In this section, we test for selection on observables in a second context of broad economic interest: microfinance. Specifically, we examine what types of microfinance institutions (MFIs) partner with major academic organizations to carry out randomized field experiments. Unlike with Opower, we do not have a set of ATEs for the same treatment across different MFIs, as microfinance field experiments have tested a wide array of treatments. Therefore, we do not show, as we had with Opower, that treatment effects are correlated with selection probabilities. Instead, we simply show that MFIs that carry out RCTs differ on observables that could moderate the effects of a variety of interventions.

There are two reasons why microfinance is a convenient area to quantitatively examine partner selection. First, there are many microfinance field experiments with many partners. Second, there is a centralized global database of MFIs that both defines the set of potential partners and contains relevant partner characteristics.

The database we use is called the Microfinance Information Exchange (MIX), which includes information on the characteristics and performance of 1903 MFIs in 115 countries. These MFIs are our population of sites $\mathcal{R}$. We consider characteristics W_r 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 different presentations of loan offer letters as in Bertrand *et al.* (2010), variation in consumer loan interest rates as in Karlan and Zinman (2009), the opportunity to take out a microfinance loan as in Banerjee, Duflo, Glennerster, and Kinnan (2009), or any other intervention.

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 control trials and thus have $D_r = 1$.

Table 8 presents the means and standard deviations of these characteristics by partner status. It is analogous to Table 5, which presents characteristics of Opower partners and non-partners. The first column presents statistics for all MFIs, the second column for partner MFIs, and the third column for non-partner MFIs. The fourth column presents a t-test of the difference in means between partners and non-partners. At the bottom of the fourth column, we report the F-test of a joint regression of partner status on all characteristics. Field experiment partner MFIs clearly differ on site-level observables.

Specifically, we see that for-profit, larger, and older MFIs are substantially more likely to carry out randomized trials. This is quite natural: experiments require stable, well-managed partners and large sample sizes. In some situations, a more established or larger MFI might implement a treatment more or less effectively, or might have more or less trust from borrowers. MFIs with smaller average loan balances are also substantially more likely to be experiment partners, and loan size could affect baseline repayment rates. MFIs with more borrowers per staff member and, relatedly, lower cost per borrower are more likely to be experiment partners. The number of staff per borrower could affect baseline repayment rates through improved monitoring and could also influence the efficacy of interventions that require attention from MFI personnel. Two other correlations are not statistically significant but are suggestive: partner MFIs have less Portfolio at Risk, which corresponds to better 30-day repayment rates, and a larger share of women borrowers. Both of these factors could moderate the effects of a variety of microfinance interventions. We note that these results are not driven by the countries in which RCTs are carried out: they are robust to limiting the population of sites to MFIs in countries where there is at least one partner.

There are two basic takeaways from this section. First, MFIs that partner with academics to carry out randomized field experiments differ from the broader population of MFIs on observables that could moderate the effects of the interventions. Second, Table 8 provides a template for testing partner selection bias. Just as it is common to compare treatment and control group individual-level observables as a suggestive test of unconfoundedness, it is possible to compare Sample and Target site-level observables as a suggestive test of external unconfoundedness in distribution. In the next section, we propose a second statistical test, which provides suggestive evidence on strong external unconfoundedness.

7 F-Test of Sub-Site Heterogeneity

7.1 Overview

In Section 4, we showed that there is significant site-level heterogeneity in the Opower setting. Documenting this required having data from multiple sites. The difficulty of generalizing, of course, is that we do not know the parameter value in the Target. Given data from a Sample site, we must decide whether to assume strong external unconfoundedness, without being able to explicitly test this assumption. In this section, we present a suggestive test of strong external unconfoundedness. As with the common suggestive tests of internal validity, it is an imperfect test of an untestable assumption about unobservables, and we will be clear about the possibilities for Type I and Type II errors.

The test is an F-test for whether the treatment effect varies by "sub-sites" within a site. The intuition is that strong external unconfoundedness requires that observable characteristics capture all heterogeneity between sites. But a "site" is often defined by a particular level of disaggregation of some variable, and within each site there are "sub-sites" defined by more disaggregation of the same variable. For example, Opower programs are implemented in cities that have distinct zip codes. Similarly, educational interventions may be implemented in districts with a number of different schools. The test we propose is based on the idea that unexplained treatment effect heterogeneity at a disaggregated level is suggestive of unexplained heterogeneity at a more aggregated level.

7.2 Procedure

More formally, allow each site to be divided into a mutually exclusive and exhaustive set of sub-sites. Our test is an explicit test of strong external unconfoundedness across sub-sites within the Sample. It is informative in contexts when this predicts the validity of strong external unconfoundedness across sites within the population of sites.

To carry out the test, define a vector of sub-site indicator variables M, leaving one sub-site as the excluded group. Then run the following regression, which interacts M and observable characteristics X with the treatment indicator and controls for lower-order interactions and pre-treatment outcome Y_{0i}:

$$Y_i = [\alpha X_i + \lambda M_i + \mu_0] \cdot T_i + \beta X_i + \pi M_i + Y_{0i} + \pi_0 + \varepsilon_i \tag{16}$$

The F-test of sub-site heterogeneity is simply a test of the joint hypothesis that all λ's are equal to zero. This equation is the simplest implementation of this regression appropriate for the Opower context, and there are other possible versions.

There are two differences between our test and a more generic test of treatment effect heterogeneity. First, we propose testing for a specific type of heterogeneity: heterogeneity conditional on the same variable across which the extrapolation occurs, except at a more disaggregated level. When considering extrapolating from one geographical area to another, we are interested in heterogeneity by geographical sub-area, not by income or any other observed characteristics. Second, this heterogeneity cannot be addressed statistically, even with parametric out-of-sample predictions. For example, when extrapolating treatment effects from a low-income to a high-income population, one could under some assumptions estimate how treatment effects vary by income within the low-income population and project the treatment effect onto the high-income population. By contrast, because our proposed test is of heterogeneity conditional on a set of binary sub-site indicator variables with no overlap, there is no statistical way to project the heterogeneous sub-site effects from the Sample onto the Target.

Of course, because strong external unconfoundedness is untestable in the absence of data from the Target site, the test is only suggestive. Both Type I and Type II errors are possible. An common reason for a false failure to reject is that geographic heterogeneity could occur at a level higher than the sub-site. A treatment with homogeneous effects across sub-sites in Kenya could still have different effects in India. A treatment carefully implemented across many sub-sites by a partner in Tennessee might be poorly implemented by another potential partner in California. Furthermore, if the statistical power of the test is low, perhaps because there are few observations within each sub-site, the test could also falsely fail to reject. Therefore, a failure to reject may not be good evidence that strong external unconfoundedness holds.

However, the converse is more likely to be true: rejecting than $\lambda = 0$ more forcefully suggests that strong external unconfoundedness does not hold. Certainly, false rejections are possible: Sample and Target could both have sub-site heterogeneity, but the distribution of sub-site heterogeneity could be identical in the two sites. For example, if a treatment effect is a function of teacher quality, and two school districts have the same distribution of teacher quality, the test could reject $\lambda = 0$, yet the average treatment effect in each district could be the same. However, rejecting equality puts a burden on the analyst who wants to assume strong external unconfoundedness: the analyst must argue that the distribution of unobserved sub-site effects is somehow the same.

Of course, the common suggestive tests of internal validity similarly generate Type I and Type II errors. Treatment and Control groups could be balanced on exogenous observables but unbalanced on unobservables, and in principle they could also be unbalanced on observables but balanced on unobservables. However, rejecting covariate balance between Treatment and Control places a perceived burden of proof on the analyst who wants to assume unconfoundedness. The overidentification test has false rejections, when all instruments are valid but act on different sets of compliers with different Local Average Treatment Effects, as well as false failures to reject, when all instruments are equally biased. In the Regression Discontinuity context, it is common to test whether control variables are discontinuous around the cutoff (Lee and Lemieux 2009). There could be false failures to reject: even if no observable characteristics are discontinuous at the cutoff, there often may be unobservables that are discontinuous. In principle, there could also be false rejections: there could be discontinuities in observables, which can in principle be controlled for, but no discontinuities in unobservables. However, finding discontinuities in exogenous covariates around the cutoff again places a burden on the analyst who wants to assume continuity of the conditional regression function.

7.3 Sub-Site Heterogeneity in Opower Experiments

We now present the results of the F-test for sub-site heterogeneity in the context of the 14 Opower experimental sites. In separate tests, we define sub-sites at two different levels: Census tract and zip code. In each experiment, we control for the set of observed X variables that vary at the individual household level. Standard errors are robust and clustered by household. In finite sample, the distribution of errors may no longer be normal when a sub-site has very few individual units. We therefore group sub-sites with fewer than 50 households together with the omitted sub-site.

As a visual illustration of the test, Figure 8 presents the distribution of sub-site heterogeneity in Experiment 3, when a "sub-site" is defined to be each of the 86 Census tracts within the site with more than 50 households. In other words, the figure plots the elements of the $\widehat{\lambda}$ vector, which have a standard deviation of 1.9 percent. Of course, this figure is illustrative only: it is not a statistical test of whether $\lambda = 0$. This is because the distribution of point estimates in $\widehat{\lambda}$ depends on both the true underlying distribution and sampling error.

Table 9 presents the statistical results of the F-tests of sub-site heterogeneity for each of the 14 Opower sites. The first pair of columns presents results for sub-sites defined by zip codes, and the second pair of columns presents results for Census tracts. At each site, the results tend to be roughly consistent between the two levels of geographical aggregation. At sites 5 and 6, the F-tests reject that the λ's are equal with greater than 90 percent confidence at both levels of aggregation. At sites 3 and 13, the F-tests reject equality of Census tract-level effects and nearly reject (with greater than 85 percent confidence) equality of zip code-level effects. At site 10, the F-tests reject equality of the zip code-level effects and nearly reject (with 85 percent confidence) equality of the Census tract-level effects. At site 14, the test rejects equality at the tract level but not at the zip code level. At all other sites, the tests fail to reject equality at both levels.

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

8 Conclusion

While external validity has long been a fundamental concern to empiricists in economics and other fields, there have been few opportunities to quantitatively assess the ability to generalize parameter estimates from the same treatment across different settings. This paper analyzes a remarkable series of nearly-identical energy conservation field experiments run by Opower in a number of different sites across the U.S. We document statistically and economically significant heterogeneity in treatment effects across sites that cannot be explained by observed covariates. 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, implying negative selection on observables. This suggests that replicating the Opower program with additional partners has not given an unbiased estimator of the potential effects of the program in non-partner sites.

While our quantitative results are specific to this set of energy conservation programs, we think that the concept of partner selection bias may be general to empirical analyses in other settings. As an additional example, we document suggestive evidence in the context of microfinance. We show that MFIs that partner with academics to carry out RCTs differ on observables that could be correlated with the effects of different interventions, suggesting that we should continue to exercise caution in generalizing results from partner MFIs to non-partner MFIs.

Our analysis suggests a specific set of steps that could be taken regularly when generalizing empirical results. First, the analyst would clearly define the Target site or population of interest. Second, the analyst would formally compare the observable characteristics of the population, the partner, the intervention, and the economic environment between Sample and Target. Part of this comparison might resemble our Tables 5 and 8, which are analogous to tables commonly used to compare observable characteristics of Treatment and Control groups in a suggestive test of internal validity. Third, the analyst could use theory to understand potential unobserved differences between Sample and Target sites that could moderate the treatment effects. Fourth, if appropriate in a given setting, the analyst could use formal statistical tests that provide suggestive evidence on external validity, including our F-test of sub-site heterogeneity, the untreated outcomes test of Hotz, Imbens, and Mortimer (2005) and Hotz, Imbens, and Klerman (2006), and the conditional LATE test of Angrist and Fernandez-Val (2010).

At the project design stage, there are a wider array of options to improve external validity. First, when treatment effects are difficult to generalize, these results suggest the importance of RCTs with representative samples of the Target population of policy interest. For example, each of the Opower experiments is at a site where the ATE is of some interest *per se*, as the partner utility decides whether to continue running the program. If the effects of large-scale social programs need to be measured accurately, it might be especially important for policymakers to implement the program itself - not just a pilot in a particular location - using an RCT. Second, researchers can attempt to replicate similar field experiments in multiple locations. Importantly, these locations would ideally be run at sites with attributes, including partner characteristics, that lie in different parts of the distribution of factors that moderate the treatment effect. Third, some have argued that theoretically-motivated "mechanisms" are often more generalizable than average treatment effects of specific programs (Deaton 2010a, Deaton 2010b, Ludwig, Kling, and Mullainathan 2011). If this is the case, researchers can focus on identifying mechanisms upon which policy decisions hinge and designing empirical studies to tease them out.

In conclusion, we re-emphasize that the empirical results from our context simply cannot be used to argue against using randomized control trials. To the contrary, the Opower experiments are excellent examples of the importance of RCTs. Allcott (2011) shows that non-experimental estimates perform extremely poorly in the Opower context. Indeed, within the set of 14 initial sites, internally-valid estimators from other Sample sites predict true ATEs at a Target site far better than non-experimental estimators from the same Target. Allcott and Mullainathan (2010) and Allcott and Greenstone (2012) highlight Opower as an example of the importance of using randomized control trials to evaluate energy conservation programs. Discussion of external validity certainly does not diminish the importance of internal validity.

References

[1] Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas Kane, and Parag Pathak (2009). "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots." NBER Working Paper No. 15549 (November).

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

[3] Allcott, Hunt (2011). "Social Norms and Energy Conservation." *Journal of Public Economics*, Vol. 95, No. 9-10 (October), pages 1082-1095.

[4] Allcott, Hunt and Michael Greenstone (2012). "Is There an Energy Efficiency Gap?" *Journal of Economic Perspectives*, Vol. 26, No. 1 (Winter), pages 3-28.

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

[6] Angrist, Joshua (2004). "Treatment Effect Heterogeneity in Theory and Practice." *The Economic Journal*, Vol. 114, No. 494 (March), pages C52-C83.

[7] Angrist, Joshua, and Ivan Fernandez-Val (2010). "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework." NBER Working Paper No. 16566 (December).

[8] 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.

[9] Angrist, Joshua, Parag Pathak, and Christopher Walters (2011). "Explaining Charter School Effectiveness." NBER Working Paper No. 17332 (August).

[10] 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.

[11] Arimura, Toshi, Shanjun Li, Richard Newell, and Karen Palmer (2011). "Cost-Effectiveness of Electricity Energy Efficiency Programs." Resources for the Future Discussion Paper 09-48 (May).

[12] 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).

[13] Banerjee, Abhijit (2009). "Big Answers for Big Questions." In Cohen, Jessica, and William Easterly (Eds.), What Works in Development? Thinking Big and Thinking Small. Washington, DC: Brookings Institution Press.

[14] 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.

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

[16] Belot, Michele, and Jonathan James (2012). "Selection into Policy Relevant Field Experiments." Working Paper, Oxford University (June).

[17] 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.

[18] Bertrand, Marianne, Dean Karlan, Sendhil Mullainathan, Eldar Shafir, and Jonathan Zinman (2010). "What's Advertising Content Worth? Evidence from a Consumer Credit Marketing Field Experiment." *Quarterly Journal of Economics*, Vol. 125, No. 1 (February), pages 263-306

[19] 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.

[20] 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.

[21] Campbell, Donald (1957). "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin*, Vol. 54, No. 4 (July), pages 297-312.

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

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

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

[25] 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.

[26] Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora (2012). "Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment." Working Paper, Centre de Recherche en Economie et Statistique (June).

[27] 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).

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

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

[30] Deaton, Angus (2010b). "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives*, Vol. 24, No. 3 (Summer), pages 3-16.

[31] 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.

[32] Dehejia, Rajeev, and Sadek Wahba (1999). "Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs." *Journal of the American Statistical Association*, Vol. 94, pages 1053–1062.

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

[34] 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).

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

[36] 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).

[37] Heckman, James (1979). "Sample Selection Bias as a Specification Error." *Econometrica*, Vol. 47, No. 1 (January), pages 153-161.

[38] Heckman, James (1992). “Randomization and social policy evaluation”. In Charles Manski and Irwin Garfinkel (Eds.), Evaluating Welfare and Training Programs. Harvard Univ. Press: Cambridge, MA, pages 201-230.

[39] Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd (1998). "Characterizing Selection Bias Using Experimental Data." *Econometrica*, Vol. 66, No. 5 (September), pages 1017-1098.

[40] Heckman, James, Hidehiko Ichimura, and Petra Todd (1997). "Matching as an Econometric Evaluation Estimator: Evidence from Evaluating a Job Training Program." *Review of Economic Studies*, Vol. 64, No. 4, (October), pages 605-654.

[41] Heckman, James, Robert Lalonde, and Jeffrey Smith (1999). "The Economics and Econometrics of Active Labor Market Programs." In Orley Ashenfelter and David Card (Eds.) Handbook of Labor Economics, Chapter 31, pages 1865-2097.

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

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

[44] Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). "Understanding Instrumental Variables in Models with Essential Heterogeneity." The Review of Economics and Statistics, Vol. 88, No. 3 (August), pages 389-432.

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

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

[47] Heckman, James, and Edward Vytlacil (2007a). ""Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In James Heckman and Edward Leamer (Eds), Handbook of Econometrics, Vol. 6B. Amsterdam: Elsevier, pages 4779-4874.

[48] Heckman, James, and Edward Vytlacil (2007b). "Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast their Effects in New Environments." In James Heckman and Edward Leamer (Eds), Handbook of Econometrics, Vol. 6B. Amsterdam: Elsevier, pages 4875-5144.

[49] Hotz, Joseph (1992). "Designing Experimental Evaluations of Social Programs: The Case of the U.S. National JTPA Study." University of Chicago Harris School of Public Policy Working Paper 9203 (January).

[50] Hotz, Joseph, Guido Imbens, and Jacob Klerman (2006). “Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program.” *Journal of Labor Economics*, Vol. 24, No. 3, pages 521–66.

[51] Hotz, Joseph, Guido Imbens, and Julie Mortimer (2005). "Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations." *Journal of Econometrics*, Vol. 125, No 1-2, pages 241-270.

[52] Imbens, Guido (2010). "Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009)." *Journal of Economic Literature*, Vol. 48, No. 2 (June), pages 399-423.

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

[54] 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).

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

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

[57] Levitt, Steven D. and John A. List (2009). “Field Experiments in Economics: The Past, the Present, and the Future." *European Economic Review*, Vol. 53, No. 1 (January), pages 1-18.

[58] Ludwig, Jens, Jeffrey Kling, and Sendhil Mullainathan (2011). "Mechanism Experiments and Policy Evaluations." *Journal of Economic Perspectives*, Vol. 25, No. 3 (Summer), pages 17-38.

[59] Manning, Willard, Joseph Newhouse, Naihua Duan, Emmett Keeler, Bernadette Benjamin, Arleen Leibowitz, Susan Marquis, and Jack Zwanziger (1988). "Health Insurance and the Demand for Medical Care." Santa Monica, California: The RAND Corporation.

[60] Manski, Charles (2011). "Policy Analysis with Incredible Certitude." *The Economic Journal*, Vol. 121, No. 554 (August), pages F261–F289.

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

[62] 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.

[63] Murphy, Kevin M., and Robert Topel (1985). "Estimation and Inference in Two-Step Econometric Models." *Journal of Business and Economic Statistics*, Vol. 3, No. 4 (October), pages 370-379.

[64] 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.

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

[66] 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.

[67] Rosenbaum, Paul, and Donald Rubin (1983). “The Central Role of the Propensity Score in Observational Studies for Causal Effects.” *Biometrika*, Vol. 70, No. 1, pages 41–55.

[68] 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.

[69] 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.

[70] Sanbonmatsu, Lisa, Jens Ludwig, Lawrence Katz, Lisa Gennetian, Greg Duncan, Ronald Kessler, Emma Adam, Thomas McDade, and Stacy Tessler Lindau (2011). "Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation." Available from http://isites.harvard.edu/fs/docs/icb.topic964076.files/mto_final_exec_summary.pdf

[71] 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.

[72] Smith, Jeffrey, and Petra Todd (2004). "Does Matching Address LaLonde’s Critique of Nonexperimental Estimators?" *Journal of Econometrics*, Vol 125 pages 305-353.

[73] Stuart, Elizabeth, Stephen Cole, Catherine Bradshaw, and Philip Leaf (2011). "The Use of Propensity Scores to Assess the Generalizability of Results from Randomized Trials." *Journal of the Royal Statistical Society*, Vol. 174, Part 2, pages 369-386.

[74] 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.

[75] 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

[76] 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

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

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

Tables

Table 1: Overview of Opower Experiments

Site *Number*	*Region*	*Start Date*	*Households*	*Treated* *Households*	*Observations*
1	Urban Midwest	July, 2009	36,603	18,790	1,264,375
2	Urban Midwest	July, 2009	54,475	28,027	1,873,482
3	Rural Midwest	January, 2009	78,129	39,024	3,421,306
4	Suburban Mountain	October, 2009	11,593	7,254	394,525
5	Suburban Mountain	October, 2009	27,117	16,947	914,344
6	West Coast	October, 2009	33,506	23,906	570,386
7	Rural Midwest	April, 2009	17,728	9,861	794,457
8	Urban Northeast	September, 2009	49,522	24,808	1,712,530
9	West Coast	October, 2008	79,017	34,893	3,121,879
10	West Coast	January, 2009	25,150	5,570	985,148
11	West Coast	January, 2009	17,669	3,852	672,629
12	West Coast	September, 2009	39,334	19,663	671,990
13	West Coast	March, 2008	59,664	24,761	1,809,427
14	West Coast	April, 2008	24,291	9,903	735,663
Combined		March, 2008	553,798	267,259	18,942,141

Table 2: Household-Level Observed Characteristics

	Weather		**Energy Use**	**Census Tract**			**House**						
Site Number	*Heating Degrees*	*Cooling Degrees*	*Baseline Comparison*	*Mean Age*	*Median Income ($000s)*	*Percent Caucasian*	*Electric Heat*	*House Age*	*Value ($000s)*	*Has Pool*	*Rental House*	*Single Family*	*Square Footage (000s)*
1	13.6	2.7	-0.47	49.4	68.2	0.76	-	-	-	-	0.1	0.75	-
	(13.4)	(3.9)	(1.88)	(5.3)	(30.3)	(0.21)	-	(20.6)	-	-	(0.2)	(0.32)	(1.27)
2	13.6	2.7	-0.47	49.4	68.2	0.76	-	-	-	-	0.1	0.75	-
	(13.4)	(3.9)	(1.17)	(5.7)	(25.9)	(0.3)	-	-	-	-	(0.27)	(0.44)	-
3	16.6	1.7	-0.05	45.1	62.2	0.96	-	31.6	393.9	-	-	-	1.66
	(16.4)	(2.5)	(1.41)	(2.4)	(9.4)	(0.02)	-	(28)	(139.4)	-	-	-	(0.45)
4	17.1	1.7	-1.08	43.9	56.2	0.91	0.2	23.4	-	-	0.23	0.87	2.25
	(14.8)	(2.7)	(1.55)	(4.2)	(16.5)	(0.04)	(0.35)	(15.3)	-	-	(0.42)	(0.34)	(0.76)
5	17.5	1.7	0.3	43.3	50.5	0.9	0.11	26.5	-	-	0.36	0.7	1.91
	(14.8)	(2.7)	(0.93)	(4.4)	(16.6)	(0.05)	(0.22)	(17.9)	-	-	(0.48)	(0.46)	(0.53)
6	4.1	2.3	-0.3	50.3	49.5	0.66	-	59.2	-	0.1	0.35	0.5	1.69
	(3.7)	(2.6)	(1.23)	(3)	(21.6)	(0.1)	-	(16.7)	-	(0.31)	(0.4)	(0.5)	(0.52)
7	18.8	0.7	0.09	52.9	38.8	0.95	0.31	-	-	-	0.05	-	-
	(17.7)	(1.1)	(1.11)	(2.1)	(6.5)	(0.07)	(0.46)	-	-	-	(0.21)	-	-
8	13.6	2.4	-0.28	51	65.9	0.93	-	58.7	-	0.02	0.06	-	2.03
	(12.4)	(3.6)	(1.26)	(2.7)	(23)	(0.1)	-	(36.5)	-	(0.15)	(0.22)	-	(0.72)
9	13.3	0.5	-0.26	47.2	71.5	0.8	0.07	31.2	361.1	-	0.03	-	2.2
	(8.4)	(1.2)	(1.18)	(3.5)	(19.8)	(0.09)	(0.25)	(15.6)	(175.8)	-	(0.16)	-	(0.67)
10	2.7	10	0.05	55.3	43.8	0.79	-	27.1	-	0.37	0.01	1	1.95
	(4.1)	(10.3)	(1.35)	(8)	(12.8)	(0.14)	-	(15.3)	-	(0.48)	(0.08)	-	(0.67)
11	2.8	10.3	0.27	57	42.1	0.81	-	28.6	-	0.06	-	0.08	1.74
	(4.2)	(10.5)	(1.24)	(7.5)	(16.6)	(0.15)	-	(6.6)	-	(0.23)	-	(0.27)	(0.51)
12	12.5	0.4	-0.41	48.1	54.1	0.72	0.17	65.1	436.6	-	0.06	-	1.83
	(6.2)	(0.6)	(1.63)	(3.5)	(13.8)	(0.22)	(0.38)	(25.4)	(293.2)	-	(0.22)	-	(0.77)
13	5.9	2.9	-	49.7	60.7	0.77	0.32	34.1	229.3	0.28	0.01	-	1.84
	(6.6)	(3.1)	-	(4.4)	(14)	(0.15)	(0.47)	(16.4)	(148.7)	(0.45)	(0.1)	-	(0.6)
14	6	3	-	49.9	56.9	0.77	0.12	43.1	174.4	0.04	0.01	-	1.49
	(6.6)	(3.2)	-	(4.2)	(12.7)	(0.15)	(0.32)	(20.9)	(116.7)	(0.19)	(0.09)	-	(0.43)

Notes: This table presents the site-level means of observed characteristics, with standard deviations in parenthesis. Heating and cooling degrees are observed at the household-by-month level; all other variables are at the household level. Dashes represent cases when a variable is not observed at a site.

Table 3: Average Treatment Effects

		Frequency	
Site	*Monthly*	*BiMonthly*	*Quarterly*
1	-1.94 (0.2)		
2		-1.43 (0.2)	-1.45 (0.19)
3	-2.65 (0.15)		-2.17 (0.17)
4		-2.55 (0.41)	
5			-1.79 (0.32)
6		-2.56 (0.22)	
7		-2.56 (0.34)	
8		-1.73 (0.14)	
9	-2.09 (0.14)		-1.47 (0.2)
10	-1.62 (0.32)		
11			-1.47 (0.48)
12		-1.88 (0.2)	
13	-2.84 (0.16)		
14			-1.49 (0.32)
Mean	-2.23	-2.12	-1.64

Notes: This table presents the Average Treatment Effect for each site as a percent of control group post-period usage. Standard errors are in parenthesis. The columns group the experiments by monthly, bi-monthly, and quarterly frequency of treatment.

Table 4: Tests for Site Effects

Interactions with T_i	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*
Monthly Frequency		-0.54				-0.52	-0.55
		(0.14)***				(0.14)***	(0.14)***
Bi-Monthly Frequency		0.02				-0.01	
		(0.22)				(0.21)	
Duration<6 Months			0.52			0.50	0.47
			(0.13)***			(0.15)***	(0.14)***
Cooling Degrees				-0.074		-0.08	-0.060
				(0.030)**		(0.03)**	(0.031)*
Heating Degrees				-0.005		-0.01	
				(0.008)		(0.01)	
Baseline Comparison					1.25	1.26	1.27
					(0.13)***	(0.13)***	(0.13)***
Tract Mean Age					-0.027	-0.027	-0.034
					(0.015)*	(0.015)*	(0.015)**
Tract Median Income					-0.002	-0.002	
					(0.004)	(0.004)	
Tract Percent Caucasian					-0.10	-0.11	
					(0.50)	(0.50)	
Electric Heat House					-0.70	-0.71	-0.67
					(0.28)**	(0.28)**	(0.28)**
House Age					-0.001	-0.001	
					(0.003)	(0.003)	
House Value					-0.001	-0.001	
					(0.001)	(0.001)	
House Has Pool					-1.38	-1.38	-1.41
					(0.30)***	(0.30)***	(0.30)***
Rental House					0.20	0.19	
					(0.33)	(0.33)	
Single Family House					-1.04	-1.05	-1.14
					(0.30)***	(0.30)***	(0.27)***
House Square Footage					-0.32	-0.31	-0.44
					(0.13)**	(0.13)**	(0.10)***
X_it Main Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes
N (millions)	8.01	8.01	8.01	8.01	8.01	0.00	0.00
R^2	0.53	0.53	0.54	0.62	0.63	0.63	0.61
F Stat (Regression)	16085	15338	13897	18616	7950	7551	8610
Site Effects Tests							
F Stat (Site Indicators)	4.25	3.65	4.13	4.55	9.51	9.79	10.35
F-test p-Value (x10^{-6})	0.38	0.00	0.00	0.00	0.00	0.00	0.00
SD (Estimated Site Indicators)	0.48	0.50	0.48	0.62	0.91	1.02	1.05
Sampling SD (Site Indicators)	0.28	0.31	0.29	0.30	0.33	0.35	0.32
Implied True SD (Site Indicators)	0.39	0.39	0.39	0.54	0.85	0.96	1.00

Notes: This table presents tests of whether site indicator variables estimated in Equation (13) are different from each other. Columns I through VII condition on different sets of observables. The estimated $\widehat{\mu}$ coefficients are presented in Online Appendix Table A1. Robust standard errors, clustered by household, are in parenthesis. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 5: Opower Partner Characteristics

	All	*Partners*	*Non-Partners*	*Difference*
Investor-Owned Utility	0.18	0.62	0.16	0.46
	(0.39)	(0.49)	(0.37)	(0.07)***
Municipality-Owned Utility	0.25	0.27	0.25	0.02
	(0.43)	(0.45)	(0.43)	(0.07)
Other Government-Owned Utility	0.04	0.02	0.04	-0.02
	(0.19)	(0.15)	(0.20)	(0.02)
Cooperative-Owned Utility	0.48	0.07	0.50	-0.43
	(0.50)	(0.25)	(0.50)	(0.04)***
log(Residential Customers)	3.66	5.69	3.56	2.13
	(1.26)	(1.63)	(1.15)	(0.24)***
Residential Electricity Price (cents/kWh)	10.59	12.21	10.51	1.70
	(3.20)	(4.08)	(3.13)	(0.61)***
Mean Electricity Usage (MWh/year)	12.4	9.4	12.6	-3.1
	(3.4)	(2.4)	(3.4)	(0.4)***
Energy Efficiency Spending ($/customer-year)	15.4	25.1	14.9	10.2
	(138.2)	(29.7)	(141.4)	(6.4)
Estimated Energy Conserved (kWh/customer-year)	25.3	48.3	24.1	24.2
	(178.1)	(69.9)	(181.8)	(12.0)**
Green Energy Market Share (Percent)	0.72	1.58	0.67	0.90
	(4.98)	(3.89)	(5.02)	(0.60)
State Has Renewables Portfolio Standard	0.52	0.84	0.50	0.34
	(0.50)	(0.37)	(0.50)	(0.06)***
State Has Energy Efficiency Portfolio Standard	0.6	0.9	0.5	0.4
	(0.5)	(0.3)	(0.5)	(0.0)***
State Percent Democrat Voters	49.1	55.5	48.8	6.7
	(9.4)	(8.6)	(9.3)	(1.3)***
State Median Income ($000s)	49.3	55.9	49.0	6.9
	(6.8)	(5.5)	(6.7)	(0.8)***
State Percent College Graduates	26.0	29.5	25.8	3.7
	(4.4)	(3.8)	(4.4)	(0.6)***
N	939	45	894	
F Test p-Value				0.000 ***

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

Table 6: Opower Partner Selection

	I	*II*	*III*	*IV*
Probit Selection Equation				
Investor-Owned Utility	Yes	Yes		
Municipality-Owned Utility	Yes	Yes		
Other Government-Owned Utility	Yes	Yes		
Cooperative-Owned Utility	Yes	Yes		
log(Residential Customers)	Yes	Yes		
Residential Electricity Price (cents/kWh)	Yes			Yes
Mean Electricity Usage (MWh/year)	Yes		Yes	
Energy Efficiency Spend ($/cust.-yr)	Yes		Yes	
Energy Conserved (kWh/cust.-yr)	Yes		Yes	
Green Energy Market Share (Percent)	Yes			Yes
State Renewables Portfolio Standard	Yes			Yes
State Energy Efficiency Portfolio Standard	Yes			Yes
State Percent Democrat Voters	Yes			Yes
State Median Income ($000s)	Yes			Yes
State Percent College Graduates	Yes			Yes
N	938	939	939	938
OLS Regression of $\widehat{\tau}$ on $\widehat{Pr}(T=1)$				
$\widehat{\theta}$ Coefficient	1.26	1.46	2.54	0.73
Murphy-Topel (1985) Standard Error	(0.25)***	(0.34)***	(4.83)	(3.33)
Robust Standard Error, Clustered by Utility	(0.11)***	(0.16)***	(4.68)	(3.30)
OLS Standard Error	(0.37)***	(0.44)***	(3.92)	(3.67)

Notes: The top section of this table presents the results of estimating the Opower partner selection function from Equation (14). Robust standard errors, clustered by state, are in parenthesis. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively. The bottom section of this table presents the estimation results from Equation (15). The "Murphy-Topel (1985) Standard Error" is robust, clustered by utility, with the Murphy-Topel (1985) adjustment.

Table 7: Additional Partner Selection Results

	Univariate	*Univariate*	*Univariate*	$\widehat{\theta}$ *Coefficient*
	Correlation	$\widehat{\rho}$ *Selection*	$\widehat{\theta}$	*when*
	with ATE	*Coefficient*	*Coefficient*	*Omitted*
	I	*II*	*III*	*IV*
Probit Selection Equation				
Investor-Owned Utility	0.58	1.04	4.07	1.26
	(0.11)***	(0.19)***	(2.30)*	(0.26)***
Municipality-Owned Utility	-0.25	0.05	-49.16	1.32
	(0.20)	(0.20)	(793.85)	(0.29)***
Other Government-Owned Utility	-0.17	-0.27	7.97	1.24
	(0.14)	(0.38)	(56.09)	(0.25)***
Cooperative-Owned Utility	-0.57	-1.11	7.21	1.28
	(0.13)***	(0.24)***	(8.82)	(0.27)***
log(Residential Customers)	0.18	0.46	1.49	2.71
	(0.033)***	(0.070)***	(0.37)***	(1.19)**
Residential Electricity Price (cents/kWh)	0.025	0.055	4.15	1.24
	(0.036)	(0.021)***	(8.75)	(0.24)***
Mean Electricity Usage (MWh/year)	-0.062	-0.134	2.29	1.24
	(0.092)	(0.027)***	(4.87)	(0.25)***
Energy Efficiency Spending ($/customer-year)	0.0059	0.0002	296	1.26
	(0.0052)	(0.0002)	(1460)	(0.25)***
Estimated Energy Conserved (kWh/customer-year)	0.0018	0.0003	67.7	1.26
	(0.0016)	(0.0002)	(177.7)	(0.25)***
Green Energy Market Share (Percent)	-0.032	0.011	-27.7	1.24
	(0.027)	(0.008)	(83.1)	(0.25)***
State Has Renewables Portfolio Standard		0.74		1.26
		(0.23)***		(0.26)***
State Has Energy Efficiency Portfolio Standard		1.03		1.27
		(0.25)***		(0.25)***
State Percent Democrat Voters	0.013	0.030	1.86	1.30
	(0.004)***	(0.011)***	(1.36)	(0.26)***
State Median Income ($000s)	-0.054	0.066	-4.51	1.23
	(0.016)***	(0.012)***	(2.58)*	(0.25)***
State Percent College Graduates	-0.007	0.081	-0.05	1.28
	(0.039)	(0.018)***	(2.33)	(0.25)

Notes: Column I presents the coefficient of a regression of the ATE τ_r on each W_r variable. Column II presents the $\widehat{\rho}$ coefficient from estimating the Opower partner selection function from Equation (14) using each W_r variable individually. Column III presents the $\widehat{\theta}$ from the selection estimation using each W_r variable individually. Column IV presents the $\widehat{\theta}$ from the selection estimation from Equation (15) leaving out each W_r variable individually. Robust standard errors, clustered by state, are in parenthesis. Columns III and IV include robust standard errors, clustered by utility, with the Murphy-Topel (1985) adjustment. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Table 8: MFI Partner Characteristics

	All	*Partners*	*Non-Partners*	*Difference*
Average Loan Balance ($000's)	1.42	0.58	1.44	-0.86
	(3.07)	(0.51)	(3.10)	(0.12)***
Borrowers/Staff Ratio (10^3)	0.13	0.22	0.13	0.09
	(0.21)	(0.19)	(0.21)	(0.03)***
Cost per Borrower ($000's)	0.18	0.10	0.18	-0.08
	(0.19)	(0.08)	(0.19)	(0.01)***
MFI Age (Years)	13.99	21.86	13.84	8.02
	(10.43)	(11.21)	(10.36)	(1.88)***
Non-Profit	0.63	0.37	0.64	-0.27
	(0.48)	(0.49)	(0.48)	(0.08)***
Number of Borrowers (10^6)	0.06	0.85	0.05	0.80
	(0.40)	(1.84)	(0.27)	(0.31)***
Percent Portfolio at Risk	0.083	0.068	0.083	-0.015
	(0.120)	(0.066)	(0.121)	(0.012)
Percent Women Borrowers	0.62	0.69	0.62	0.07
	(0.27)	(0.27)	(0.27)	(0.05)
N	1903	35	1868	
F Test p-Value				0.000 ***

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

Table 9: Results of F-Tests of Sub-Site Heterogeneity

	Zip Code		**Census Tract**	
Site Number	*Degrees of Freedom*	*F-Test p-Value*	*Degrees of Freedom*	*F-Test p-Value*
1	40	0.279	176	0.948
2	40	0.557	267	0.358
3	31	0.104	85	0.025 **
4	4	0.903	33	0.655
5	4	0.079 *	36	0.020 **
6	9	0.017 **	36	0.000 ***
7	56	0.540	35	0.905
8	25	0.098 *	95	0.229
9	40	0.928	178	0.720
10	4	0.010 ***	33	0.149
11	3	0.264	30	0.360
12	20	0.364	120	0.754
13	28	0.132	84	0.000 ***
14	27	0.451	79	0.030 **
Mean	24	0.338	92	0.3682

Notes: This table shows the results of the F-test of sub-site heterogeneity for each of the 14 Opower sites with sub-sites defined as Zip Codes and Census Tracts. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.

Figures

Figure 1: Home Energy Reports: Social Comparison Module

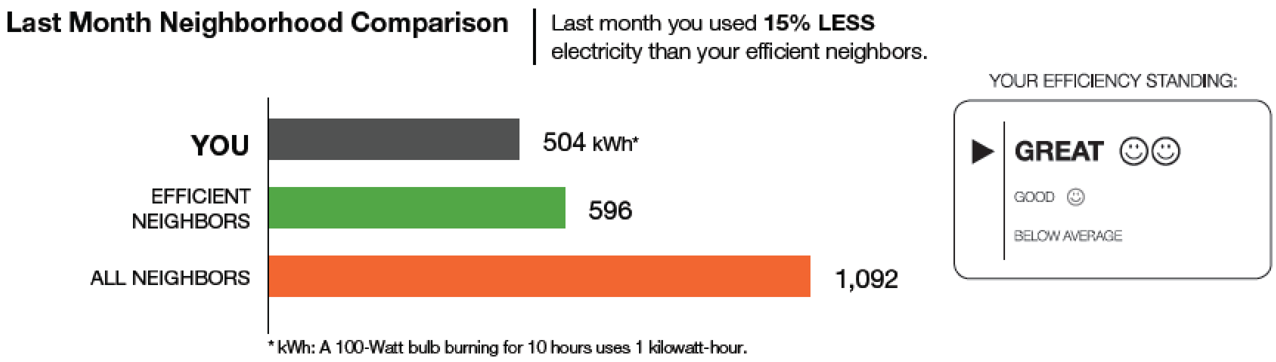
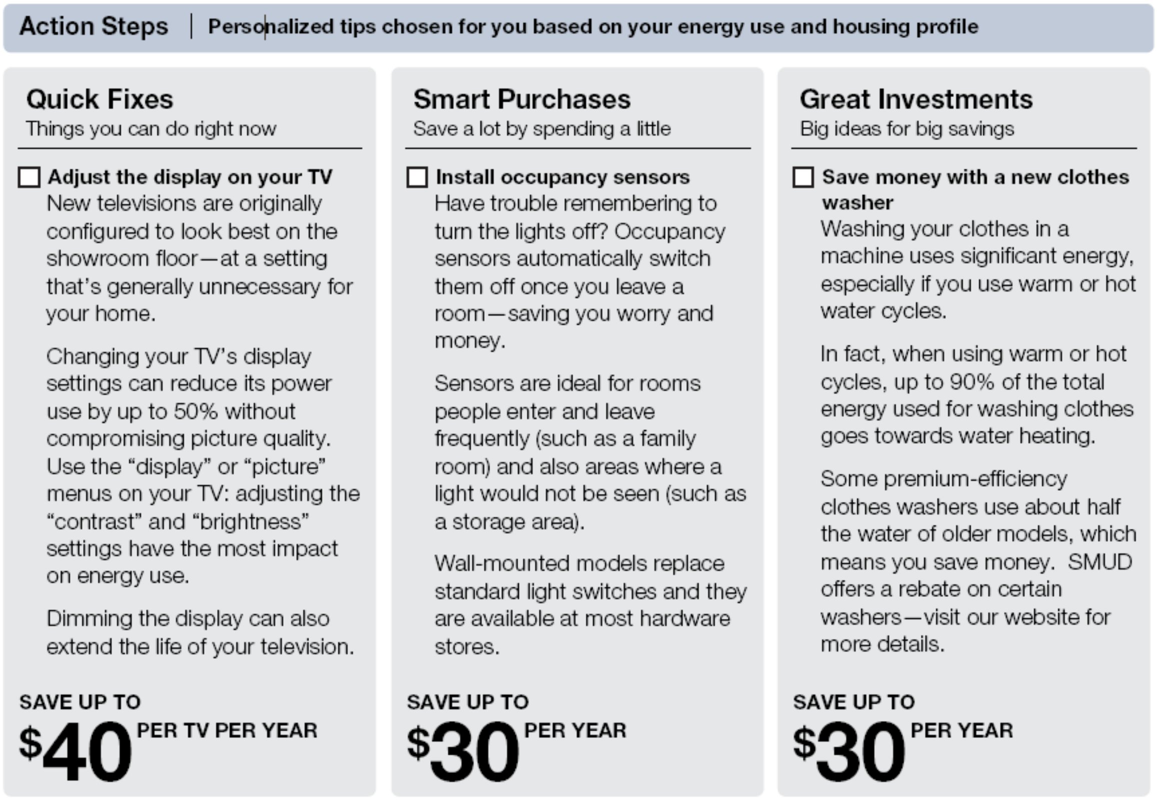


Figure 2: Home Energy Reports: Action Steps Module



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

Quick Fixes
Things you can do right now

□ **Adjust the display on your TV**

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

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

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

SAVE UP TO
$40 PER TV PER YEAR

Smart Purchases
Save a lot by spending a little

□ **Install occupancy sensors**

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

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

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

SAVE UP TO
$30 PER YEAR

Great Investments
Big ideas for big savings

□ **Save money with a new clothes washer**

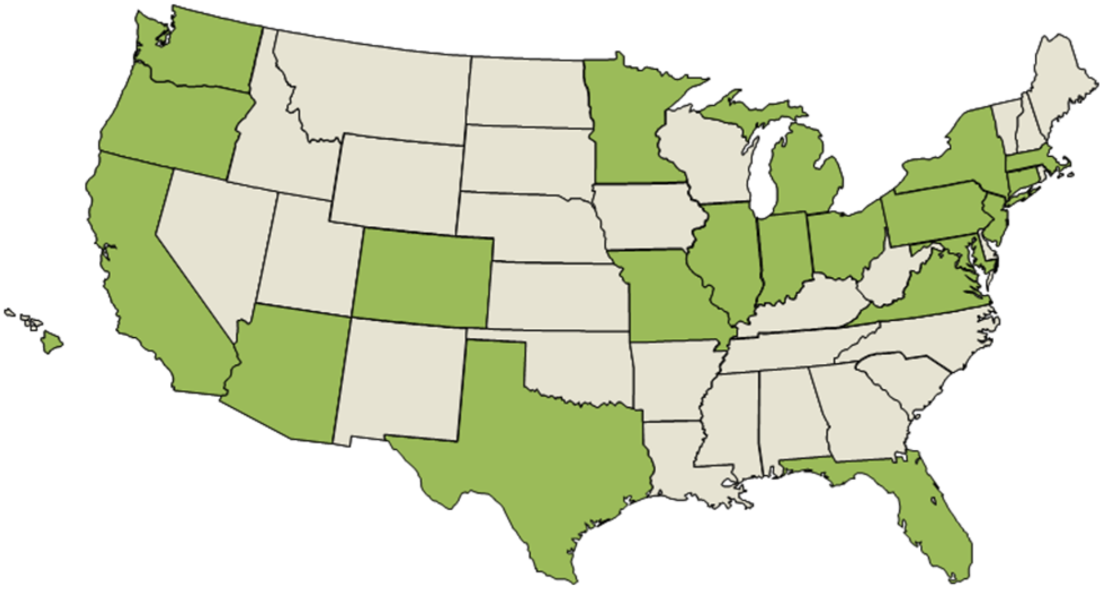
Washing your clothes in a machine uses significant energy, especially if you use warm or hot water cycles.

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

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

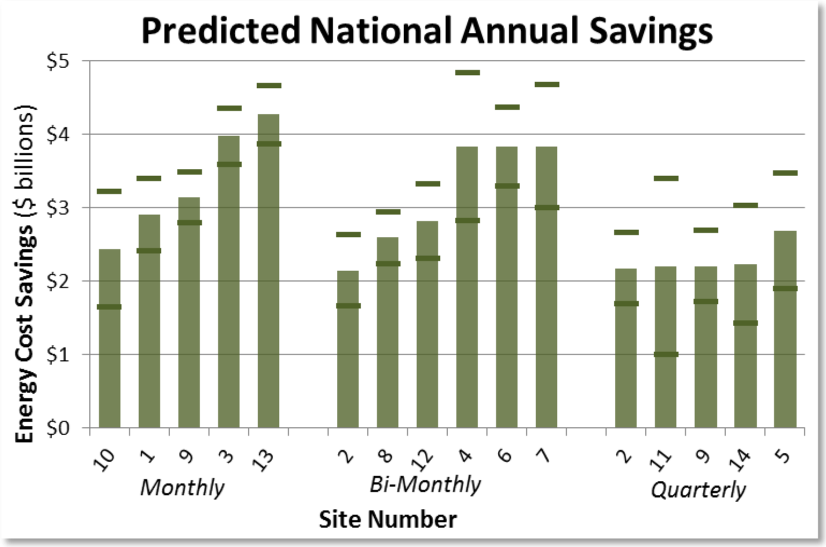
SAVE UP TO
$30 PER YEAR

Figure 3: Map of Opower Partner Utilities



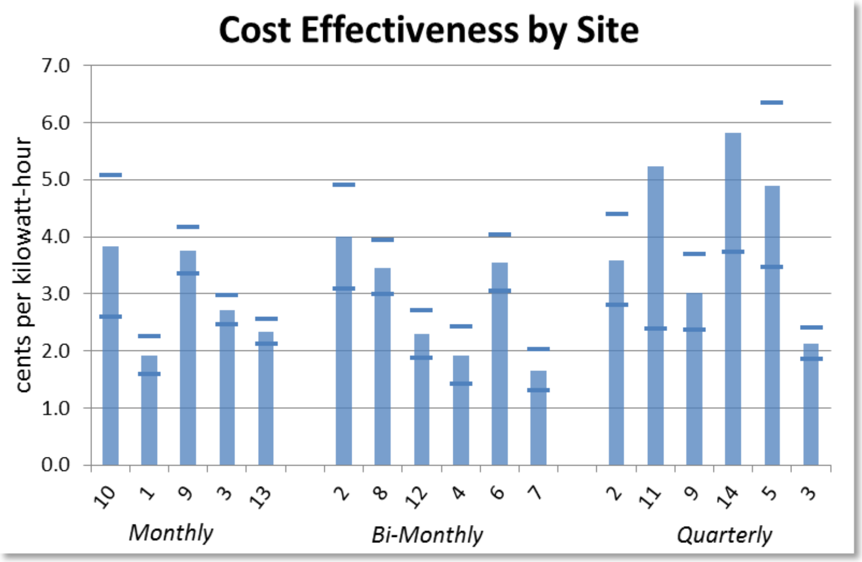
Notes: Highlighted states are those where Opower has a partner.

Figure 4: Predicted National Annual Savings



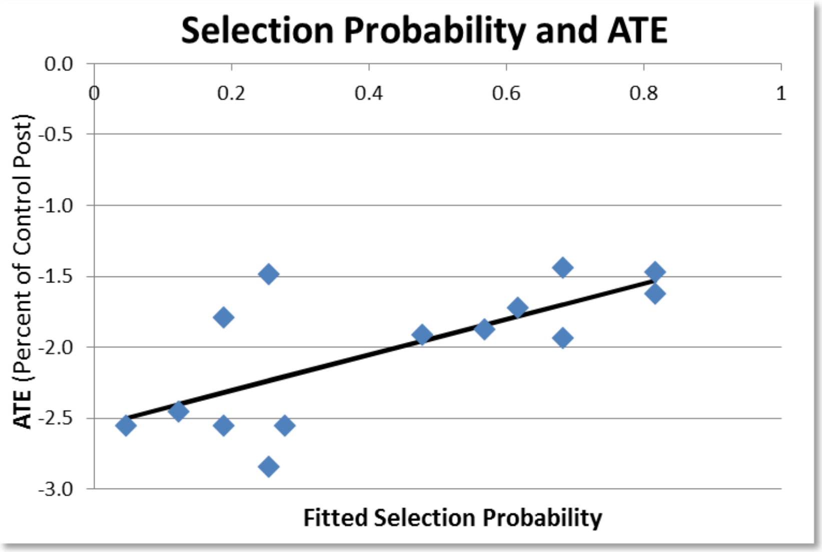
Notes: This figure illustrates the U.S. annual electricity cost savings that would be predicted from generalizing the treatment effect from each of Opower's first 14 sites. Lines illustrate 90% confidence intervals.

Figure 5: Cost Effectiveness by Site



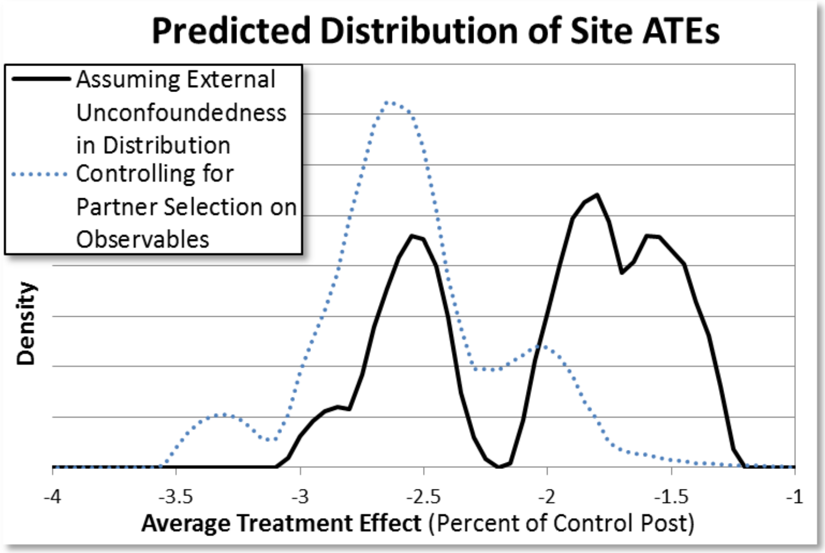
Notes: This figure presents the cost effectiveness at each site in cents of program cost per kilowatt-hour conserved. Lines illustrate 90% confidence intervals.

Figure 6: Partner Selection Bias



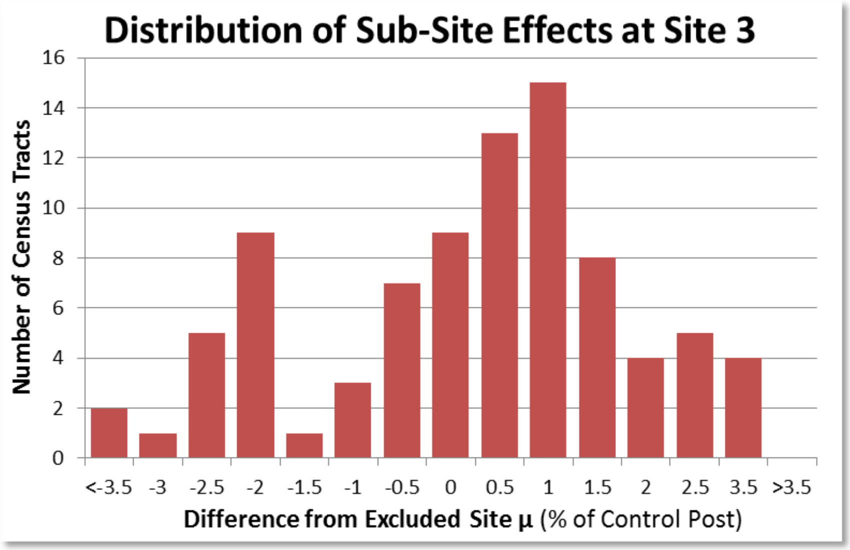
Notes: This figure shows the fitted relationship between the estimated unconditional Average Treatment Effect and the fitted probability of being an Opower partner, for the first 14 sites.

Figure 7: Predicted Distributions of ATEs in the Population of Sites



Notes: The black line is the predicted distribution of Opower ATEs in the population of potential partner sites assuming that partners are randomly drawn from the distribution of potential partners. The dotted blue line is the predicted distribution of ATEs after adjusting for partner selection on observables using the estimates of Equation (15) in Column I of Table 4. These distributions are estimated using a kernel density estimator with bandwidth of 0.1 percentage points.

Figure 8: Distribution of Sub-Site Effects at Site 3



Notes: This figure is a histogram of the point estimates of Census Tract-level heterogeneous treatment effects at Opower site 3.

Appendix: For Online Publication

External Validity and Partner Selection Bias

Hunt Allcott and Sendhil Mullainathan

Tables

Online Appendix Table A1: Estimated Site Effects for Table 4

Interactions with TxPost	*I*	*II*	*III*	*IV*	*V*	*VI*	*VII*
Site 1	-1.94	-1.39	-2.07	-1.64	4.05	4.478	4.554
	(0.20)***	(0.24)***	(0.20)***	(0.25)***	(0.93)***	(0.950)***	(0.921)***
Site 2	-1.44	-1.45	-1.60	-1.17	3.10	3.027	2.950
	(0.16)***	(0.19)***	(0.17)***	(0.21)***	(0.86)***	(0.872)***	(0.839)***
Site 3	-2.46	-2.13	-2.57	-2.25	1.29	1.558	1.385
	(0.13)***	(0.16)***	(0.14)***	(0.21)***	(0.80)	(0.806)*	(0.737)*
Site 4	-2.55	-2.57	-2.72	-2.30	2.67	2.566	2.547
	(0.41)***	(0.46)***	(0.41)***	(0.43)***	(0.92)***	(0.949)***	(0.879)***
Site 5	-1.79	-1.79	-1.99	-1.57	1.16	1.044	1.001
	(0.32)***	(0.32)***	(0.32)***	(0.36)***	(0.83)	(0.837)	(0.774)
Site 6	-2.56	-2.58	-2.77	-2.36	1.16	0.976	1.047
	(0.22)***	(0.31)***	(0.22)***	(0.23)***	(0.85)	(0.892)	(0.831)
Site 7	-2.56	-2.58	-2.70	-2.44	1.21	1.167	1.166
	(0.34)***	(0.40)***	(0.34)***	(0.36)***	(0.93)	(0.960)	(0.894)
Site 8	-1.73	-1.75	-1.90	-1.48	2.40	2.293	2.233
	(0.14)***	(0.26)***	(0.15)***	(0.20)***	(0.88)***	(0.909)**	(0.843)***
Site 9	-1.92	-1.53	-2.01	-1.81	2.06	2.394	2.393
	(0.13)***	(0.16)***	(0.13)***	(0.17)***	(0.78)***	(0.794)***	(0.764)***
Site 10	-1.62	-1.08	-1.77	-0.81	3.54	3.940	4.026
	(0.32)***	(0.35)***	(0.32)***	(0.42)*	(1.00)***	(1.026)***	(1.005)***
Site 11	-1.47	-1.47	-1.62	-0.61	1.97	1.838	1.780
	(0.48)***	(0.48)***	(0.49)***	(0.56)	(1.06)*	(1.074)*	(1.055)*
Site 12	-1.88	-1.90	-2.06	-1.81	2.32	2.216	2.096
	(0.20)***	(0.30)***	(0.21)***	(0.22)***	(0.82)***	(0.852)***	(0.790)***
Site 13	-2.84	-2.30	-2.94	-2.62	1.57	2.030	2.169
	(0.26)***	(0.30)***	(0.25)***	(0.25)***	(0.83)*	(0.843)**	(0.808)***
Site 14	-1.49	-1.49	-1.57	-1.31	2.15	2.094	2.188
	(0.37)***	(0.37)***	(0.38)***	(0.38)***	(0.90)**	(0.910)**	(0.880)**

Notes: This table presents tests of whether site indicator variables estimated in Equation (13) are different from each other. Columns I through VII condition on different sets of observables. These are the estimated $\widehat{\mu}$ coefficients that correspond to the same columns in Table 4. Robust standard errors, clustered by household, are in parenthesis. *, **, ***: Statistically significant with 90%, 95%, and 99% confidence, respectively.