NBER WORKING PAPER SERIES

SITE SELECTION BIAS IN PROGRAM EVALUATION

Hunt Allcott

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

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

This paper replaces an earlier version which was titled "External Validity and Partner Selection Bias." The earlier draft was jointly authored with Sendhil Mullainathan, and the project benefited substantially from his insight and collaboration. I thank Josh Angrist, Amitabh Chandra, Lucas Davis, Kyle Dropp, Meredith Fowlie, Xavier Gine, Chuck Goldman, Matt Harding, Joe Hotz, Guido Imbens, Larry Katz, Chris Knittel, Dan Levy, Jens Ludwig, Konrad Menzel, Emily Oster, Rohini Pande, Todd Rogers, Piyush Tantia, Ed Vytlacil, Heidi Williams, five anonymous referees, and seminar participants at the ASSA meetings, Berkeley, Columbia, Harvard, MIT, NBER Labor Studies, NBER Energy and Environmental Economics, NEUDC, the UCSB/UCLA Conference on Field Experiments, and the World Bank for insights and helpful advice. Thanks also to Tyler Curtis, Marc Laitin, Alex Laskey, Alessandro Orfei, Nate Srinivas, Dan Yates, and others at Opower for fruitful discussions. Christina Larkin provided timely research assistance. I am grateful to the Sloan Foundation for financial support of this paper and related research on the economics of energy efficiency. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2012 by Hunt Allcott. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Site Selection Bias in Program Evaluation
Hunt Allcott
NBER Working Paper No. 18373
September 2012, Revised September 2014
JEL No. C93,D12,L94,O12,Q41

ABSTRACT

"Site selection bias" can occur when the probability that a program is adopted or evaluated is correlated with its impacts. I test for site selection bias in the context of the Opower energy conservation programs, using 111 randomized control trials involving 8.6 million households across the U.S. Predictions based on rich microdata from the first ten replications substantially overstate efficacy in the next 101 sites. Several mechanisms caused this positive selection. For example, utilities in more environmentalist areas are more likely to adopt the program, and their customers are more responsive to the treatment. Also, because utilities initially target treatment at higher-usage consumer subpopulations, efficacy drops as the program is later expanded. The results illustrate how program evaluations can still give systematically biased out-of-sample predictions, even after many replications.

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

1 Introduction

Program evaluation has long been a important part of economics, from the Negative Income Tax experiments to the wave of recent randomized control trials (RCTs) in development, health, and other fields. Often, evaluations from one or more sample sites are generalized to make a policy decision for a larger set of target sites. Replication is valued because program effects can often vary across sites due to differences in populations, treatment implementation, economic environments. As Angrist and Pischke (2010) write, "A constructive response to the specificity of a given research design is to look for more evidence, so that a more general picture begins to emerge." If a program works well in a number of different replications, we might advocate that it be scaled up.

Formally, this logic involves an "external unconfoundedness" assumption which requires that sample sites are as good as randomly selected from the population of target sites. In practice, however, there are often systematic reasons why sites are selected for empirical analysis. For example, because RCTs often require highly-capable implementing partners, the set of actual RCT partners may have more effective programs than the average potential partner. Alternatively, potential partners with existing programs that they know are effective are more open to independent impact estimates (Pritchett 2002). Both of these mechanisms would generate positive *site selection bias*: treatment effects in sample sites would be larger than in target sites. On the other hand, innovative organizations that are willing to test new programs may already have many other effective programs in the same area. If there are diminishing returns, a new program with an actual partner might have lower impact than with the average potential partner. This would cause negative site selection bias. Site selection bias implies that even with a large number of internally valid replications, policymakers could still draw systematically biased inference about a program's impact at full scale.

While site selection bias is intuitive and potentially important, there is little empirical evidence on this issue or the potential mechanisms in any context. The reason is simple: since this type of selection operates at the level of the site instead of the individual unit, one needs a large sample of *sites* with internally valid evaluations of the same treatment. Then, one must define a population of potential partner sites and somehow infer treatment effects in sites where evaluations have not yet been carried out. Given the cost of RCTs, it is unusual for the same intervention to be rigorously evaluated at more than a small handful of sites. By contrast, as in LaLonde (1986), Dehejia and Wahba (1999), Heckman, Ichimura, Smith, and Todd (1998), Smith and Todd (2004), and many other studies, providing evidence on individual-level selection bias simply requires a large sample of *individuals*.

The Opower energy conservation program provides an exceptional opportunity to study a site selection process. The treatment is to mail "Home Energy Reports" to residential energy consumers that provide energy conservation tips and compare their energy use to that of their neighbors. As

of February 2013, the program had been implemented in 111 RCTs involving 8.6 million households at 58 electric utilities across the United States.

This paper's initial organizing question is, *what would be the effects if the Opower program were scaled nationwide?* This out-of-sample prediction problem is particularly policy-relevant. In recent years, "behavior-based" energy conservation programs such as Home Energy Reports have received increasing attention as alternatives to traditional approaches such as subsidies and standards for energy efficient capital stock. Consultancy McKinsey & Co. recently released a study predicting "immense" potential for behavior-based conservation in the U.S., with potential savings amounting to 16 to 20 percent of current residential energy consumption (Heck and Tai 2013). Policymakers use such predictions, as well as evaluations of early pilot programs, to help determine the stringency of energy conservation mandates.[1]

I begin by using microdata from Opower's first ten sites to predict aggregate nationwide effects. This is a highly promising setting for extrapolation: there are large samples totaling 508,000 households, ten replications spread throughout the country, and a useful set of individual-level covariates to adjust for differences between sample and target populations. Using standard econometric approaches, I predict savings of about 1.7 percent, or $2.3 billion in retail electricity costs, in the first year of a nationally-scaled program.

Aside from the microdata, I also have Opower's "metadata": impact estimates from all 111 RCTs that began before February 2013. As an "in-sample" test of external validity, I use the microdata from the first ten sites to predict first-year effects at the 101 later sites. The microdata over-predict efficacy by approximately 0.5 percentage points, or $690 million worth of retail electricity. This shows that even in such a highly promising setting for extrapolation, estimates are not externally valid: early sites were strongly positively selected from later sites through mechanisms associated with the treatment effect.

I then use the metadata to explain this positive selection. It occurs both between utilities and within utilities at early vs. later customer sub-populations. Much of the within-utility trend reflects successful initial targeting of higher-usage households that are more responsive to treatment. If a program works well in an initial sub-population, many utilities later expand it to additional sub-populations within their service area. The between-utility trend is partially explained by two other mechanisms, neither of which reflects explicit targeting on gains. First, there was selection on "population preferences": high-income and environmentalist consumer populations both encourage utilities to adopt energy efficiency programs and are more responsive to the Opower program once it is implemented. Second, there was selection on utility ownership structure: for-profit investor-

[1] ENERNOC (2013), KEMA (2013), and Quackenbush (2013) are examples of state-level energy efficiency potential assessments that include predictions for behavioral energy efficiency programs based partially on results from RCTs in pilot locations. The studies were commissioned by utilities and state public utilities commissions as part of the process of setting Energy Efficiency Resource Standards. Allcott and Greenstone (2012) discuss the economic rationale for these types of policies.

owned utilities (IOUs) were less likely to adopt the program until early results from other utilities demonstrated its efficacy and until conservation mandates became more stringent. Although more IOUs have now adopted they program, they tend to experience lower efficacy, perhaps because their customers are less engaged and thus less responsive to utility-provided information.

The 111-site metadata can also help predict efficacy in a nationally-scaled program. Opower's current partners are still higher-income and more environmentalist than the average utility, which suggests lower efficacy. On the other hand, current partners are now disproportionately IOUs with smaller treatment effects. On net, current samples are still positively selected from the national population on site-level observables. But because there is also evidence of selection on site-level unobservables, an unbiased prediction may still not be possible - even after 111 replications.

This paper does not argue that site selection bias reflects sub-optimal behavior: just as individual-level selection into job training, education, or other treatments reflects rational choices by potential participants, site-level endogenous selection also reflects rational choices by potential partners. Indeed, beginning with the most responsive populations maximizes cost effectiveness if there is limited scaling capacity or uncertainty over efficacy. Instead, the point of the paper is that site-level selection can systematically bias inference and policy decisions, just as individual-level selection can. This paper also does not argue that RCTs should be de-emphasized in favor of less costly non-experimental approaches that could perhaps be implemented in a more general sample of sites: in the Opower context, Appendix C shows that it is still more informative to extrapolate RCT results from other sites than to rely on non-experimental estimates from the same site. Furthermore, site selection bias need not be limited to RCTs: for example, sites that collect high-quality data necessary for quasi-experimental analyses may also have systematically different institutions or economic environments which could generate different parameter estimates.

This paper builds on distinguished existing work on multi-site program evaluations, selection bias, external validity, and energy economics. In particular, the Job Training Partnership Act of 1982 (JTPA) provides closely-related evidence on external validity and site selection bias. The JTPA initiated job training programs at 600 sites, of which 200 were approached to do RCTs and 16 eventually agreed. Hotz (1992), Heckman (1992), Heckman and Vytlacil (2007), and others discuss the fact that these sites were non-randomly selected and propose that this could lead experimental estimates to differ from the true nationwide effects.[2] However, Heckman (1992) writes that the evidence from JTPA on external validity is "indirect" and "hardly decisive." Given average sample sizes of 270 people per site, Heckman and Smith (1997) show that it is not even possible to reject that the JTPA treatment effects are homogeneous across sites. With much larger samples and many more sites, the Opower experiments allow a clearer analysis of ideas proposed in the discussion of JTPA.

[2]Non-random site selection is part of what Heckman (1992) calls "randomization bias," although his discussion focuses also on other issues, such as how operational demands of RCTs could cause program performance to decline and how the need for control groups requires expansion of the pool of eligible individuals.

Also closely related are the large body of academic papers[3] and consulting reports[4] on Opower programs. Nolan *et al.* (2008) and Schultz *et al.* (2007) provided the academic "proof of concept" for the Home Energy Report. Although their experiment is not part of my meta-analysis, it is strikingly consistent with site selection bias. Their treatment was to hand-deliver door-hangers with energy use neighbor comparisons to about 300 homes in a wealthy California suburb, and the treatment effects are three to six times larger than even the first ten Opower programs.

The paper proceeds as follows. Section 2 presents case studies from microfinance and clinical trials of how RCT sites differ systematically from policy-relevant target sites. Section 3 formalizes a model of external validity and site selection bias. Section 4 gives an overview of the Opower experiments, and Section 5 presents the data. Section 6 uses the Opower microdata for extrapolation, while Section 7 uses the metadata to explain the site selection bias shown in Section 6. Section 8 concludes.

2 Motivation: Examples of Site Selection on Observables

I begin with two simple examples of how randomized control trial sample sites differ from policy-relevant populations of target sites. For both examples, I define a target population of sites and then compare sample to non-sample sites on observable characteristics that theory suggests could moderate treatment effects.

2.1 Microfinance Institutions

In the past ten years, there have been many randomized control trials with microfinance institutions (MFIs). Are MFIs that that partner with academics for RCTs representative of the MFIs that might learn from RCT results?

I define the population of sites as all MFIs included in the Microfinance Information Exchange (MIX) global database, which includes characteristics and performance of 1903 MFIs in 115 countries. Partners are defined as all MFIs listed as RCT partners on the Jameel Poverty Action Lab, Innovations for Poverty Action, and Financial Access Initiative websites. About two percent of MFIs in the database are RCT partners. I focus on eight MFI characteristics that might theoretically be correlated with empirical results. Average loan balance, percent of portfolio at risk of default, and the percent of borrowers who are female could be correlated with default rates, a common outcome variable. An MFI's structure (as measured by age, non-profit status, and size) could influence the strength of the MFI's relationship with its clients, which might in turn affect the

[3]These include Allcott (2011), Allcott and Rogers (2014), Ayres, Raseman, and Shih (2013), Costa and Kahn (2013), and others.

[4]Integral Analytics (2012), KEMA (2012), Opinion Dynamics (2012), Perry and Woehleke (2013), Violette, Provencher, and Klos (2009), and many other consulting reports evaluate individual programs for regulatory accounting purposes.

MFI's ability to implement or monitor an intervention. Similarly, staff availability and expenditures per borrower could affect implementation or monitoring ability.

Table 1 presents means and standard deviations by partner status. Column 4 presents differences in means for partners vs. non-partners. Partners have smaller average loan balances, as well as marginally insignificantly lower percent of portfolio at risk and more female borrowers. Each of these factors suggests lower default rates, which raises the question of whether treatment effects on default rates might be larger in non-partner sites given larger baselines. Partner MFIs are also older, larger, and more likely to be for profit, perhaps because RCTs require large samples and well-managed partners. Finally, partner MFIs have statistically significantly fewer staff and lower costs per borrower. Overall, partner MFIs differ statistically on six of the eight individual characteristics, and an F-test easily rejects the hypothesis that partners are representative on observables.

2.2 Clinical Trials

Are the hospitals that carry out clinical trials representative of hospitals where interventions might eventually be implemented?

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

Table 2 compares US hospitals that have been the site of at least one clinical trial to hospitals that have never hosted a registered trial. Clinical trial sites are from the ClinicalTrials.gov registry, while hospital characteristics are from Medicare and American Hospital Association databases; see Appendix A.1 for details of data preparation. I separately consider "drug" trials, which include drugs, biological interventions, and dietary supplements, and "procedure" trials, which include both surgical and radiation procedures, because hospital characteristics are almost certainly more important moderators for procedures compared to drugs. Of 4653 US hospitals, 1722 have hosted a drug trial and 1265 have hosted a procedure trial.

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

Clinical trial sites have significantly different *structures*. They are larger and perform more surgeries per year. Chandra and Staiger (2007) show that due to productivity spillovers, surgical procedures are more effective in areas that perform more surgeries, and they point out that this may compromise the external validity of randomized control trials. The average trial site also offers

five to six more of the 21 advanced technologies and three more of the 13 patient services scored in the U.S. News Hospital Quality Rankings. If these technologies and services are complements to surgical procedures, then such interventions will be less effective at non-trial sites.

Clinical trial sites also have significantly different *processes*. They perform 0.33 to 0.35 standard deviations better on five surgical process measures included in the Hospital Safety Score (HSS) methodology, which could suggest that surgical procedures are more effective at trial hospitals. On the other hand, patient surveys show that doctors and nurses at trial site hospitals are worse at communication, including explaining medicines and what to do during recovery.

Although this may be due to patient selection instead of treatment effects, clinical trial sites perform worse on two *outcome* measures: they have higher rates of hospital acquired conditions and higher rates of the six complications included in the HSS patient safety indicator index. On the other hand, trial sites have substantially lower mortality rates when treating patients suffering from heart attack, heart failure, and pneumonia.

Finally, clinical trial sites are significantly more likely to appear in the top 50 hospitals in 12 specialties rated by the U.S. News Hospital Quality Rankings, and they have an average of 0.17 to 0.29 additional specialties ranked. These results point to "ability bias" as a site selection mechanism in clinical trials: almost mechanically, clinical trials take place at higher-quality hospitals because technology, size, and skill are complements to clinical research.

MFIs and clinical trials are rare settings where there are many sample and target sites and it is possible to gather site-level characteristics. But while suggestive, both examples are speculative and incomplete. Ideally, we could focus on one well-defined treatment and present concrete evidence on the mechanisms that drive site selection and how site selection affects out-of-sample inference. The Opower program provides a unique opportunity to do this.

3 A Model of External Validity and Site Selection Bias

3.1 External Validity

This section briefly lays out the assumptions required for external validity, closely following Hotz, Imbens, and Mortimer (2005). Consider the standard Rubin (1974) Causal Model. $T_i \in \{1, 0\}$ is the treatment indicator variable for individual i, and each individual has two potential outcomes, $Y_i(1)$ if treated and $Y_i(0)$ if not. Individual i's difference in potential outcomes is $\tau_i = Y_i(1) - Y_i(0)$. X_i is a vector of observable covariates. Individuals are either in a sample population which was exposed to treatment or a target population for which we wish to infer treatment effects. $D_i \in \{1, 0\}$ is an indicator that takes value one if individual i is in the sample.

The average treatment effect (ATE) in a target population can be consistently estimated under four assumptions:

Assumption 1: Unconfoundedness. $T_i \perp (Y_i(1), Y_i(0)) \,|X_i$

Assumption 2: Overlap. $0 < Pr(T_i = 1|X_i = x) < 1$

Assumption 3: External unconfoundedness. $D_i \perp (Y_i(1) - Y_i(0)) \, | X_i$

Assumption 4: External overlap. $0 < Pr(D_i = 1|X_i = x) < 1$

The external unconfoundedness and external overlap assumptions are just sample-target analogues of the familiar assumptions required for internal validity. If Assumptions 1-4 hold in the support of X in the target population, then the target ATE can be estimated from sample data, after controlling for differences in X between treatment and control and between sample and target:

$$E[\tau_i|D_i = 0] = E\left[E[Y_i|T_i = 1, D_i = 1, X_i] - E[Y_i|T_i = 0, D_i = 1, X_i]|D_i = 0\right]. \quad (1)$$

This argument is closely comparable to Lemma 1 in Hotz, Imbens, and Mortimer (2005).[5]

3.2 Sites, Replication, and Site Selection Bias

External unconfoundedness requires conceptually different assumptions in single-site vs. multi-site evaluations. Specifying these assumptions both clarifies the importance of replication and defines site selection bias.

Define a "site" as a setting in which one program might be implemented or evaluated. Sites are indexed by s, and the integer variable S_i indicates the site of which individual i is a member. A site consists of three elements: a population of individuals, a treatment (as implemented, for example, by an MFI, job training center, or hospital), and an economic environment (for example, market interest rates, labor market conditions, or disease prevalence).[6] Defining F_s and V_s as vectors of characteristics of the treatment and economic environment, respectively, $\tau_s(x) = E[\tau_i|X_i = x, S_i = s] = E[\tau_i|X_i = x, F_s, V_s]$ is the average treatment effect at site s conditional on $X_i = x$.

In the Opower example, the decision to implement or evaluate a program is made at the site level, so I assume that either all individuals in a site are in sample or all are in target.[7] $D_s \in \{1, 0\}$

[5]The proof follows Hotz, Imbens, and Mortimer (2005) almost identically. The two unconfoundedness assumptions imply that for any value x of the covariates, the target treatment effect is estimated by the treatment-control difference in outcomes in the sample: $E[\tau_i|D_i = 0, X_i = x] = E[\tau_i|D_i = 1, X_i = x] = E[Y_i|D_i = 1, T_i = 1, X_i = x] - E[Y_i|D_i = 1, T_i = 0, X_i = x]$. Then, the two overlap assumptions imply that it is feasible to estimate the target ATE by taking the expectation of this difference over the distribution of X in the target population.

There are two minor differences, however. First, unlike their Lemma 1, Equation (1) does not require random assignment of treatment within sample sites, so it is relevant for quasi-experimental analyses as well. Second, external unconfoundedness is a weaker version of their "unconfounded location" assumption, which is $D_i \perp (Y_i(1), Y_i(0)) \, | X_i$. The external unconfoundedness assumption clarifies that only the *difference* in potential outcomes need be independent of D_i. The stronger assumption can be used to motivate tests of $D_i \perp Y_i(0)|X_i$ as evidence of external unconfoundedness, but $D_i \perp Y_i(0)|X_i$ is in theory neither necessary nor sufficient. In Appendix C, I show that this test is empirically uninformative in the Opower context, because the ability to predict untreated outcomes $Y(0)$ depends largely on weather variation, while treatment effects τ differ across sites for many other reasons.

[6]The idea of a site connects to Heckman and Vytlacil (2005), who discuss extrapolation from a "history" of "policy-environment pairs." The exogeneity assumption in their Equation (A-9) is conceptually analogous to external unconfoundedness.

[7]This simplifying assumption is appropriate for Opower but inappropriate for some other settings. Heckman and Vytlacil (2005) show how the marginal treatment effects approach can be used when individual-level selection

is an indicator that takes value one if s is a sample site. This model could reflect sites choosing whether to adopt a new program, as with Opower, or whether to evaluate an existing program, as with JTPA.

Consider two alternative assumptions:

Assumption 3A: Homogeneous site effects. $\tau_{s'}(x) = \tau_{s''}(x)$ for a pair of sites s' and s''

Assumption 3B: No site selection bias. $E[\tau_s(x)|D_s = 1] = E[\tau_s(x)|D_s = 0]$ over a large number of sites

When extrapolating from single sample site to a single target site, external unconfoundedness is equivalent to the homogeneous site effects assumption. In practice, however, it is rarely plausible that two different sites have the same treatment effects. This would hold if individuals were somehow randomly (or quasi-randomly) assigned between the two sites *and* if there were no site-level differences in treatment implementation F_s or economic environments Z_s. Nevertheless, this assumption is made (either implicitly or explicitly) whenever results from a single-site analysis are used to infer effects out of sample.[8]

By contrast, when extrapolating from many sample sites to many target sites, external unconfoundedness is equivalent to Assumption 3B. This is a weaker than Assumption 3A, because it allows heterogeneity in $\tau_s(x)$ across sites as long any site-level heterogeneity averages out. The plausibility of Assumption 3B depends on the assignment mechanism that allocates sites to sample. It would be guaranteed if a large number of sites were randomly assigned to sample. 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 assumption would also hold with quasi-random site assignment, which could arise in a multi-site evaluation if evaluators choose sample sites to maximize external validity. For example, the Moving to Opportunity and RAND Health Insurance experiments were implemented in multiple cities chosen for diversity in size and geographic region (Sanbonmatsu *et al.* 2011, Manning *et al.* 1988).

This discussion formalizes the appeal of replication but also highlights the limitation: replication allows external unconfoundedness to hold even when there is site-specific heterogeneity - as long as replication sites are chosen randomly or quasi-randomly. This discussion also formalizes *site selection bias*: the failure of external unconfoundedness when sites are assigned to sample through mechanisms other than random or quasi-random assignment. Notice that site selection bias is quite distinct from the treatment effect heterogeneity relevant to defining local average treatment effects (Angrist and Imbens 1994): within-sample heterogeneity in τ_i is neither necessary nor sufficient for

mechanisms vary across sites.

[8] As an example of how the no site effects assumption has been made explicitly, 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 (LFA) (Dehejia 2003, Hotz, Imbens, and Klerman 2006). These analyses require that there are no unobservable factors other than the use of LFA that moderate the treatment effect and differ across sites.

site selection bias.[9] Notice also that site selection bias does not mean that the estimated sample treatment effects are biased away from the true sample treatment effects. Instead, the word "bias" underscores that sample effects can be systematically different from target effects due to systematic site selection mechanisms.

The next several sections test Assumption 3B in the specific context of Opower and give intuition for the forces that generate site selection bias in that context.

4 Opower: Overview and Site Selection Mechanisms

4.1 The Home Energy Report Program

The Home Energy Report is a two-page letter with two key components. The Neighbor Comparison Module at the top of the first page compares the household's energy use to its 100 geographically-nearest neighbors in similar house sizes. The Action Steps Module, which is typically on the second page, includes energy conservation tips targeted to the household based on its historical energy use patterns and observed characteristics. The envelope and report are branded with the utility's name, as this is believed to increase open rates, perceived credibility, and the utility's customer satisfaction. Appendix Figures A2 and A3 present an example report.

Except at a few utilities whose customer bases are too small for precise impact estimates, all Opower programs are implemented as randomized control trials, because it is easy to hold out a randomized control group from a mail-based program. The treatment group is sent reports at frequencies that vary within and between households and sites. For example, of the first ten programs, two randomized households between monthly and quarterly frequencies, while three others targeted heavier users with monthly reports and lighter users with quarterly. One common pattern is to start with three monthly reports and then decrease to a bimonthly frequency.

The reports vary within-household over time: for example, the information and tips are updated each month to reflect the customer's most recent energy bills and season-specific energy conservation tips. The reports also vary somewhat across sites, at a minimum because they carry different utility names. However, the basic design and implementation are highly consistent, and 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 effects of "contract teachers" could depend markedly on the teacher's ability and even who employs them (Bold *et al.* 2013). This suggests that after accounting for differences in treatment frequency, other variation in treatment is relatively unlikely to cause substantial site-level heterogeneity. The more likely causes would thus be variation in treated

[9] One might draw the analogy between a site and a set of compliers in the LATE framework. In this analogy, site selection bias would arise if the kinds of instruments available tended to identify *systematically* different populations - i.e., that LATEs from different instruments were not only heterogeneous but were systematically different from the ATE in a target population.

populations and "economic environments" V_s, which tangibly include several factors discussed below.

Aside from treatment fidelity, there are two other useful features of the Opower experiments. First, in the taxonomy of Harrison and List (2004), these are "natural field experiments," meaning that people are in general not aware that they are being studied. Second, these are "opt-out" experiments, and opting out requires actively calling the utility and canceling. In the average program, only about 0.6 percent of the treatment group opts out over the first year. Thus, there is no need to model essential heterogeneity or household-level selection into the treatment (Heckman, Urzua, and Vytlacil 2006), and the treatment effect is a Policy-Relevant Treatment Effect in the sense of Heckman and Vytlacil (2001).

4.2 Site Selection Mechanisms

For the Opower program, there are two levels of site selection. First, a utility contracts with Opower. In theory, the partnership decision is an equilibrium outcome of Opower's sales outreach efforts and utility management decisions. In practice, most of the selection derives from demand-side forces, as Opower will implement the program with any utility willing to pay for it, and the company's initial sales efforts were largely targeted at utilities that were most likely to be interested. As recounted in personal communication (Laskey 2014), Opower's early outreach efforts sound remarkably similar to an economist searching for a field experiment partner: the founders started with existing personal connections, cold called other utilities that they thought might be interested, and then moved forward with any partners that agreed. The founders initiated discussions with 50 to 100 utilities in order to land the first ten (Laskey 2014), and by now, the the program is very well-known nationwide. Thus, I focus on selection mechanisms that make utilities interested in the program, with less attention to Opower's outreach process.

Discussions with Opower executives and utility industry practitioners suggest five potential utility-level selection mechanisms that could also moderate treatment effects:

- **Usage.** Utilities use metrics such as cost effectiveness (measured in kilowatt-hours saved per dollar spent) to make program adoption decisions, and the program's potential savings are larger at utilities with higher usage.

- **Population Preferences**. Environmentalist states are more likely adopt Energy Efficiency Resource Standards (EERS) that require utilities to run energy conservation programs, and even in the absence of such regulation, utility managers from environmentalist areas might be more likely to prioritize conservation. If environmentalism or related cultural factors also make consumers more responsive to conservation messaging, this would generate positive selection.

- **Complementary or Substitute Programs.** Utilities that prioritize energy conservation should be more likely to adopt the Opower program. Depending on whether a utility's other programs are complements or substitutes to Opower, this would generate positive or negative selection. Complementarity is possible because one way that consumers respond to the Opower treatment is by participating in other utility programs, such as energy efficient insulation and lighting replacement (Allcott and Rogers 2014). However, such programs could instead be substitutes, because households that have already installed energy efficient insulation or lighting would save less energy when adjusting the thermostat or turning off lights in response to the Opower treatment.

- **Size.** Larger utilities have economies of scale with Opower because of fixed costs of implementation and evaluation. This could cause negative selection, because larger utilities tend to be in urban areas where people are less likely to know their neighbors and are thus potentially less responsive to neighbor energy use comparisons.

- **Ownership.** Different types of utilities implement energy conservation programs for different reasons. For-profit investor-owned utilities (IOUs) typically have little incentive to run energy efficiency programs in the absence of EERS policies. By contrast, municipally-owned utilities and rural electric cooperatives are more likely to maximize welfare instead of profits, so they run energy efficiency programs if they believe the programs benefit customers. Ownership structure could also be associated with treatment effects: for-profit IOUs average lower customer satisfaction rankings in the JD Power (2014) survey, and related forces may cause IOU customers to be less likely to trust and use utility-provided information.

After a utility contracts with Opower, the second level of site selection occurs when the utility, with guidance from Opower, chooses a sample population of residential consumers within the utility's service territory. Some small utilities choose to include the entire residential consumer base, while other utilities target specific local areas where reduced electricity demand could help to delay costly infrastructure upgrades. Simple theory, along with empirical results in Schultz *et al.* (2007), suggests that relatively high-usage households would conserve more in response to the treatment, both because they have more potential usage to conserve and because the neighbor comparisons induce them to decrease usage toward the norm. Thus, some utilities include only relatively heavy users in a sample population.

Opower differs in two ways from some other programs evaluated in the economics literature. First, Opower's for-profit status meant that the company could benefit from early successes.[10] However, this does not make their site selection incentives qualitatively different: social programs

[10] All of Opower's first ten sites had fee-for-service contracts without performance incentives. This has largely continued to be the case, although a small number of contracts include additional payments for larger effects. Regardless of contract structure, efficacy at previous sites affects subsequent utility adoption decisions.

and non-profits depend on government or foundation funds that can also hinge on the results of early evaluations. Pritchett (2002) shows how such incentives could lead to an equivalent of site selection bias.

Second, because the program is opt-out instead of opt-in, utilities can explicitly target more responsive households. It is ambiguous whether this generates stronger or weaker selection on gains than an opt-in program such as job training: this depends on whether individuals' perceived net utility gains have higher or lower covariance with τ_i than site managers' targeting decisions. While Opower now has substantial data with which to predict treatment responsiveness, utilities have been reticent to target based on observables other than high energy use because of concerns over customer equity.

5 Data

This section provides a brief overview of the three main datasets: utility-level data, microdata from Opower's first ten sites, and metadata from all 111 sites that began before February 2013. Appendix A provides substantial additional information.

5.1 Utility-Level Data

I define the policy-relevant consumer population to be all residential consumers at all 882 large electric utilities in the United States.[11] Table 3 shows characteristics of Opower partner utilities and utilities from the population of potential partners. The 58 partner utilities include all US electric utilities that had started Home Energy Report RCTs by February 2013.[12]

I consider variables that proxy for the five utility-level selection mechanisms proposed in Section 4.2 - that is, variables that might moderate both selection and treatment effects. Utility-specific data are from the Energy Information Administration (EIA) Form 861 for calendar year 2007 (EIA 2013), the year before the first Opower programs began. I also merge additional variables by taking population-weighted means of county-level data for the counties in each utility's service territory.

Utility Mean Usage is daily average residential electricity usage. For context, one kilowatt-hour (kWh) is enough electricity to run either a typical new refrigerator or a standard 60-watt incandescent lightbulb for about 17 hours.

The next seven variables proxy for population preferences. Mean Income and Share College Grads are from 2000 Census county-level data, while Hybrid Auto Share uses each county's share

[11] This figure excludes utilities with fewer than 10,000 residential consumers and power marketers in states with deregulated retail markets, as Opower has no clients in these two categories. About five percent of utilities operate in multiple states. To reflect how state-level policies affect utilities' program adoption decisions, a utility is defined as a separate observation for each state in which it operates.

[12] Three additional utilities started Home Energy Report programs before that date but did not evaluate them with RCTs because the customer populations were too small to include randomized control groups.

of registered vehicles that were hybrid-electric as of 2013. Green Party Share is the county-level share of votes in the 2004 and 2008 presidential elections that were for the Green party candidate, while Democrat Share is the share of Democratic and Republican votes that were for the Democratic candidate, both from Leip (2013). Energy Efficiency Resource Standard is an indicator for whether the utility is in a state with an EERS, using data from the Pew Center (2011). Green Pricing Share is the share of residential consumers that have voluntarily enrolled in "green pricing programs," which sell renewably-generated energy at a premium price. For the empirical analysis, I use the variable "Normalized Population Preferences" as a single proxy for higher income and environmentalism. This is just the sum of these seven variables, after normalizing each variable to mean zero, standard deviation one.

The next two variables measure complementary or substitute programs: the ratio of estimated electricity conserved by residential energy conservation programs to total residential electricity sold ("Residential Conservation/Sales") and the ratio of total spending on energy conservation programs to total revenues ("Conservation Cost/Total Revenues"). I construct a single proxy called "Normalized Other Programs" by adding these two variables after normalizing each to mean zero, standard deviation one.

The final three variables measure utility size and ownership. Utilities that are neither IOUs nor municipally-owned are either rural electric cooperatives or other government entities such as the Tennessee Valley Authority.

Table 3 shows that Opower's partner utilities are clearly different from non-partners: they use less electricity, have higher socioeconomic status and stronger environmentalist preferences, have more existing energy efficiency programs, and are much larger and more likely to be investor-owned. All of the 13 utility-level covariates are unbalanced with more than 90 percent confidence.

5.2 Microdata

I have household-level microdata through the end of 2010 for each of the ten Opower programs that began before December 2009. This includes 21.3 million electricity meter reads from 508,295 households, of which 5.4 million are in the first year post-treatment. The dataset includes Census tracts, which I use to merge in tract-level data, as well as household-level demographic data from public records and marketing data providers. Columns 1, 2, and 3 of Table 4 present observation counts, means, and standard deviations, respectively. Every variable has at least some missing observations. Most missing observations are missing because the variable is unavailable for the entire site.

I consider 12 X covariates that proxy for four mechanisms that theory suggests could moderate treatment effects. The first three mechanisms connect to the first three site-level selection mechanisms detailed in Section 4.2. "First Comparison" is the usage difference in kWh/day between a household and its mean neighbor, as reported in the Social Comparison Module on the first report.

(Opower also constructs this for control households.) The mean of 1.46 implies that these first ten sites consisted of slightly above-mean usage households due to utilities' program targeting decisions.

The next four variables proxy for population preferences. Mean Income, Share College Grads, and Share Hybrid Autos are Census tract means from the same source as their utility-level analogues in Table 3, while Green Pricing Participant is at the household level.[13]

"EE Program Participant" is an indicator for whether the household had received a loan or rebate for an energy efficient appliance, insulation, or a heating, ventilation, and air conditioning system through another utility program before the Opower program began. This and the Green Pricing Participant indicator are only available at one site.

The final six variables measure characteristics of housing stock. While I do not hypothesize that site-level variation in these factors directly affects site selection, there are clear theoretical reasons why each of these six characteristics could moderate the treatment effect. One natural way for households to respond to treatment is to lower thermostat temperatures in the winter, and having electric heat (instead of gas or oil heat) implies that this would reduce electricity use. Because building codes have been progressively tightened over the past 30 years, older homes are less energy efficient and offer more low-cost opportunities to conserve. Replacing pool pumps can save large amounts of energy. Renters have less ability and incentive to invest in energy efficient capital stock in their apartments. Occupants of single family dwellings have more control over their electricity use.[14]

Section 6 will condition on these variables for out-of-sample prediction, and columns 4 and 5 of Table 4 present the target population means to which the effects are fitted. Column 4 is the national mean across all 882 potential partner utilities, weighted by the number of consumers in each utility. This weighting means that the extrapolated effect will reflect the total potential savings if the treatment were scaled nationwide. Column 5 is the unweighted mean across the "later sites," which refers to the 101 Opower programs that started after the 10 programs in the microdata sample.

Table 4 shows that individuals in the microdata sample differ on observable proxies for population preferences: they have higher income, more college graduates, own more hybrid autos, and are more likely to participate in green pricing programs. Their houses also have somewhat different physical characteristics, with much less electric heat, fewer renters, and more single-family homes.

[13] I do not include tract-level Democrat vote share in the primary analysis because its association with the treatment effect is not robust to the inclusion of other covariates and is actually often negative, which is inconsistent with the sign at the site level. Appendix B provides intuition for why this happens and presents results including Democratic vote share.

[14] I must also gather utility-level data to construct target means of these housing stock characteristics. Mean square footage and share of homes with pools are from the American Housing Survey state-level averages, and share using electric heat, mean house age, share rented instead of owner-occupied, and share single family are from the county-level American Community Survey 5-year estimates for 2005-2009.

5.3 Metadata

Due to contractual restrictions, Opower cannot share microdata from many of their recent partners. Instead, they have provided site-level metadata, including average treatment effects and standard errors, control group mean usage, and number of reports sent for each post-treatment month of each RCT. Some utilities have multiple sites, typically because they began with one customer sub-population and later added other sub-populations in separate RCTs. As of February 2014, there were 111 sites with at least one year of post-treatment data at 58 different utilities. I focus on the ATEs over each site's first year in order to average over seasonal variation and eliminate duration heterogeneity while including as many sites as possible.

Opower's analysts estimated first-year ATEs using mutually-agreed procedures and code; see Appendix A for details. The 111 site-level populations average about 77,200 households, of which an average of 53,300 are assigned to treatment. The total underlying sample size for the meta-analysis is thus 8.57 million households, or about one in every 12 in the United States.

When analyzing the metadata, I consider ATEs both measured in levels (kWh/day) and normalized into a percent of control group post-treatment usage. Savings in kWh/day is relevant because it is more closely (although not perfectly) related to electricity cost savings and externality reductions, which are outcomes that enter cost effectiveness and welfare calculations. On the other hand, ATEs in percent terms are much more informative for out-of-sample prediction, because control group usage strongly predicts ATEs.[15] In other words, it is more useful for prediction to know that first-year effects in the metadata average 1.31 percent than it is to know that first-year effects average 0.47 kWh/day.[16] Because of this, Opower typically quotes ATEs in percent terms when reporting results publicly.

5.3.1 Dispersion of Site Effects

Is the heterogeneity across sites statistically significant? If effects do not vary across sites, then there is no possibility for site selection bias. Put formally, if Assumption 3A holds across all sites, this is sufficient for Assumption 3B. In reality, the 111 ATEs vary substantially, from 0.10 to 1.47 kWh/day and from 0.50 to 2.63 percent. This is statistically significant, in the sense that it is much larger than can be explained by sampling error: Cochran's Q test rejects that the effects are homogeneous with a p-value of less than 0.001. The percent ATEs have standard deviation of 0.45 percentage points, while the average standard error is only 0.18 percentage points.

Is this site-level heterogeneity also economically significant? One measure of economic significance is the dollar magnitude of the variation in predicted effects at scale. Figure 1 presents a

[15] The relationship between treatment effects in kWh/day and control group usage in kWh/day has a t-statistic of 11.35 and an R^2 of 0.54 across the 111 sites.

[16] Appendix Table A3 shows this formally: the coefficient of variation for ATEs is 55 percent higher when measured in kWh/day instead of percent.

forest plot of the predicted electricity cost savings in the first year of a nationwide program at all households in all potential partner utilities. Each dot reflects the prediction using the percent ATE from each site, multiplied by annual national residential retail electricity costs. The point estimates of first-year savings vary by a factor of 5.2, from $695 million to $3.62 billion, and the standard deviation is $618 million. Appendix A also documents a wide dispersion in cost effectiveness, which suggests that extrapolating effects from other sites could lead to ex-post program adoption errors.

This site-specific heterogeneity implies that Assumption 3A does not hold when not conditioning on X. The next section explores whether Assumption 3B holds: even if there are site effects, is it possible to condition on X and extrapolate from 10 replications?

6 Microdata: Extrapolation Under External Unconfoundedness

6.1 Empirical Strategy

Under Assumptions 1-4 in Section 3, microdata from the first ten Opower replications could identify the average treatment effect in a target population. Furthermore, because there are a relatively large number of replications, external unconfoundedness could hold under Assumption 3B (no site selection bias) even if Assumption 3A fails and there is site-specific treatment effect heterogeneity. Using the microdata, I first predict the average treatment effect for the program if it were scaled nationwide. I then test external unconfoundedness "in sample" by extrapolating from the microdata to the remainder of the sites in the metadata.

I address missing data using standard multiple imputation commands in Stata. I use the chained equations ("MICE") approach and estimate with 25 imputations, combining coefficients and standard errors according to Rubin (1987).[17]

The econometric techniques that can be used to condition on observables are limited by the fact that I observe only the means of X in the target populations. However, I can still use two simple off-the-shelf procedures commonly used in applied work: linear prediction and re-weighting to match means. In both procedures, I condition only on the subset of X variables that statistically significantly moderate the treatment effect. This increases precision in the re-weighting estimator, because it reduces extreme weights that match samples on Xs that don't actually moderate the treatment effect.

[17] Five imputations is standard in some applications, and 25 is certainly sufficient here: due to the large samples, parameter estimates are very similar in each individual imputation. Multiple imputation is consistent under the Missing at Random assumption. In earlier drafts, I instead used the missing indicator method, which is only unbiased under stronger assumptions (Jones 1996) but gives very similar results.

6.1.1 Determining the Set of Conditioning Variables

Y_{is} is household i's mean daily electricity use (in kWh/day) over the first year post-treatment, C_s is the control group mean usage in site s over that same year, and $y_{is} = \frac{100Y_{is}}{C_s}$. $\bar{X}_{D=1}$ is the vector of sample means of the covariates reported in Table 4, where the mean is taken across all 25 imputations. $\tilde{X}_{is} = X_{is} - \bar{X}_{D=1}$ is the vector of demeaned covariates. Y_{0i} is a vector of three baseline usage controls: average daily usage over the entire 12-month baseline period, the baseline winter (December-March), and the baseline summer (June-September). Heterogeneous treatment effects are estimated using the following equation:

$$y_{is} = -(\alpha\tilde{X}_i + \alpha_0)T_i + \sum_s \left(\beta_s\tilde{X}_i + \gamma_s Y_{0i} + \pi_s\right) + \varepsilon_{is}. \tag{2}$$

Equation (2) is analogous to the equation used to estimate ATEs for the metadata, but it also includes interactions with X.

The treatment causes energy use to decrease. By convention, I multiply the first term by -1 so that more positive α imply higher efficacy. The normalization of y_{is} is such that treatment effects can be interpreted as the percentage point effect on electricity use. For example, $\tau_s = 1$ would reflect a one percent effect.[18] Because $\tilde{X}$ are normalized to have mean zero in the sample, in expectation the constant term α_0 equals the sample ATE that would be estimated if $\tilde{X}$ were not included in the regression.

Standard errors are robust and clustered by the unit of randomization. In sites 1-9, randomization was at the household level. In site 10, households were grouped into 952 "block batch groups" - about the same size as Census block groups - that were then randomized between treatment and control.

I determine the set of conditioning variables using the "top-down" procedure of Crump, Hotz, Imbens, and Mitnik (2008). I start with the full set of X, estimate Equation (2), drop the one covariate with the smallest t-statistic, and continue estimating and dropping until all remaining covariates have t-statistic greater than or equal to 2 in absolute value. I denote this set of remaining covariates as X^*.

[18] While dividing Y_{is} by sample mean control group usage would be a purely presentational normalization, dividing Y_{is} by the site-specific C_s substantially improves prediction. The reason is that treatment effects in kWh/day depend both on a household's absolute level of usage (measured by average baseline usage) and relative usage compared to neighbors (measured by First Comparison). These two variables are highly collinear, making it difficult to separately identify their effects. (When interacting both with the treatment effect, the coefficients often have opposite signs depending on the other included covariates, whereas theory clearly predicts that they should both should increase the treatment effect.) Normalizing by site-specific C_s allows me to predict a percent ATE in a target site and also estimate how the effect varies within-site as a function of First Comparison. This allows predictions for target sites that both have different average absolute usage levels and also differentially target relatively heavy-usage households.

6.1.2 Linear Prediction

One approach to extrapolation is to assume that treatment effects are linear functions of X^* plus a constant:

Assumption 5: Linear treatment effects. $E[\tau_i|X_i = x] = \alpha x + \alpha_0$

I denote sample and target average treatment effects as $\tau_{D=1}$ and $\tau_{D=0}$, respectively. $\overline{X}^*_{D=0}$ is the vector of target mean covariates. Assuming external unconfoundedness and linear treatment effects, an unbiased estimator of the target treatment effect is:

$$\hat{\tau}_{D=0} = \hat{\tau}_{D=1} + \hat{\alpha}(\overline{X}^*_{D=0} - \overline{X}^*_{D=1}). \tag{3}$$

To implement this, I insert the estimated sample ATE $\hat{\tau}_{D=1}$ and the $\hat{\alpha}$ parameters from Equation (2) estimated with X^* only. Standard errors are calculated using the Delta method.

6.1.3 Re-Weighting

A second approach to extrapolation is to re-weight the sample population to approximate the target means of X^* using the approach of Hellerstein and Imbens (1999). Given that only the target means of X^* are observed, I assume that the target probability density function of observables $f_{D=0}(x)$ is the sample distribution $f_{D=1}(x)$ rescaled by λ, a vector of scaling parameters:

Assumption 6: Rescaled distributions. $f_{D=1}(x) = f_{D=0}(x) \cdot (1 + \lambda(x - \bar{X}^*_{D=0}))$

Under this assumption, observation weights $w_i = 1/(1 + \lambda(X^*_i - \bar{X}^*_{D=0}))$ re-weight the sample to exactly equal the target distribution of X^*.

Following Hellerstein and Imbens (1999), I estimate w_i using empirical likelihood, which is equivalent to maximizing $\sum_i \ln w_i$ subject to the constraints that $\sum_i w_i = 1$ and $\sum_i w_i X^*_i = \bar{X}^*_{D=0}$. In words, the second constraint is that the re-weighted sample mean of X^* equals the target mean. Given that the sum of the weights is constrained to 1, Jensen's inequality implies that maximizing the sum of $\ln w_i$ penalizes variation in w from the mean. Thus, the Hellerstein and Imbens (1999) procedure amounts to finding observation weights that are as similar as possible while still matching the target means.

6.1.4 Frequency Adjustment

Because treatment frequency varies across sites, with reports sent on monthly, bimonthly, quarterly, or other frequencies, I adjust for frequency when extrapolating and comparing ATEs. To do this, I estimate ϕ, the causal impact of frequency on the treatment effect, by exploiting the two sites in the microdata where frequency was randomly assigned between monthly and quarterly. A "frequency-adjusted treatment effect" $\widetilde{\tau}$ is adjusted to match the mean frequency $\overline{F}$ across all 111 sites in the metadata, which is 0.58 reports per month. Denoting the frequency at site s as F_s, the adjustment

is:

$$\tilde{\tau}_s = \hat{\tau}_s + \hat{\phi}(\overline{F} - F_s) \tag{4}$$

Standard errors are calculated using the Delta method.

6.2 Results

6.2.1 Heterogeneous Treatment Effects

Table 5 presents heterogeneous treatment effects using combined microdata from the first ten sites. Column 1 shows that the average treatment effect $\hat{\tau}_{D=1}$ across the first ten sites is 1.707 percent of electricity use. Because column 1 excludes the X covariates, this is the only column in Table 5 that does not use multiple imputation. The R^2 is 0.86, reflecting that fact that the lagged outcomes Y_{0i} explain much of the variation in y_{is}.

Column 2 presents estimates of Equation (2) including all $\tilde{X}$ variables. Column 3 presents the results from the last regression of the Crump, Hotz, Imbens, and Mitnik (2008) "top-down" procedure, including only the $\tilde{X}^*$ that statistically significantly moderate the treatment effect. Column 4 adds a set of 10 site indicators interacted with T. This identifies the α parameters only off of within-site variation, not between-site variation. Column 5 repeats column 3 after adding the interaction between T and average baseline usage. This tests whether the final α coefficients reflect an omitted association between X and baseline usage.

The $\hat{\alpha}$ coefficients are remarkably similar across columns. Furthermore, Appendix Table A8 presents estimates of Equation (2) for each of the 10 sites individually. None of the coefficients for the combined sample are solely driven by any one site, and there is only one $\hat{\alpha}$ from one site that is statistically significant and has a sign opposite the $\hat{\alpha}$ in the combined data.

The signs and magnitudes are also sensible. The first social comparison interaction is positive: informing a household that it uses ten kilowatt-hours per day more than its neighbors (which is about 1/3 of the mean) is associated with just less than a one percentage point larger treatment effect. Homes with electric heat conserve over one percentage point more, suggesting that reduced heating energy use is an important effect of the program. Homes that have pools or are 1000 square feet larger both have approximately 0.5 percentage point larger effects. Because these estimates condition on First Comparison, the α parameters for physical characteristics reflect the extent to which the characteristic is associated with the treatment effect relative to some other household characteristic that would use the same amount of electricity.

Appendix Table A9 presents the empirical likelihood estimates for the re-weighting estimator. As suggested by comparing sample and target means in Table 4, they imply lower weights for households with higher First Comparison and higher weights for households with electric heat.

Appendix Table A10 presents the estimated frequency adjustment. The estimated $\hat{\phi}$ is 0.515 percent of electricity use per report/month, and the estimates from each of the two sites alone are

economically and statistically similar. The point estimate implies that a one-standard deviation change in reports per month across the 111 sites (0.11 reports/month) would change the ATE by 0.515×0.11=0.056 percentage points. Frequency adjustment does not meaningfully impact the analyses, both because frequency is uncorrelated with other factors and because the adjustment is small relative to the variation in effect sizes.

6.2.2 Predicting Target Treatment Effects

Figure 2 presents the extrapolation results, with 90 percent confidence intervals. The left panel presents the frequency-adjusted sample ATE $\tilde{\tau}_{D=1}$. This is simply the estimate in column 1 of Table 5 adjusted to match the 111-site mean reports/month using Equation (4). The middle panel presents the predicted effects if the program were scaled "nationwide" to all households at all potential partner utilities. Per Equation (3), the "Linear Fit" is simply the frequency-adjusted sample ATE $\tilde{\tau}_{D=1}$ adjusted by the differences in sample and target mean X^* (the fourth minus the second column of Table 4) multiplied by the $\hat{\alpha}$ estimates (column 3 of Table 5). This linear adjustment is primarily composed of an increase of $(0.34 - 0.12) \times 1.196\% \approx 0.26\%$ predicted by the difference in Electric Heat, plus a decrease of $(0 - 1.47) \times 0.092\% \approx -0.14\%$ predicted by the difference in First Comparison. On net, the linear fit predicts a slightly larger nationwide ATE, while the weighted fit is close to the unadjusted sample ATE.

Using these standard approaches and assuming external unconfoundedness, the predicted nationwide effects in the program's first year would be about 1.7 percent of electricity use. This amounts to 21 terawatt-hours, or about the annual output of three large coal power plants. At retail prices, the first-year electricity cost savings would be $2.3 billion.

The results from the 101 later sites provide an opportunity to explicitly test the external unconfoundedness assumption. The right panel of Figure 2 shows the linear and weighted fits to the unweighted mean of X^* in the 101 later sites, along with the unweighted mean of the true ATEs in the metadata. The predicted ATEs are 0.64 and 0.40 percentage points larger than the true ATE. When scaled to the national level, a misprediction of 0.50 percentage points would amount to an overstatement of the first-year effects by 6.3 terawatt-hours, or $690 million in retail electricity cost savings. As suggested by the confidence intervals on the figure, the overpredictions are highly statistically significant, with p-values<0.0001.[19] If the only goal were to predict the average target site ATE, this $690 million measures the improved inference from randomly sampling a sufficiently large number of replication sites instead of allowing non-random site selection.

Predictions can also be made for each of the 101 later sites. Figure 3 compares the site-specific linear predictions from Equation (3) to each site's true ATE $\tilde{\tau}_s$. If all predictions were perfect, all dots would lie on the 45 degree line. Black dots vs. gray dots distinguish predictions that are vs. are not statistically different from the true $\tilde{\tau}_s$ with 90 percent confidence; the 22 non-significant

[19] See Appendix D for formal details on this test.

differences naturally tend to be closer to the 45 degree line. The graph has two key features. First, most of the sites are below the 45 degree line. This confirms that early site data systematically overpredict later ATEs and that this is not driven by any one particular site. Second, there is no correlation between predicted and actual ATEs, meaning that the adjustments on observable covariates are not correlated or perhaps negatively correlated with the site-specific heterogeneity. This echoes the result from Figure 2 that observables are not very informative about unobservables in this context.[20] Thus, the logic of inferring the direction and magnitude of bias from unobservables (Altonji, Elder, and Taber 2005) would not work well here.

6.2.3 Explaining the Prediction Failure

So far, I have documented a systematic failure of two simple approaches to predict efficacy in later sites. Does this happen due to a violation of external unconfoundedness, lack of overlap, or lack of knowledge of the full distribution of X in target sites?

Following Imbens and Rubin (2014), define a "normalized difference" for a single covariate as $\Delta = \frac{\overline{X}_{D=0} - \overline{X}_{D=1}}{\sqrt{S^2_{X,D=0} + S^2_{X,D=1}}}$, where $S^2_{X,D=d}$ is the variance of X in the population with $D = d$. Imbens and Rubin (2014) suggest that as a rule of thumb, linear regression methods tend to be sensitive to the specification when normalized differences are larger than 1/4. Although target variance $S^2_{X,D=0}$ is unknown, under the natural assumption that $S^2_{X,D=0} = S^2_{X,D=1}$, all but one of the 101 individual target sites in Figure 3 satisfy the $\Delta < \frac{1}{4}$ rule of thumb on both continuous variables in X^*. When predicting to the 101-site means, inspection of Table 4 shows that both continuous variables in X^* would easily satisfy this rule of thumb even under the most conservative assumption that $S^2_{X,D=0} = 0$.

Because the target distribution of X $f_{D=0}(x)$ is unobserved, I cannot test for overlap on continuous variables other than with this suggestive normalized difference test, and I must impose either Assumption 5 (linearity) or Assumption 6 (rescaled distributions). Appendix C tests whether prediction can be improved when $f_{D=0}(x)$ is known by predicting the ATE for each of the ten sites in the microdata, using the other nine sites as the "sample." Results show that predictions from the linear approach can be marginally improved (reducing root mean squared prediction error by 10-15 percent) by using a polynomial in X that also includes squares and interactions, and/or by predicting effects only for the target sub-population with improved overlap.

This marginal improvement should be interpreted with three caveats. First, while the within sample tests in Appendix C are informative about how well the approaches in this section control for individual-level observables, conditioning on X cannot address site selection bias due to individual-level unobservables or site-level observables that do not vary within the sample. Thus, even if

[20] This is not the only context in which individual-level observables are not very useful for prediction: Hotz, Imbens, and Mortimer (2005) similarly find that "once we separate the sample into those with and without recent employment experience, the results are remarkably insensitive to the inclusion of additional variables."

improved conditioning on X had dramatically improved prediction between the sample sites, it might still be impossible to predict the positive selection of early sites from later sites. Second, even if prediction can be improved by knowing $f_{D=0}(x)$, in applied settings it is not uncommon to only have an estimate of target means. In developing countries, for example, knowing $f_{D=0}(x)$ might require census microdata or researcher-conducted baseline surveys that do not always exist. Third, the predictiveness of observed covariates is in any event context-specific, so conditioning on observables might generate better or worse out-of-sample predictions in other contexts. The more basic implication of this section is that some adjustment is clearly necessary for the microdata to successfully predict impacts in later sites. As suggested by Heckman, Ichimura, Smith, and Todd (1998) and Smith and Todd (2005) in the context of individual-level selection bias, such adjustments might be possible, but only under particular conditions.

7 Metadata: Explaining Site Selection Bias

Why were Opower's first 10 sites positively selected from the full set of 111 sites? And is the current 111-site sample positively or negatively selected from the nationwide consumer population? In this section, I empirically test site selection mechanisms using site-level metadata. Building on the discussion in Section 4.2, I first separate within-utility vs. between-utility selection and then use utility-level covariates to test the hypothesized utility-level mechanisms.

7.1 Empirical Strategy

7.1.1 Cohort Trends and Within- vs. Between-Utility Selection

The microdata analysis compared the ten initial sites to all later sites, which exploits only a coarse binary measure of early vs. late selection. The metadata allow me to use M_s, the program start date for site s measured continuously in years, in the following continuous test:

$$\tilde{\tau}_s = \eta M_s + \kappa + \epsilon_s. \tag{5}$$

If η is positive (negative), this implies that earlier sites are negatively (positively) selected from all sample sites. As dependent variables, I use ATEs measured both in levels (kWh/day) and in percent. In some specifications, I also condition on individual-level observables by using the frequency- and X-adjusted ATE $\tilde{\tau}_s|X = \tilde{\tau}_s + \hat{\alpha}(\overline{X}^* - \overline{X}^*_s)$, where $\overline{X}$ is the mean of $\overline{X}_s$ across all 111 sites in the metadata. Using $\tilde{\tau}_s$ measures whether there is systematic site selection, while using $\tilde{\tau}_s|X$ measures site selection bias unexplained by observables. Of course, the results of Section 6 suggest that $\hat{\eta}$ will be negative when measuring $\tilde{\tau}_s$ in percent and that using $\tilde{\tau}_s|X$ will not make much difference.

I weight observations by analytic weights $1/\hat{Var}(\tilde{\tau}_s)$. This improves precision by weighting more heavily the $\tilde{\tau}_s$ which are more precisely estimated. I also present results using random effects meta-regression, which assumes that ϵ_s is the normally-distributed sum of variance from both unexplained site level heterogeneity and sampling error in $\tilde{\tau}_s$.

To isolate within-utility site selection mechanisms, I condition on utility and estimate the within-utility trend in ATEs. Denote ω_u as a vector of 58 indicator variables for utilities u. Within each utility, I number sites in order of start dates and define this integer variable as L_{su}. I estimate:

$$\tilde{\tau}_{su} = \lambda L_{su} + \omega_u + \epsilon_{su}. \tag{6}$$

In this equation, λ measures how treatment effects increase or decrease as utilities expand the program to additional households. The λ parameter should be interpreted carefully: utilities' decisions to expand the program were endogenous, and utilities that did not start additional sites may have expected a less favorable efficacy trend. This would cause λ to be larger (or less negative) than if all utilities proceeded with additional sites.

Section 4.2 hypothesized one systematic within-utility site selection mechanism, which is that utilities initially target higher-usage populations. If this mechanism dominates, then $\lambda < 0$, and including control group mean post-treatment usage C_s in Equation (6) should attenuate λ.

7.1.2 Testing Utility-Level Selection Mechanisms

The test of utility-level selection mechanisms is straightforward: does a variable that moderates selection also moderate treatment effects? I estimate both selection and outcome equations as a function of utility-level covariates Z_u that proxy for the selection mechanisms hypothesized in Section 4.2.

I assume that the utility-level selection decision D_u depends on a linear combination of Z_u plus a normally-distributed unobservable υ_u:

$$D_u = 1(\rho Z_u + \upsilon_u \geq 0) \tag{7}$$

I consider selection on two different margins. First, I consider selection into early partnership from the set of all partners with results in the metadata. Here, the first ten utilities have $D_u = 1$, while the remaining 48 partner utilities have $D_u = 0$. This type of selection could help explain why microdata from early sites overestimate later site ATEs. Second, I consider selection of the 58 current partner utilities from the target population of 882 utilities. This helps to assess how a nationally-scaled program might have different effects than observed so far.

To assess whether Z_u also moderates the treatment effect, I then estimate the outcome equation:

$$\tilde{\tau}_{su} = \theta Z_u + \lambda L_{su} + \zeta C_s + \kappa + \xi_{su} \tag{8}$$

This equation includes all 111 sites. L_{su} and C_s are included to control for within-utility selection mechanisms, and C_s is also a potential moderator of ATEs across utilities. If ρ and θ have the same sign for a given Z variable, that mechanism causes positive selection. If ρ and θ have opposite signs, that mechanism causes negative selection. Because Z_u vary only at the utility level, standard errors are clustered by utility. As in Equations (5) and (6), I weight observations by $1/\hat{Var}(\tilde{\tau}_{su})$ and also present robustness checks using random effects meta-regression.

7.2 Results

7.2.1 Cohort Trends and Within- vs. Between-Utility Selection

Table 6 presents trends for earlier vs. later sites. The top panel uses frequency-adjusted ATE $\tilde{\tau}$ normalized into percent, while the bottom panel uses $\tilde{\tau}$ in kWh/day. Column 1 presents the results of Equation (5), showing a statistically and economically significant decline in frequency-adjusted ATEs over time. Sites that start one year later average 0.173 percentage points (0.077 kWh/day) smaller ATEs.

Figure 4 illustrates this regression for $\tilde{\tau}_s$ in percent. Each of the first 11 sites had a frequency-adjusted ATE of 1.34 percent or larger. Sixty-seven of the next 100 sites had a smaller ATE than that. This further corroborates the results from Section 6 that extrapolating from early sites would overstate efficacy in later sites. An alternative specification in Appendix Table A12 using frequency- and X-adjusted ATE $\tilde{\tau}_s|X$ gives an almost identical results, which corroborates the result that the decline in efficacy cannot be explained by individual-level observables.

Column 2 presents estimates of Equation (6), which isolates within-utility trends. The regression excludes single-site utilities, so the sample size is 73 instead of 111. On average, a utility's next site performs 0.091 percentage points (0.062 kWh/day) worse than its previous site. Column 3 repeats column 2 but also conditions on control group mean usage C_s to test within-utility targeting of higher-usage households. As predicted, the within-utility trend attenuates toward zero, implying that some or all of the within-utility trend results from intentional decisions by utilities to initially target on gains.

Columns 4 and 5 focus on between-utility selection by adding L_{su} and C_s to Equation (5) as controls for within-utility selection. Both columns suggest that earlier utilities were positively selected from later utilities. Comparing column 4 to column 5 in the bottom panel suggests that controlling for usage makes the kWh/day ATE trend more negative. Additional regressions in Appendix Table A12 explain why. The program's expansion has taken it to increasingly high-usage utilities, such as those further away from temperate coastal areas. On top of that, the ratio of control group mean usage at utilities' first sites to utility-wide mean usage grows over time, implying that utilities are increasingly targeting heavy-usage customer subpopulations. In sum, the program has gradually expanded to utilities with consumers that are increasingly heavy users

but increasingly less responsive in percent terms.

Appendix Table A11 shows that the estimates in Table 6 are essentially identical when using random effects metaregression.

Could the systematic failure of external validity be due to time effects instead of cohort effects? In other words, is it possible that the same individuals exposed to treatment in 2008 and 2009 would have been less responsive in later years? While it is impossible to fully distinguish time effects from cohort effects in this setting, Appendix Table A13 presents a series of results suggesting that time effects are an unlikely explanation. First, the results in column 2 of Table 6 are robust to including site start date M_s, meaning that the within-utility trend is certainly not due to time effects. Second, likely forces that might drive time effects would be trends correlated with the Great Recession - for example, if a systematic decrease in average consumption left less scope for the treatment to further reduce consumption. However, the results are identical when controlling for percent changes in wages in counties served by utility u between 2007 and the first post-treatment year in site s, and results are also identical when controlling for analogously-calculated percent changes in utility average electricity usage. While there is substantial variation across sites in these two variables, neither is associated with treatment effects.[21] Furthermore, Allcott and Rogers (2014) show that effects tend to increase over a program's duration, which while certainly not dispositive, is also not suggestive of a negative time effect.

7.2.2 Testing Utility-Level Selection Mechanisms

Table 7 tests utility-level selection mechanisms. Columns 1 and 2 present the selection estimates from Equation (7), with column 1 studying selection of early partners from current partners, and column 2 studying selection of current partners from all potential partners. In most cases, the same mechanisms drive both early and overall partner selection, and the directions are as hypothesized in Section 4.2 and as documented in the unconditional comparisons in Table 3. Utilities with higher-income and more environmentalist populations are more likely partners, as are larger utilities. Furthermore, point estimates suggest that pre-existing energy efficiency programs are positively associated with selection, although this is not statistically significant.

Ownership structure, however, has different associations early (column 1) vs. overall (column 2). This is consistent with anecdotal evidence (Laskey 2014): initially, the program was unproven at scale, and the company relied on innovative and non-profit utilities for business. As initial RCTs

[21] Furthermore, this is not a spurious result of focusing only on electricity as the dependent variable: there is no trend in the proportion of "dual fuel" partner utilities that sell both electricity and gas, nor is there a trend in the share of homes using electric heat. Finally, there is no indication that the trend is driven by a lack of treatment fidelity. There is no statistically or economically significant time trend in treatment frequency (reports per month) for utilities' first sites, and the $\hat{\eta}$ coefficient estimate is almost exactly the same when not adjusting for treatment frequency. In fact, discussions with Opower managers suggest that the treatment may actually be improving over time due to a series of incremental changes. While this is difficult to quantify systematically, it only would strengthen the argument that the later populations would be less responsive to an exactly identical treatment.

gave positive results, and also as EERS policies expanded, the more conservative and heavily-regulated IOUs increasingly adopted the program.[22] Interestingly, anecdotal evidence suggests that very little of the early selection was based on intentionally targeting gains: both the company and potential partner utilities had little idea of whether the program would be at all feasible at scale, let alone how the effects would vary across utilities (Laskey 2014). Instead, this between-utility selection seems to have been based on other "unintentional" mechanisms.

Columns 3-6 present outcomes estimates from Equation (8). Effects are fairly consistent across columns and are as hypothesized. Because the specifications also condition on control mean usage C_{su}, higher utility mean usage implies that the sample of households used *less* electricity relative to others in the utility, which should decrease effects of the energy use comparison treatment. Higher income and environmentalist populations have larger treatment effects, while IOUs have smaller effects, perhaps due to lack of customer engagement. Municipally-owned utilities also have lower effects than the omitted ownership category (coops and other non-profits), but point estimates suggest larger effects than IOUs.

Point estimates suggest that effects are smaller at larger utilities. Appendix Table A14 presents alternative estimates of the outcome equations using random effects meta-regression, and the only substantive difference is that three of the four negative coefficients on ln(Residential Consumers) are statistically significant. As suggested in Section 4.2, large utilities seem to have lower efficacy largely because they are in urban areas, where neighbor comparisons might be less effective: Appendix Table A15 shows that the utility's urban population share is strongly negatively associated with treatment effects, and including this coefficient attenuates the negative coefficients on ln(Residential Consumers).

How much site selection is explained by individual-level and utility-level observables? Column 5 adds site start date M_s to Equation 8, which is also equivalent to adding the Z_u variables to column 5 of Table 6. Adding the Z_u variables attenuates the η coefficient on M_s from -0.175 to -0.122 percentage points per year, suggesting that site-level observables explain just under 1/3 of the decline in efficacy between earlier and later sites. Column 6 uses frequency- and X-adjusted ATE $\tilde{\tau}_s|X$. Point estimates are similar and suggest even greater selection on unobservables, consistent with previous results that individual-level observables are negatively correlated with unobservables driving earlier site selection.

Including utility-level data explains more of site selection than individual level data for two reasons. First, some of the selection is associated with factors that vary only at the site level, such as ownership and size. Second, site-level data better captures some population characteristics, as suggested by the stark case study of the Democrat vote share variable in Appendix B. In

[22]This is one case where Opower's sales outreach may have impacted selection independently of utilities' demand for the program. Laskey (2014) says that Opower's earliest sales outreach preferentially targeted municipally-owned and other non-profit utilities because he and others believed that IOUs would be less interested, but he believes they were wrong in retrospect given demand from IOUs reflected in column 2.

other words, part of the prediction failure with *individual-level* observables in Section 6 is due to extrapolating to sites with different *site-level* observables.[23] Unfortunately, corrections for site-level observables are not possible in a typical program evaluation without so many sites.

Variables that moderate both selection and outcomes suggest mechanisms of site selection bias. Population preferences and ownership structure are the two that are consistently statistically significant in estimates of both equations. Figure 5 presents simple graphical intuition, showing the unconditional relationship between the frequency-adjusted percent ATE and normalized population preferences for each utility's first site. The figure has two interesting features. First, while population preferences is normalized to mean zero across the 882 potential partner utilities, the sample mean is approximately one, implying strong selection. Second, the best fit line slopes upward, illustrating larger treatment effects at higher-income and environmentalist utilities. As the figure suggests, $\hat{\rho}$ and $\hat{\theta}$ have the same sign, suggesting that population preferences has caused positive selection from the 882 potential partner utilities nationwide. On the other hand, utility ownership structure generates negative selection from the 882 potential partners: current partners are more likely to be IOUs, and IOUs have conditionally lower ATEs.

What do these metadata results predict would be the first-year effects of a nationwide scaled program? To answer this, I use the outcome equation estimates (Equation (8)) to predict total effects across all consumers at all 882 potential partner utilities nationwide. To do this, I set control mean usage C_s equal to utility mean usage to reflect inclusion of all residential consumers, set within-utility start number L_{su} equal to the sample mean, and sum the predicted $\tilde{\tau}_s$, multiplying by each utility's number of residential consumers and then by the national average retail electricity price. The coefficient estimates in column 3 of Table 7 predict national first-year retail electricity cost savings of \$1.45 billion; the estimates in levels from column 4 predict a very similar \$1.42 billion. Of course, these predictions rely on the assumption that $\upsilon \perp \xi$, i.e. that no unobserved factors that moderate treatment effects affected selection of the 58 current partners from the set of 882 potential partners. Column 5 of Table 7 suggests that this may not be true: much of the downward trend in efficacy within the 111-site sample is unexplained by utility-level observables. Thus, even these predictions with a large sample of 111 sites may be biased due to unobserved site selection mechanisms.

By comparing these predictions to a prediction that does not adjust for any site-level observables, it is possible to test whether current sites are selected on observables. Assuming that the mean percent ATE from the 111 sample sites will hold for all consumers nationwide predicts first-year savings of \$1.80 billion, and assuming the sample mean ATE in levels predicts first-year savings of \$2.13 billion. These substantial overpredictions of \$350 million and \$710 million show that current

[23] One way to document that later sites differ on site-level observables is to fit "early site propensity scores" based on column 1 of Table 7. Sixty-one of the 101 later sites are outside the support of the scores for the first ten sites, meaning that they are different on site-level observables. When predicting site-specific ATEs from the ten-site microdata as in Figure 3, these 61 sites have larger absolute prediction errors.

sample sites are positively selected on observables from the nationwide population.

8 Conclusion

Replication is crucial for program evaluation because it gives a sense of the distribution of effects in different contexts. However, in the absence of randomly-selected evaluation sites, site-level selection mechanisms can generate a sample where program impacts differ systematically from target sites. The Opower energy conservation programs are a remarkable opportunity to study these issues, given a large sample of microdata with high-quality covariates plus results from 111 RCTs. There is evidence of both positive and negative selection mechanisms, involving both intentional targeting on gains (via within-utility targeting) and unintentional forces (such as population preferences and utility ownership and size). While the within-utility trend could have been predicted qualitatively with the knowledge that utilities initially target high-usage consumers, the large samples of rich individual-level microdata from the first ten sites do not even predict the *direction* of overall site-level selection, let alone its magnitude.

How can researchers address site selection bias? First, one might propose additional econometric approaches to control for observables. In the Opower example, however, econometric approaches are unhelpful, and no econometric approach can possibly work without external unconfoundedness. Second, researchers can continue efforts to replicate in sites that differ on hypothesized moderators. In the Opower example, however, this may not have been effective - there were ten replications in sites that did differ on potentially-relevant site level factors. Third, when reporting results, we can clearly define policy-relevant target populations and compare sample and target on observables, as in Tables 1, 2, and 3. While this can help to diagnose site selection bias, however, it does not solve the problem.

The only guaranteed solutions to site selection bias are "design-based" approaches. Just as the profession has de-prioritized econometric strategies in favor of randomized and quasi-randomized research designs to address individual-level selection bias, we may wish to further prioritize design-based strategies to address site-level selection bias. With a very large budget, a program could be evaluated in the entire population of sites to which it might be expanded, as in the Department of Labor YouthBuild evaluation and the Crepon *et al.* (2013) evaluation of job placement assistance in France. If only a few sites can be evaluated, sample sites can be randomly selected within strata of potentially-relevant site-level observables, as was originally envisioned for the JTPA evaluation. At least in the Opower example, which is an unparalleled empirical setting to study these issues, some form of randomized site selection would substantially improve inference.

References

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

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

[3] Allcott, Hunt, and Todd Rogers (2014). "The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation." *American Economic Review*, forthcoming.

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

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

[6] Angrist, Joshua, and Jorn-Steffen Pischke (2010). "The Credibility Revolution in Empirical Economics: How Better Research Design is Taking the Con out of Econometrics." *Journal of Economic Perspectives*, Vol. 24, No. 2 (Spring), pages 3-30.

[7] Ayres, Ian, Sophie Raseman, and Alice Shih (2013). "Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage." *Journal of Law, Economics, and Organization*, Vol. 29, No. 5 (October), pages 992-1022.

[8] Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng'ang'a, and Justin Sandefur (2013). "Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education." Working Paper, Goethe University Frankfurt (August).

[9] Chandra, Amitabh, and Douglas Staiger (2007). "Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks." Journal of Political Economy, Vol. 115, No. 1, pages 103-140.

[10] Costa, Dora, and Matthew Kahn (2013). "Energy Conservation "Nudges" and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment." *Journal of the European Economic Association*, Vol. 11, No. 3 (June), pages 680-702.

[11] 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).

[12] Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik (2008). "Nonparametric Tests for Treatment Effect Heterogeneity." *Review of Economics and Statistics*, Vol. 90, No. 3 (August), pages 389-405.

[13] 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.

[14] 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.

[15] Donabedian, Avedis (1988). "The Quality of Care: How Can It Be Assessed?" *Journal of the American Medical Association*, Vol. 260, No. 12, pages 1743-1748.

[16] EIA (U.S. Energy Information Administration) (2013). "Form EIA-861 data files." Available from http://www.eia.gov/electricity/data/eia861/

[17] ENERNOC (2013). "New Jersey Market Assessment, Opportunities for Energy Efficiency." White Paper (July).

[18] Harrison, Glenn, and John List (2004). "Field Experiments." *Journal of Economic Literature*, Vol. 42, No. 4 (December), pages 1009-1055.

[19] Heck, Stefan, and Humayun Tai (2013). "Sizing the Potential of Behavioral Energy-Efficiency Initiatives in the US Residential Market." White Paper, McKinsey & Company.

[20] 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.

[21] 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.

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

[23] 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.

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

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

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

[27] Hellerstein, Judith, and Guido Imbens (1999). "Imposing Moment Restrictions from Auxiliary Data by Weighting." *Review of Economics and Statistics*, Vol. 81, No 1 (February), pages 1-14.

[28] 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).

[29] 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.

[30] 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.

[31] Imbens, Guido, and Donald Rubin (2014). Causal Inference in Statistics and the Social Sciences. Cambridge and New York: Cambridge University Press.

[32] Integral Analytics (2012). "Sacramento Municipal Utility District Home Energy Report Program." http://www.integralanalytics.com/ia/Portals/0/FinalSMUDHERSEval2012v4.pdf

[33] JD Power (2014). "Electric Utility Residential Customer Satisfaction Study." Available from www.jdpower.com.

[34] Jones, Michael (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression." *Journal of the American Statistical Association*, Vol. 91, No. 433 (March), pages 222-230.

[35] KEMA (2012). "Puget Sound Energy's Home Energy Reports Program: Three Year Impact, Behavioral and Process Evaluation." Madison, WI: DNV KEMA Energy and Sustainability.

[36] KEMA (2013). "Update to the Colorado DSM Market Potential Assessment (Revised)." White Paper (June).

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

[38] Laskey, Alex (2014). Personal Communication. August, 2014.

[39] Leip, David (2013). "Dave Leip's Atlas of U.S. Presidential Elections." Available from http://uselectionatlas.org/

[40] 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.

[41] 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.

[42] Opinion Dynamics (2012). "Massachusetts Three Year Cross-Cutting Behavioral Program Evaluation Integrated Report." Waltham, MA: Opinion Dynamics Corporation.

[43] Perry, Michael, and Sarah Woehleke (2013). "Evaluation of Pacific Gas and Electric Company's Home Energy Report Initiative for the 2010-2012 Program." San Francisco, CA: Freeman, Sullivan, & Co.

[44] Pew Center (2011). "Energy Efficiency Standards and Targets." Accessed May 2011. Now available from http://www.c2es.org/us-states-regions/policy-maps/energy-efficiency-standards

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

[46] Quackenbush, John (2013). "Readying Michigan to Make Good Energy Decisions: Energy Efficiency." White Paper, Michigan Public Service Commission (October).

[47] 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.

[48] Rubin, Donald (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

[49] 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

[50] 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.

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

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

[53] Wennberg, David, F. L. Lucas, John Birkmeyer, Carl Bredenberg, and Elliott Fisher (1998). "Variation in Carotid Endarterectomy Mortality in the Medicare Population." *Journal of the American Medical Association*, Vol. 279, No. 16, pages 1278-1281.

Tables and Figures

Table 1: **Microfinance Institution Characteristics: RCT Partners and Non-Partners**

	(1)	(2)	(3)	(4)
	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)***
Percent of Portfolio at Risk	0.083	0.068	0.083	-0.015
	(0.120)	(0.066)	(0.121)	(0.012)
Percent Female Borrowers	0.62	0.69	0.62	0.07
	(0.27)	(0.27)	(0.27)	(0.05)
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)***
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)***
N	1903	35	1868	
F Test p-Value				0.00002***

Notes: The first three columns present the mean characteristics for all global MFIs, field experiment partners, and field experiment non-partners, respectively, with standard deviations 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 percent 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. "F Test p-Value" is from a regression of a partner indicator on all characteristics.

Table 2: **Hospital Characteristics: Clinical Trial Sites and Non-Trial Sites**

	(1)	(2)	(3)
	Population Mean	Difference: Drug Trial Sites - Other Hospitals	Difference: Procedure Trial Sites - Other Hospitals
County Percent with College Degree	0.23	0.09	0.08
	(0.10)	(0.00)***	(0.00)***
County Income per Capita	37.6	7.7	7.4
	(10.7)	(0.3)***	(0.4)***
In Urban Area	0.57	0.47	0.42
	(0.49)	(0.01)***	(0.01)***
Bed Count	179	238	256
	(214)	(7)***	(8)***
Annual Number of Admissions (000s)	7.4	11.0	11.9
	(9.6)	(0.3)***	(0.4)***
Annual Number of Surgeries (000s)	5.8	8.0	8.7
	(7.5)	(0.2)***	(0.3)***
Uses Electronic Medical Records	0.62	0.13	0.15
	(0.31)	(0.01)***	(0.01)***
U.S. News Technology Score	4.92	5.27	5.75
	(4.78)	(0.14)***	(0.16)***
U.S. News Patient Services Score	4.42	2.87	3.16
	(3.16)	(0.09)***	(0.10)***
Surgical Care Process Score	0.00	0.35	0.33
	(1.00)	(0.03)***	(0.03)***
Patient Communication Score	0.00	-0.36	-0.23
	(1.00)	(0.03)***	(0.03)***
Hospital-Acquired Condition Score	0.00	0.13	0.14
	(1.00)	(0.03)***	(0.03)***
Patient Safety Indicator Score	0.00	0.21	0.25
	(1.00)	(0.03)***	(0.04)***
Surgical Site Infections from Colorectal Surgery	0.00	-0.02	0.03
	(1.00)	(0.06)	(0.05)
Mortality Rate Score	0.00	-0.34	-0.37
	(1.00)	(0.03)***	(0.03)***
Ranked as U.S. News Top 50 Hospital	0.04	0.04	0.07
	(0.21)	(0.01)***	(0.01)***
Number of Specialties in U.S. News Top 50	0.20	0.17	0.29
	(1.25)	(0.04)***	(0.05)***
N	4653		
F Test p-Value		0.0000***	0.0000***

Notes: The first column presents the mean characteristic for all US hospitals, with standard deviations in parenthesis. The second and third columns present differences in means between clinical trial sites and non-trial sites, with robust standard errors in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively. 1722 hospitals have hosted drug trials, and 1265 have hosted procedure trials. "F Test p-Value" is from a regression of a trial site indicator on all characteristics.

Table 3: **Utility Characteristics: Opower Partners and Non-Partners**

	(1)	(2)	(3)	(4)
	All	Partners	Non-Partners	Difference
Utility Mean Usage (kWh/day)	34.7	28.3	35.2	-6.8
	(9.0)	(7.5)	(9.0)	(1.0)***
Mean Income ($000s)	50.2	59.0	49.6	9.4
	(10.1)	(9.7)	(9.9)	(1.3)***
Share College Grads	0.21	0.27	0.21	0.06
	(0.07)	(0.06)	(0.07)	(0.01)***
Hybrid Auto Share	0.0073	0.0112	0.0070	0.0042
	(0.0042)	(0.0050)	(0.0040)	(0.0007)***
Democrat Share	0.44	0.53	0.44	0.10
	(0.11)	(0.10)	(0.11)	(0.01)***
Green Party Share	0.0046	0.0052	0.0046	0.0007
	(0.0033)	(0.0028)	(0.0033)	(0.0004)*
Energy Efficiency Resource Standard	0.58	0.97	0.55	0.41
	(0.49)	(0.18)	(0.50)	(0.03)***
Green Pricing Share	0.0045	0.0100	0.0041	0.0059
	(0.0151)	(0.0187)	(0.0147)	(0.0025)**
Residential Conservation/Sales	0.0007	0.0035	0.0005	0.0029
	(0.0028)	(0.0063)	(0.0022)	(0.0008)***
Conservation Cost/Total Revenues	0.0027	0.0092	0.0022	0.0069
	(0.0065)	(0.0110)	(0.0058)	(0.0015)***
Municipally-Owned Utility	0.26	0.17	0.27	-0.10
	(0.44)	(0.38)	(0.44)	(0.05)*
Investor-Owned Utility	0.19	0.74	0.15	0.59
	(0.39)	(0.44)	(0.35)	(0.06)***
ln(Residential Customers)	10.5	12.8	10.4	2.5
	(1.3)	(1.3)	(1.1)	(0.2)***
N	882	58	824	
F Test p-Value				0.0000***

Notes: The first three columns of this table present the means of utility-level characteristics for all US 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 percent confidence, respectively. "F Test p-Value" is from a regression of a partner indicator on all characteristics.

Table 4: **Microdata Covariates**

	(1)	(2)	(3)	(4)	(5)
	Microdata Sample Size	Microdata Sample Mean	Microdata Sample Std. Dev.	National Mean	Later Sites Mean
First Comparison (kWh/day)	475,278	1.47	15.33	0.00	1.34
Tract Mean Income ($000s)	508,082	73.8	28.2	57.0	59.3
Tract Share College Grads	508,082	0.35	0.17	0.25	0.27
Tract Share Hybrid Autos	506,367	0.018	0.012	0.010	0.011
Green Pricing Participant	82,836	0.096	0.292	0.006	0.009
EE Program Participant	82,715	0.06	0.24	-	-
Electric Heat	313,076	0.12	0.35	0.34	0.28
House Age (Years)	407,469	41.5	27.7	39.2	41.2
Has Pool	207,885	0.18	0.35	0.17	0.17
Rent	272,308	0.10	0.32	0.33	0.33
Single Family	241,332	0.76	0.40	0.63	0.64
Square Feet (000s)	380,296	1.83	0.74	1.86	1.83

Notes: Columns 1, 2, and 3, respectively, present the observed sample sizes, means, and standard deviations of household characteristics in the ten-site microdata. Sample means and standard deviations are taken across the 25 imputations. The total microdata sample size is 508,295. Column 4 presents the national means, which are the means across the 882 potential partner utilities, weighted by number of residential consumers. Column 5 presents the unweighted mean across "later sites," the 101 more recent sites not included in the microdata. The Energy Efficiency Program Participant variable is not observed outside of the microdata.

Table 5: **Heterogeneous Treatment Effects**

	(1)	(2)	(3)	(4)	(5)
Treatment	1.707	1.762	1.760		1.760
	(0.056)***	(0.055)***	(0.058)***		(0.058)***
T x First Comparison		0.089	0.092	0.095	0.099
		(0.009)***	(0.009)***	(0.009)***	(0.013)***
T x Tract Mean Income		0.001			
		(0.003)			
T x Tract College Share		-0.516			
		(0.678)			
T x Tract Hybrid Auto Share		0.940			
		(7.437)			
T x Green Pricing		0.009			
		(0.235)			
T x EE Program Participant		0.064			
		(0.300)			
T x Electric Heat		1.125	1.196	1.015	1.314
		(0.229)***	(0.223)***	(0.224)***	(0.269)***
T x House Age		-0.000			
		(0.002)			
T x Has Pool		0.460	0.491	0.415	0.571
		(0.211)**	(0.208)**	(0.214)*	(0.224)**
T x Rent		-0.185			
		(0.252)			
T x Single Family		0.048			
		(0.221)			
T x Square Feet		0.498	0.494	0.652	0.565
		(0.128)***	(0.109)***	(0.121)***	(0.127)***
T x Baseline Usage					-0.010
					(0.013)
N	508,295	508,295	508,295	508,295	508,295
T x Site Indicators	No	No	No	Yes	No

Notes: This table presents estimates of Equation (2) with different X characteristics. The dependent variable is household i's post-treatment electricity use normalized by the site s control group post-treatment average. Missing data are imputed by multiple imputation. Robust standard errors, clustered at the level of randomization (household or block batch), are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table 6: **Cohort Trends and Within- vs. Between-Utility Selection**

Dependent Variable: Frequency-Adjusted ATE (Percent)

	(1)	(2)	(3)	(4)	(5)
Site Start Date (Years)	-0.173			-0.174	-0.175
	(0.032)***			(0.035)***	(0.035)***
Within-Utility Start Number		-0.091	-0.059	0.003	0.006
		(0.033)***	(0.028)**	(0.027)	(0.030)
Control Mean Usage (kWh/day)			0.017		0.001
			(0.004)***		(0.002)
R2	0.22	0.65	0.76	0.22	0.22
N	111	73	73	111	111

Dependent Variable: Frequency-Adjusted ATE (kWh/day)

	(1)	(2)	(3)	(4)	(5)
Site Start Date (Years)	-0.077			-0.044	-0.054
	(0.013)***			(0.013)***	(0.010)***
Within-Utility Start Number		-0.062	-0.002	-0.046	-0.000
		(0.016)***	(0.010)	(0.009)***	(0.007)
Control Mean Usage (kWh/day)			0.020		0.012
			(0.002)***		(0.001)***
R2	0.17	0.65	0.93	0.28	0.72
N	111	73	73	111	111
Utility Indicator Variables	No	Yes	Yes	No	No
Sample:	All Sites	Multi-Site Utilities	Multi-Site Utilities	All Sites	All Sites

Notes: Column 1 presents estimates of Equation (5), columns 2 and 3 present estimates of Equation (6), and columns 4 and 5 add controls to Equation (5). The average percent ATE is 1.31 percent, and the average ATE in levels is 0.47 kWh/day. Observations are weighted by inverse variance. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

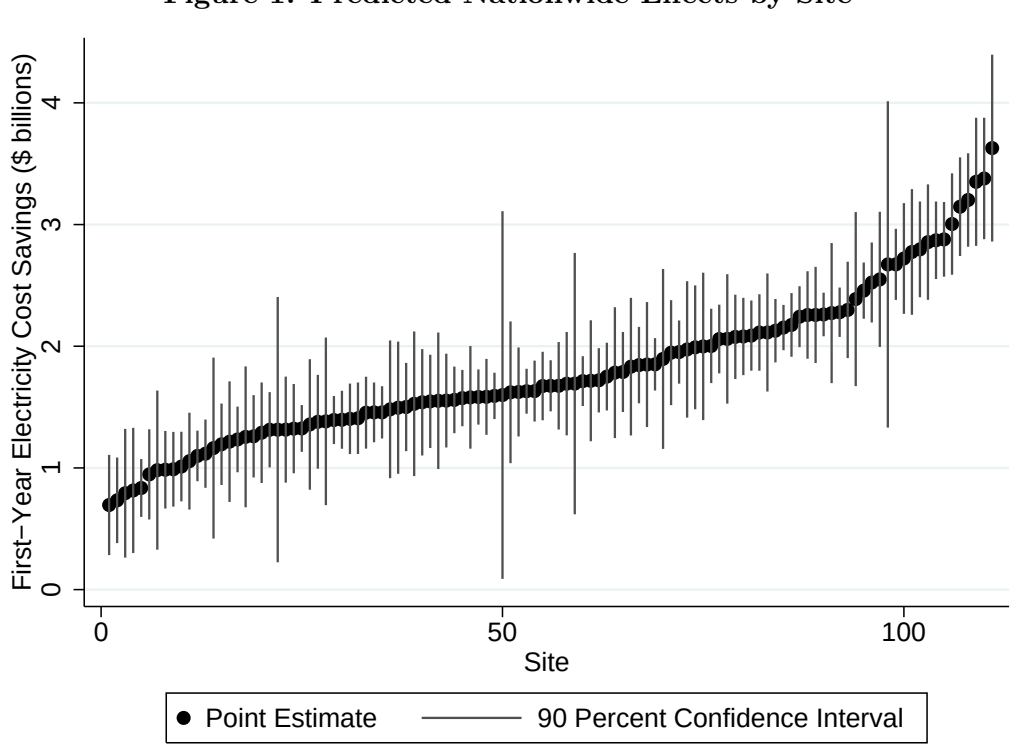
Table 7: **Utility-Level Selection**

	Selection		Outcomes			
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:	1(Early Partner)	1(Partner)	Frequency-Adjusted ATE (%)	Frequency-Adjusted ATE (kWh/day)	Frequency-Adjusted ATE (%)	Frequency- and X-Adjusted ATE (%)
Utility Mean Usage (kWh/day)	-0.068	-0.007	-0.040	-0.013	-0.038	-0.062
	(0.046)	(0.012)	(0.009)***	(0.003)***	(0.007)***	(0.008)***
Normalized Population Preferences	0.808	0.337	0.122	0.050	0.094	0.055
	(0.359)**	(0.091)***	(0.058)**	(0.019)**	(0.043)**	(0.044)
Normalized Other Programs	0.093	0.071	-0.002	-0.002	0.008	-0.010
	(0.144)	(0.059)	(0.011)	(0.004)	(0.011)	(0.013)
Municipally-Owned Utility	-2.145	0.405	-0.367	-0.188	-0.331	-0.635
	(1.109)*	(0.264)	(0.171)**	(0.046)***	(0.129)**	(0.202)***
Investor-Owned Utility	-3.695	0.557	-0.485	-0.185	-0.382	-0.381
	(1.111)***	(0.302)*	(0.175)***	(0.049)***	(0.149)**	(0.141)***
ln(Residential Customers)	0.541	0.494	-0.043	-0.013	-0.060	-0.084
	(0.440)	(0.078)***	(0.037)	(0.012)	(0.042)	(0.059)
Within-Utility Start Number			-0.070	-0.012	-0.031	-0.013
			(0.016)***	(0.004)***	(0.018)*	(0.023)
Control Mean Usage (kWh/day)			0.015	0.018	0.016	0.014
			(0.003)***	(0.001)***	(0.003)***	(0.003)***
Site Start Date (Years)					-0.122	-0.141
					(0.036)***	(0.042)***
Psuedo R2	0.43	0.44				
N	58	882	111	111	111	111
R2			0.47	0.80	0.56	0.60
Estimator:	Probit	Probit	OLS	OLS	OLS	OLS
Sample:	Partner Utilities	All Utilities	All Sites	All Sites	All Sites	All Sites

Notes: Columns 1 and 2 present estimates of Equation (7), while columns 3-6 present estimates of Equation (8). Observations are weighted by inverse variance. Robust standard errors, clustered by utility, are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

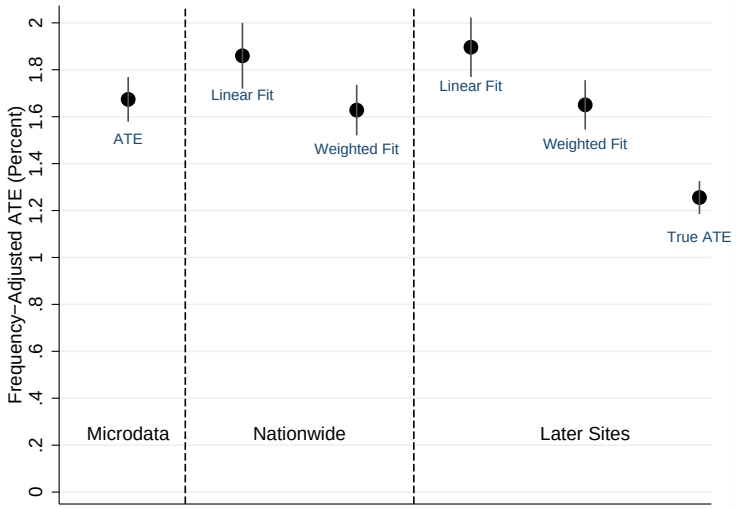
Figures

Figure 1: Predicted Nationwide Effects by Site



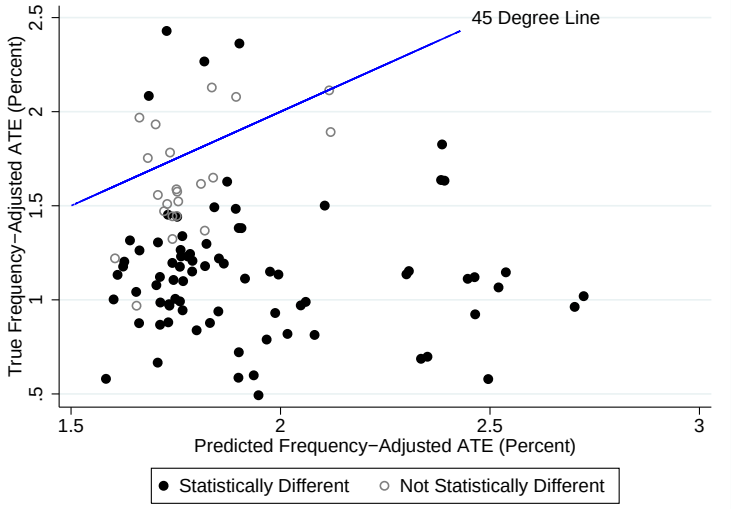
Notes: This figure presents the national electricity cost savings that would be predicted by extrapolating the percent average treatment effect from the first year of each Opower site to all US households.

Figure 2: Predicted Effects Using Microdata



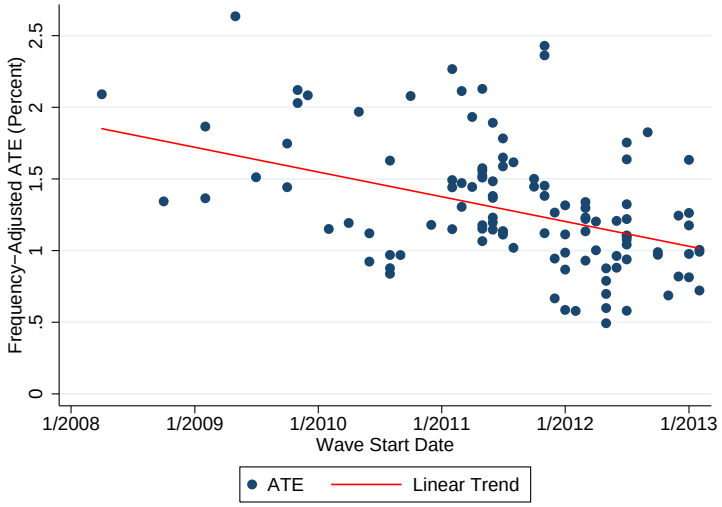
Notes: This figure presents the average treatment effects (ATEs) of the Opower program as predicted by microdata from the first ten sites. ATEs are "frequency adjusted" to match the average number of home energy reports per month in across the 111 sites in the metadata. The left panel is the sample ATE in the microdata, the middle panel is the nationwide prediction, and the right panel is the prediction for the 101 later sites that are in the metadata but not the microdata. The "True ATE" is the unweighted mean ATE for the later sites.

Figure 3: Site-Specific Predictions from Microdata



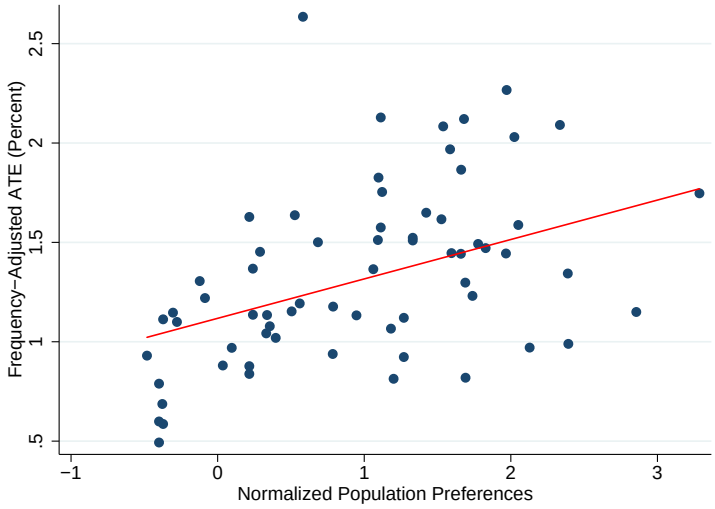
Notes: This figure plots the actual average treatment effects (ATEs) for each of the 101 later sites against a linear prediction from sample microdata using Equation (3). All ATEs are "frequency adjusted" to match the average number of home energy reports per month across the 111 sites in the metadata. "Statistically different" means that predicted and true ATEs differ with 90 percent confidence.

Figure 4: Efficacy Trend Across Sites



Notes: This figure plots the data and fitted regression line for column 1 of Table 6. In this regression, observations are weighted by inverse variance.

Figure 5: Site Selection on Population Preferences



Notes: This figure plots the regression of frequency-adjusted ATE on normalized population preferences, which is the sum of Income, Share College Grads, Hybrid Auto Share, Democrat Share, Green Party Share, Energy Efficiency Resource Standard, and Green Pricing Share, after normalizing each to mean zero, standard deviation one. In estimating the best fit line, observations are weighted by inverse variance.

Appendix: For Online Publication

Site Selection Bias in Program Evaluation

Hunt Allcott

A Data Appendix

A.1 Clinical Trial and Hospital Data

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

My hospital characteristics database combines three major data sources: the Center for Medicare & Medicaid Services (CMS) Provider of Services (POS) files for 2011 (CMS 2013a), the American Hospital Association (AHA) Annual Survey Database for 2011 (AHA 2012), and the CMS Hospital Compare database (CMS 2013b). Hospitals are linked between the databases using the six-digit CMS provider identification number.

From the POS files, I extract the hospital name, county, zip code, urban location indicator variable, and bed count.

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

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

Patient Communication Score=
Percent of patients who reported that their nurses "Always" communicated well
$+\frac{1}{2}\cdot$Percent of patients who reported that their nurses "Usually" communicated well
+Percent of patients who reported that their doctors "Always" communicated well
$+\frac{1}{2}\cdot$Percent of patients who reported that their doctors "Usually" communicated well
+Percent of patients who reported that staff "Always" explained about medicines
$+\frac{1}{2}\cdot$Percent of patients who reported that staff "Usually" explained about medicines
+Percent of patients who reported that they were given information about what to do during recovery

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

The next four variables from Hospital Compare were motivated directly from the Hospital Safety Score methodology, available from http://www.hospitalsafetyscore.org. The Surgical Care Process Score is the sum of five measures from the Surgical Care Improvement Project, which reports the percentage of times that surgeons at the hospital followed accepted practices, from giving prophylactic antibiotic within one hour

of surgical incision to giving appropriate venous thromboembolism. For each of the five specific measures, I normalized the percentages to have mean zero, standard deviation one across hospitals so as to not overweight variation coming from any one measure. I then summed the normalized measures and again normalized the sum to have mean zero, standard deviation one.

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

The Hospital Safety Score includes the incidence rates per 1000 discharges of four Hospital Acquired Conditions: foreign object retained after surgery, air embolism, pressure ulcers, and falls and trauma. Each of these individual rates is normalized to mean zero, standard deviation one. The Hospital Acquired Condition Score is the sum of these four normalized measures.

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

A.2 Opower Microdata

Appendix Table A1 provides an overview of Opower's first ten sites, which are the sites for which I have microdata. Due to confidentiality restrictions, utility names and locations are masked and the sites are numbered from one to ten. The rightmost column shows that treatment and control groups at nine sites are statistically balanced on baseline usage, while there is some imbalance at site 5. Placebo tests using pre-treatment data suggest that controlling for lagged electricity use eliminates the potential bias from this imbalance, and the overall results are effectively the same when excluding site 5, which is unsurprising given that it is only a small share of the ten-site sample.[24]

Appendix Table A2 presents the means and standard deviations of each variable at each specific site. Some variables are not available for all sites, and Green Pricing and EE Program Participant are only available in site 10.

Tract Mean Income and Share College Grads are mean household income and the share of population over 25 years old that holds a college degree, both from the 2000 Census. Tract Share Hybrid Autos uses vehicle registration data from 2013.

I note that in addition to the within-utility site selection processes discussed in the body of the paper, there is one additional element of within-utility site selection that is purely technical: to be eligible for the program, a customer must have at least one year of valid pre-experiment energy use data and satisfy some additional conditions. Typically, households in Opower's experimental populations need to have valid names and addresses, no negative electricity meter reads, at least one meter read in the last three months, no significant gaps in usage history, exactly one account per customer per location, and a sufficient number of neighbors to construct the neighbor comparisons. Households that have special medical rates or photovoltaic panels are sometimes also excluded. Utility staff and "VIPs" are sometimes automatically enrolled in the reports, and I exclude these non-randomized report recipients from any analysis. These technical exclusions eliminate only a small portion of the potential population. Such technical exclusions do not contribute to site selection bias, because the excluded households would never receive the program and are thus not part of a target population.

[24]Since these early programs, Opower has institutionalized a re-randomization algorithm to ensure covariate balance before implementation.

Table A1: **Microdata Experiment Overviews**

	(1)	(2)	(3)	(4)	(5)	(6)
Site	Region	Start Date	Households	Treated Households	Electricity Usage Obs.	Baseline Usage: Treatment-Control (Std. Error)
1	Midwest	July 2009	54,259	27,914	1,869,843	0.04 (0.05)
2	Midwest	January 2009	72,687	38,930	3,182,028	0.01 (0.12)
3	Mountain	October 2009	38,502	24,088	1,304,199	0.12 (0.14)
4	West	October 2009	33,308	23,766	568,395	0.09 (0.13)
5	Rural Midwest	April 2009	17,558	9,755	791,227	1.01 (0.42)
6	Northeast	September 2009	49,165	24,631	1,704,897	-0.21 (0.13)
7	West	October 2008	78,549	34,683	3,117,229	0.02 (0.10)
8	West	January 2009	42,576	9,367	1,667,334	0.26 (0.27)
9	West	September 2009	38,855	19,406	668,419	0.00 (0.17)
10	West	March 2008	82,836	33,651	6,388,135	-0.42 (0.58)
Combined		March 2008	508,295	246,191	21,261,706	

Notes: This table presents overviews of the first ten Opower sites, which are the sites for which microdata are available. Electricity Usage Observations includes all pre- and post-treatment data, including before the one-year baseline period and after the first post-treatment year. The rightmost column presents the treatment - control difference in baseline usage in kWh/day, with standard errors in parentheses.

Table A2: **Microdata Covariates by Site**

	Energy Use		**Census Tract**			**Household-Level**							
Site	Baseline Usage (kWh/day)	First Comparison (kWh/day)	Mean Income ($000s)	Share College Grads	Share Hybrid Autos	Green Pricing Participant	EE Program Participant	Electric Heat	House Age (Years)	Has Pool	Rent	Single Family	Square Feet (000s)
1	30.9	3.33	89.9	0.40	0.013	-	-	-	50.3	-	0.09	0.77	1.91
	(5.7)	(14.6)	(41.0)	(0.21)	(0.009)	-	-	-	(26.1)	-	(0.29)	(0.42)	(0.90)
2	29.7	0.00	70.2	0.21	0.007	-	-	0.08	31.7	-	-	0.96	1.69
	(16.4)	(18.5)	(12.9)	(0.08)	(0.003)	-	-	(0.27)	(28.1)	-	-	(0.21)	(0.54)
3	25.1	0.37	62.9	0.47	0.018	-	-	0.14	25.5	-	0.32	0.74	2.01
	(13.2)	(11.1)	(18.8)	(0.11)	(0.007)	-	-	(0.35)	(20.6)	-	(0.47)	(0.44)	(0.78)
4	18.2	1.33	63.7	0.34	0.022	-	-	-	59.2	0.10	0.35	0.50	1.69
	(10.7)	(9.9)	(27.5)	(0.11)	(0.009)	-	-	-	(23.1)	(0.30)	(0.48)	(0.50)	(0.72)
5	39.5	-2.45	45.3	0.16	0.004	-	-	0.31	-	-	0.05	-	1.28
	(27.5)	(24.2)	(6.0)	(0.05)	(0.002)	-	-	(0.46)	-	-	(0.21)	-	(0.54)
6	30.0	2.49	82.5	0.39	0.013	-	-	-	58.6	0.02	0.06	-	2.03
	(14.8)	(13.3)	(30.0)	(0.16)	(0.006)	-	-	-	(42.4)	(0.15)	(0.23)	-	(0.85)
7	30.5	2.88	85.4	0.40	0.024	-	-	0.07	31.0	-	0.03	-	2.14
	(13.8)	(15.4)	(30.7)	(0.16)	(0.012)	-	-	(0.26)	(16.0)	-	(0.18)	-	(0.64)
8	31.2	-1.13	65.3	0.25	0.019	-	-	-	28.0	0.24	-	0.62	1.88
	(22.5)	(13.9)	(31.2)	(0.09)	(0.008)	-	-	-	(15.7)	(0.43)	-	(0.49)	(0.80)
9	36.4	3.48	70.6	0.47	0.035	-	-	0.17	65.0	-	0.06	-	1.83
	(17.2)	(16.3)	(23.1)	(0.18)	(0.015)	-	-	(0.38)	(25.4)	-	(0.23)	-	(0.77)
10	30.8	0.98	70.9	0.36	0.020	0.09	0.06	0.25	37.4	0.21	-	-	1.75
	(15.1)	(14.1)	(17.2)	(0.14)	(0.010)	(0.29)	(0.24)	(0.44)	(18.3)	(0.41)	-	-	(0.60)

Notes: This table presents covariate means for the first ten Opower sites, with standard deviations in parenthesis. A dash means that a variable is unavailable at that site.

A.3 Site-Level Metadata

Appendix Table A3 presents descriptive statistics for the metadata. The I^2 statistic (Higgins and Thompson 2002) shows that 86.6 percent of the variation in percent ATEs is due to true heterogeneity instead of sampling variation. Effectively none of this variation is due to variation in reports per month: the standard deviation of frequency-adjusted ATEs and their mean standard error are 0.44 percent and 0.18 percent, respectively, and the I^2 is 85.6 percent.

I focus on the ATEs over each program's first year, for several reasons. Considering full instead of partial years averages over seasonal effect variation, whereas comparing programs that have been in effect over different seasons would require location-specific seasonal adjustments. Comparing programs that have been in effect for different durations would also require duration controls, given that effect sizes tend to grow over time (Allcott and Rogers 2014). This growth in effect sizes over time means that these first-year ATEs are smaller than the ATEs that are realized over longer treatment periods. I use one year instead of two or more full years because this allows the analysis to include the largest number of sites. In the 67 sites with two years of data, the first-year ATE is highly predictive of the ATE over the first two years ($R^2 = 0.79$).

Opower's analysts estimated the ATEs using mutually-agreed procedures and code. I define M_s as the month when the first home energy reports are generated in the site. The 12 months before M_s are the "baseline" period, while the "post-treatment" period begins the first day of the month after M_s. The month M_s is excluded from the analysis, as it often will include days both before and after the first reports arrive. Y_{it} is daily average electricity usage (in kilowatt-hours per day) for household i for the period ending with a meter read on date t. Y_{0i} is a vector of three baseline usage controls: average daily usage over the entire baseline period, the baseline winter (December-March), and the baseline summer (June-September). π_t is a set of indicators for the month and year in which t falls. The first-year ATE is estimated using the following equation:

$$Y_{it} = -\tau T_i + \gamma Y_{0i} + \pi_t + \varepsilon_{it} \tag{9}$$

The treatment causes energy use to decrease. By convention, I multiply τT by -1, so that reported τ are positive and larger values imply higher efficacy. Standard errors are robust and clustered by household.

Instead of estimating in levels, one alternative approach would be to use the natural log of Y as the dependent variable. However, regressing in logs and transforming to kWh levels tends to understate the quantity of energy conserved, because regressing in logs gives higher weight to lower-usage households with smaller effect sizes. Other practical reasons to prefer logs are less important in this context: there is very little measurement error because these are administrative records, and the estimated $\widehat{\tau}$ are not affected by dropping outlying high-usage observations.

Due to various contractual and computational issues, Opower has not been able to provide the clustered standard errors for 12 of the 111 sites. However, I observe non-clustered standard errors for all sites. For the sites where clustered standard errors are not available, I have predicted them based on a regression of clustered on non-clustered standard errors in the other sites. Because intra-household correlations of electricity use are similar across sites, the prediction has an R^2 of 0.87, so this approximation seems highly unlikely to affect the results.

As documented in Appendix Table A3, there are two types of attrition. First, an average of 10 percent of households move and close their utility accounts each year. The site with the highest one-year move rate (42 percent) is at a utility in college town where most households are rentals that change hands each academic year. After an account closes, Opower ceases to send reports and no longer observes electricity bills for the physical location or the former occupant, so the unit attrits from the sample.

The second type of attrition is when a household actively calls the utility and asks to opt out of the program. An average of 0.6 percent of households opt out during the first year. These households' utility bills are observed, and they remain in the sample. I define the "treatment" as "being mailed a Home Energy Report or opting out." This definition of "treatment" gives a treatment effect of policy interest: the effect of attempting to mail Home Energy Reports to an entire site-level population. In practice, because opt-out rates are so low, the ATE is the almost exactly the same when the "treatment" is defined as "being mailed a Home Energy Report."

Opower also works with utilities that sell only natural gas and other "dual fuel" utilities that sell both natural gas and electricity. Instead of studying effects on electricity use only, one alternative approach would be to combine effects on natural gas and electricity consumption. There are two reasons why I do not do this. First, there is no equivalent of the EIA form 861 database for natural gas utilities, so it would be difficult to construct a dataset with characteristics of potential partner natural gas utilities. Second, while the treatment presumably affects natural gas and oil use in all sites where households use these fuels, Opower only observes these effects if their partner utility is the company that sells the other fuels. In many sites, the natural gas and oil retailers are separate companies from the electricity retailer. I prefer a consistently-observed measure of the effects on electricity use instead of an inconsistently-observed measure of the effects on total energy use.

A.3.1 Site-Level Variation in Cost Effectiveness

In addition to the national-level effects illustrated in Figure 1, a second measure of economic significance is the variation in cost effectiveness, as presented in Figure A1. While there are many ways to calculate cost effectiveness, I present the simplest: the ratio of program cost to kilowatt-hours conserved during the first two years.[25] As Allcott and Rogers (2014) point out, cost effectiveness improves substantially when evaluating over longer time horizons; I use two years here to strike a balance between using longer time horizons to calculate more realistic levels vs. using shorter time horizons to include more sites with sufficient post-treatment data. I make a boilerplate cost assumption of $1 per report.

The variation is again quite substantial. The most cost effective (0.88 cents/kWh) is 14 times better than the least cost effective, and the 10th percentile is four times better than the 90th percentile. The site on the right of the figure with outlying poor cost effectiveness is a small program with extremely low ATE and high cost due to frequent reports.

This variation is economically significant in the sense that it can cause program adoption errors: managers at a target site might make the wrong decision if they extrapolate cost effectiveness from another site in order to decide whether to implement the program. Alternative energy conservation programs have been estimated to cost approximately five cents per kilowatt-hour (Arimura, Li, Newell, and Palmer 2011) or between 1.6 and 3.3 cents per kilowatt-hour (Friedrich *et al.* 2009). These three values are plotted as horizontal lines on Figure A1. Whether an Opower program at a new site has cost effectiveness at the lower or upper end of the range illustrated in Figure A1 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. The program is not cost effective at all sites: for example, one early partner utility ended a program due to poor cost effectiveness.

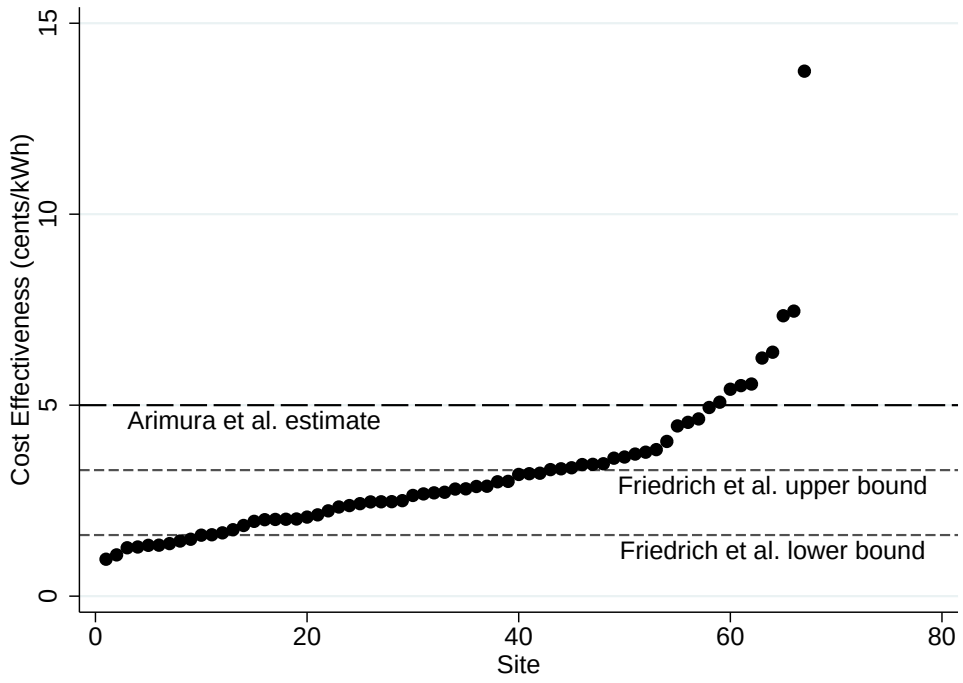
[25] Cost effectiveness would be further improved if natural gas savings were included. Of course, cost effectiveness is not a measure of welfare. The welfare effects of non-price interventions such as the Opower program are an important issue, but this is certainly distinct from this paper's argument about site selection bias.

Table A3: **Site-Level Metadata**

	Mean	Standard Deviation	Minimum	Maximum
Number of Households (000s)	77.2	70.4	5.8	435
Number of Treated Households (000s)	53.3	58.7	2.91	348
Reports/Month	0.58	0.11	0.21	1.03
Control Mean Usage (kWh/day)	36.2	14.9	12.0	90.1
Average Treatment Effect (kWh/day)	0.47	0.25	0.1	1.47
Standard Error (kWh/day)	0.062	0.032	0.017	0.19
Average Treatment Effect (Percent)	1.31	0.45	0.50	2.63
Standard Error (Percent)	0.18	0.095	0.079	0.66
Move Rate	0.10	0.059	0.018	0.42
Opt-Out Rate	0.006	0.004	0	0.032

Notes: This table presents descriptive statistics for the site-level Opower metadata. There are 111 sites at 58 different utilities.

Figure A1: Cost Effectiveness by Site



Notes: This figure presents the cost effectiveness over the first two years of each site against national benchmark estimates from Arimura *et al.* (2011) and Friedrich *et al.* (2009).

B The Democrat Share Variable

Political affiliation provides an interesting case study of challenges in estimating heterogeneous treatment effects when covariates that might moderate the treatment effect are not randomly assigned. In my ten-site microdata sample, the association between the treatment effect and Census tract Democrat vote share is not robust and is often negative.[26] A negative association is counter to the association in the site-level Opower metadata and also counter to results from other domains that environmentalism and Democrat party affiliation are positively correlated. In this appendix, I document these results and explain why they arise.

The tract-level Democrat share variable is the share (from zero to 1) of Democratic and Republican votes in the 2008 presidential elections that were for the Democratic candidate. Data are from the Harvard Election Data Archive.[27] The archive does not include voting data for the state where site 6 is located, so I substitute Democratic and Republican party registrations.

Across counties in the U.S., Democrat vote shares are positively associated with socioeconomic status (SES), as measured by variables such as income and education. Across Census tracts within Opower samples, however, Democrat share is negatively associated with SES. As shown in columns 1-5 of Appendix Table A4, Democratic Census tracts within cities use less electricity and are more urban, with lower income, less education, fewer single-family homes, and more renters. Furthermore, columns 6 and 7 the empirical association between measures of environmentalism and political ideology is ambiguous: consumers in Democratic Census tracts are more likely to participate in green pricing programs, but they are less likely to participate in the utility's other energy efficiency programs, conditional on income and education. Columns 6 and 7 restrict to Census tracts in site 10 because Green Pricing and EE Program Participant are only observed in that site. I observe households' Census block group in site 10 (only), and the results in column 6 and 7 are similar using block group-level data.

These correlations suggest two results. First, *within* a site, Democrat vote shares could likely be negatively correlated with environmentalism, and thus potentially negatively correlated with Opower treatment effects. Second, because Democrat share is correlated with other covariates, the association between this variable and the treatment effect may depend on what other covariates are included.

Appendix Table A5 reflects both of these suggested results. The table presents estimates of the heterogeneous treatment effect regression, Equation (2), also including Democrat vote share. Columns 1-3 show that Democratic neighborhoods have *smaller* treatment effects, both conditional on all other covariates and unconditional. However, simply controlling for the interaction between T and baseline usage Y_0 in column 4 eliminates this negative association. In columns 5 and 6, I limit the sample to Site 10 and use the block-group level Democratic vote shares. Column 5 replicates the approach in column 2, showing a negative but insignificant association. However, when I use $\ln Y$ as the dependent variable (multiplying by 100 to make the coefficients comparable) and condition on interactions between T and a particular set of other X variables, I can obtain a positive association between Democratic vote share and the treatment effect.

Because this within-site association is both not robust in my data and because a negative association is inconsistent with the between-site comparative static, I do not condition on Democrat share when using microdata for out-of-sample prediction in Section 6.

[26] Costa and Kahn (2013) show that the share of Democratic voters in a household's Census block group is positively associated with the treatment effect in one Opower site, conditional on interactions between T and some covariates. Their specification and available covariates differ from mine, and so this appendix is not a comment on their results. They present a series of regressions showing that their results are robust in their specifications at their site.

[27] This can be accessed at http://projects.iq.harvard.edu/eda/home.

Table A4: **Associations with Democrat Share**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable:	Baseline Usage	Mean Income	College Grads	Single Family	Rent	Green Pricing	EE Prog. Participant
Democrat Share	-14.475	-134.747	-0.367	-1.196	0.665	0.076	-0.051
	(5.097)**	(13.947)***	(0.111)***	(0.074)***	(0.073)***	(0.044)*	(0.014)***
Mean Income						-0.001	-0.000
						(0.000)*	(0.000)
Share College Grads						0.209	0.042
						(0.036)***	(0.015)***
R^2	0.11	0.28	0.06	0.32	0.25	0.51	0.26
Within-Site R2	0.11	0.28	0.06	0.32	0.25		
Between-Site R2	0.01	0.00	0.33	0.79	0.06		
N	1,386	1,385	1,385	715	1,117	85	85

Notes: This table presents associations between Democrat vote share and other variables, using data collapsed to Census tract-level averages. Columns 1-5 include all 10 microdata sites, while columns 6-7 include only site 10. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A5: **Heterogeneous Treatment Effect Estimates Including Democrat Share**

	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable:	y	y	y	y	y	100*ln(y)
Treatment	1.759				1.879	1.891
	(0.055)***				(0.170)***	(0.236)***
T x Tract Democrat Share	-2.912	-1.666	-1.606	-0.316	-0.679	2.772
	(0.572)***	(0.730)**	(0.583)***	(0.600)	(1.577)	(1.512)*
T x First Comparison	0.090	0.091			0.138	
	(0.009)***	(0.009)***			(0.017)***	
T x Tract Mean Income	-0.007	0.001			0.011	
	(0.003)**	(0.004)			(0.012)	
T x Tract College Share	0.075	-0.636			-2.033	0.113
	(0.690)	(0.789)			(1.885)	(1.123)
T x Tract Hybrid Auto Share	12.448	12.920			5.724	
	(7.590)	(10.635)			(23.610)	
T x Green Pricing	0.016	0.024			0.195	
	(0.234)	(0.233)			(0.348)	
T x EE Program Participant	0.047	0.049			0.194	
	(0.299)	(0.297)			(0.414)	
T x Electric Heat	1.094	0.914			2.267	
	(0.229)***	(0.236)***			(0.435)***	
T x House Age	0.004	0.002			0.006	-0.013
	(0.002)*	(0.003)			(0.012)	(0.013)
T x Has Pool	0.393	0.359			0.766	
	(0.211)*	(0.213)*			(0.360)**	
T x Rent	-0.182	-0.313				
	(0.253)	(0.263)				
T x Single Family	-0.130	-0.244				
	(0.228)	(0.248)				
T x Square Feet	0.510	0.546			0.899	
	(0.128)***	(0.135)***			(0.374)**	
T x Baseline Usage				0.069		0.029
				(0.011)***		(0.013)**
N	508,295	508,295	508,295	508,295	82,836	82,831
Sample Sites:	All	All	All	All	Site 10	Site 10
T x Site Indicators:	No	Yes	Yes	Yes	N/A	N/A

Notes: This table presents estimates of Equation (2) including Democrat vote share and other covariates. The dependent variable in columns 1-5 is y_{is}, household i's post-treatment electricity use normalized by the site s control group post-treatment average. In column 6, the dependent variable is $100 \cdot \ln y_{is}$. Missing data are imputed by multiple imputation. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

C Testing Alternative Extrapolation Approaches Within Sample

Prediction with microdata in Section 6 is limited by the fact that only the means of covariates X, not the full distributions, are known for the target populations. This appendix tests whether prediction is improved when the full distributions are known. To do this, I predict the average treatment effect for each of the ten sites in the microdata, using the other nine sites as the "sample." For each of the ten "target" sites, I compare the linear prediction approach from Section 6.1.2 to predictions using either a polynomial estimator or inverse probability weights. I also test whether improved prediction is possible for subsets of target populations that have improved overlap with sample data.

C.1 Procedure

I use five different approaches to predict treatment effects for each of the ten sites, using the nine other sites as the "sample."

The first approach is the unconditional prediction, i.e. simply the ATE for the nine sample sites estimated using Equation 2, without conditioning on any X covariates. The second is the frequency-adjusted linear prediction, which effectively combines Equations (3) and (4). Again denoting the target as $D = 0$ and the remaining sample of nine sites as $D = 1$, this is:

$$\hat{\tau}_{D=0} = \hat{\tau}_{D=1} + \hat{\alpha}(\overline{X}_{D=0} - \overline{X}_{D=1}) + \hat{\phi}(F_{D=0} - F_{D=1}). \tag{10}$$

As in the main estimates, $\bar{X}$ is the mean over all 25 imputations. The third approach is the same, except using only the X^* variables that survive the top-down procedure. This replicates the approach from Section 6.1.2.

The fourth approach relaxes Assumption 5 (that τ is linear in X) and instead uses a polynomial estimator. To determine the set of the variables that enter this prediction, I begin with all squares and interactions of all X variables and again use the Crump, Hotz, Imbens, and Mitnik (2008) top-down procedure to select covariates X^{**} that statistically significantly moderate the treatment effect. The polynomial prediction also uses Equation (10), except with X^{**} and the corresponding $\hat{\alpha}$.

The fifth approach is the inverse probability weight (IPW) estimator. Denote $e(x) = Pr(D_i = 1|X_i = x)$ as the "target propensity score," i.e. the probability that an individual drawn from the combined (sample and target) population is in target. I assume that $e(X_i^*) = \Phi(\xi X_i^* + \xi_0)$, where Φ is the normal cumulative density function and ξ_0 is a constant, and estimate with a probit. I then estimate the ATE with sample data, weighting observations by weights $w_i = \frac{\hat{e}(X_i^*)}{1-\hat{e}(X_i^*)}$. This ATE is then adjusted for the target-sample difference in frequency (reports/month) by adding $\hat{\phi}(F_{D=0} - F_{D=1})$.

To test whether improved predictions can be made for a subset of the target population with improved overlap, I trim the sample and target to observations with $0.1 \leq \hat{e}(X_i^*) \leq 0.9$. This parallels the rule of thumb developed in Crump, Hotz, Imbens, and Mitnik (2009) for overlap between treatment and control.[28] I then repeat each of the five approaches above, beginning with this trimmed sample.

C.2 Results

Table A6 presents estimates from the last iteration of the Crump, Hotz, Imbens, and Mitnik (2008) top-down procedure, after beginning with all squares and interactions of the X vector. Column 1 clusters at the level of randomization (household in sites 1-9 and block batch in site 10), while column 2 clusters by Census tract. Although some of the covariates vary only at the Census tract level, clustering standard errors by Census tract only slightly affects the standard errors, and all t-statistics are still larger than 2. The variables that survive are the same X^* from the linear procedure plus eight interactions. Notice that these eight

[28] The target population is on average 1/9 the size of the sample, which mechanically generates small target propensity scores. To address this, I re-weight observations in the probit estimator so that the weighted observation counts are identical in sample and target.

interactions can be grouped into four pairs, with each pair including two positively correlated variables with one positive and one negative coefficient.[29]

Table A7 presents prediction results. Row 1 show the mean and standard deviation of the ten ATEs in the ten "target" sites. Rows 2-6 present predictions from the full "samples" of nine sites to the full target population in the tenth. Rows 7-11 present predictions using trimmed sample and target populations. Columns 1 and 2 show the means and standard deviations of the ten predicted ATEs, while columns 3 and 4 show the average absolute value of the prediction errors and the root mean squared prediction error (RMSE), respectively.[30]

The linear prediction with all variables performs almost exactly the same as linear prediction with the top-down selected variables, as measured both by average absolute error and RMSE. The polynomial prediction performs slightly better than the linear predictions, and all perform slightly better after trimming to improve overlap.

When predicting the ATE in the full target population, the IPW estimator performs significantly worse than the other approaches. Inspection of the ten site-specific predictions shows that IPW predictions are very similar to the linear and polynomial predictions in eight sites. In two sites, however, there are a handful of observations with $\hat{e}(X_i^*)$ very close to one and thus very large weights w_i, which substantially reduces precision. These observations are trimmed in Row 11, and the IPW estimator performs approximately as well as the others.

Not only do all four approaches perform similarly across all sites, they give very similar predictions at each individual site: the predictions for trimmed populations in Rows 8-11 all have correlation coefficients greater than 0.87. The linear predictions with all X vs. with selected X^* are very similar for each site because the $\hat{\alpha}$ coefficients on the variables that do not survive the top-down procedure tend to be smaller. The polynomial and linear predictions are very similar for each site because the polynomial prediction simply adds the four pairs of nearly-offsetting adjustments discussed above.

The fact that the four approaches to conditioning on X give similar results suggests that they are correctly adjusting for observable population differences. However, conditioning on X does not improve predictions of target ATEs: the unconditional predictions perform substantially better. This implies that unobservable differences in populations or economic environments are negatively correlated with observable population differences within the ten sites in the microdata - just as they are when extrapolating from the ten-site microdata to the remaining 101 sites. Further inspection shows that this is not driven by any one "target" site or any one X covariate: there are four sites where the linear, polynomial, and IPW predictions tend to differ more from the true ATEs, and in these sites, no individual component of the sample-target ATE adjustment in Equation (10) tends to be substantially larger than the others.

The fact that linear prediction on X^* performs only slightly worse than polynomial prediction and/or trimming suggests that the failure in Section 6 to predict the decline of efficacy is likely not due to the inability to use more than the means of the target distribution of observables.

C.2.1 Prediction of Control Group Outcomes as a Suggestive Test of External Unconfoundedness

Hotz, Imbens, and Mortimer (2005) define "location unconfoundedness," $D_i \perp (Y_i(1), Y_i(0)) \,|X_i$, and show that it is a sufficient condition to extrapolate using Equation (1). What I call "external unconfoundedness," $D_i \perp (Y_i(1) - Y_i(0)) \,|X_i$ is conceptually similar, but it clarifies that only the *difference* in potential outcomes need be independent of assignment to sample vs. target for Equation (1) to hold. Hotz, Imbens, and Mortimer

[29]The first three pairs are interactions with Tract College Share and Tract Hybrid Auto Share, which are highly correlated - their association has a t-statistic of 30 when conditioning on site indicators and clustering by Census tract. The final pair is the interaction of Square Feet with House Age and the Rent indicator. These are also highly correlated, with a t-statistic of 16 conditional on site indicators.

[30]The mean of the target ATEs in Row 1 need not line up with the mean of predicted ATEs, as the former is an unweighted mean across sites, while the predictions will depend on the relationship of site ATEs to sample sizes and variances.

(2005) suggest that one potentially useful feature of the stronger location unconfoundedness assumption is that it motivates a suggestive test of external unconfoundedness that can be implemented in non-sample sites: $D_i \perp Y(0)|X_i$. To implement this, they test whether control group data from a sample site can predict untreated outcomes in a target site.

The framework of this appendix can be used for in-sample tests of this assumption. I focus on predicting control group usage in kWh/day levels (Y_{is}) instead of normalized usage $y_{is} = \frac{100Y_{is}}{C_s}$ because y is normalized to average exactly 100 for the control group within each of the sites, and it thus can be fitted mechanically. Using each of the nine-site "samples," I regress Y_{is} on X_i^* and baseline usage Y_{0i} for the combined control group and use the estimated coefficients to predict electricity usage $\hat{Y}_i(0)$ in the target site control group. I then calculate the mean prediction across all individuals in the target site, $E[\hat{Y}_i(0)|D_i = 0]$. Columns 1 and 2 of Row 14 in A7 present the mean and standard deviation of $E[\hat{Y}_i(0)|D_i = 0]$ across the ten sites. Columns 3 and 4 compare the $E[\hat{Y}_i(0)|D_i = 0]$ to the true control group means $E[Y_i(0)|D_i = 0]$.

On average, the prediction $E[\hat{Y}_i(0)|D_i = 0]$ differs from the true control group means $E[Y_i(0)|D_i = 0]$ by an absolute value of 0.94 kWh/day, or about three percent. This is substantially larger than the average absolute errors in fitting the treatment effect, which are between 0.4 and 0.5 percent. If the test is informative about external unconfoundedness, the prediction error for the ATE should be larger when the prediction error for control group usage is larger. However, across the ten predictions, there is no correlation between the two types of prediction errors, either in levels (p-value=0.837) or in absolute values (p-value=0.849).

In the Opower context, this test is empirically uninformative because it is not closely conceptually related to external unconfoundedness. Because weather is the most important determinant of annual average usage, this untreated outcomes prediction test is primarily a test of whether the weather changed differentially in sample vs. target. By contrast, there are many factors other than weather variation that cause ATEs to vary across sites.

C.2.2 Errors from Non-Experimental Estimates

In the Opower context, site selection bias resulted from selective adoption of a *new* program. In some other contexts such as the Job Training Partnership Act evaluations, site selection bias might have resulted from selective inability to experimentally evaluate an *existing* program. In these contexts, one potential interpretation of the site selection bias problem would be that researchers should lower the cost of evaluation by doing non-experimental studies instead of RCTs. If this allows more sites to select as research samples, this could reduce site selection bias. In other words, one might propose to sacrifice internal validity in favor of external validity. Would this approach be promising in the Opower setting?

Allcott (2011) evaluated the early Opower programs using the same microdata used in this paper, and developed two non-experimental estimators to compare to the RCT results. The first estimator used weather-adjusted pre-post treatment group differences. This estimator would be consistent if there were no other time-varying factors that affected electricity demand. The second was a difference-in-differences estimator using average consumption at other utilities in the same state as a control group. This would be consistent if there were no systematic changes at the control utilities relative to the treatment group counterfactual change.

Rows 15 and 16 present the means, standard deviations, and errors of these non-experimental estimators relative to the experimental estimates. On average, these approaches happen to overstate the true ATEs, and the average absolute errors and RMSEs are more than five times larger than the prediction errors from extrapolation in rows 2-11. Thus, within this ten-site sample, it would be much better to extrapolate RCT results from other sites than to use a non-experimental approach in-situ. Furthermore, the errors from non-experimental estimators are also large relative to the magnitude of site selection bias estimated in Section 6. Thus, these results do not support the argument that RCTs should be de-emphasized in order to improve external validity.

Table A6: **Heterogeneous Treatment Effects Estimates with Polynomial**

	(1)	(2)
T x First Comparison	0.084	0.084
	(0.008)***	(0.009)***
T x Electric Heat	1.131	1.131
	(0.220)***	(0.224)***
T x Has Pool	0.498	0.498
	(0.207)**	(0.207)**
T x Square Feet	0.425	0.425
	(0.105)***	(0.110)***
T x First Comparison x Tract College Share	-0.157	-0.157
	(0.064)**	(0.071)**
T x First Comparison x Tract Hybrid Auto Share	3.962	3.962
	(1.029)***	(1.154)***
T x Tract Mean Income x Tract College Share	-0.033	-0.033
	(0.012)***	(0.012)***
T x Tract Mean Income x Tract Hybrid Auto Share	0.446	0.446
	(0.165)***	(0.181)**
T x Single Family x Tract College Share	4.684	4.684
	(1.525)***	(1.568)***
T x Single Family x Tract Hybrid Auto Share	-75.727	-75.727
	(25.798)***	(26.823)***
T x Square Feet x House Age	-0.010	-0.010
	(0.004)**	(0.005)**
T x Square Feet x Rent	0.872	0.872
	(0.398)**	(0.397)**
N	508,295	508,295
Clustered by:	Level of Randomization	Census Tract

Notes: This table presents estimates of the final iteration of the Crump, Hotz, Imbens, and Mitnik (2008) top-down procedure after beginning with all squares and interactions of X characteristics. The dependent variable is household i's post-treatment electricity use normalized by the site s control group post-treatment average. Missing data are imputed by multiple imputation. Robust standard errors are in parenthesis; Column 1 clusters at the level of randomization (household in sites 1-9 and block batch in site 10), while column 2 clusters by Census tract. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A7: **Within-Sample Prediction**

		(1)	(2)	(3)	(4)
Row	Estimate	Mean	Standard Deviation	Average Absolute Error	Root Mean Squared Error
1	True target ATEs (%)	1.74	0.40	-	-
	Predicted ATEs with full populations (%)				
2	Unconditional	1.71	0.04	0.33	0.41
3	Linear (All variables)	1.64	0.28	0.46	0.57
4	Linear (Top-down variables)	1.64	0.29	0.46	0.58
5	Polynomial	1.66	0.30	0.41	0.53
6	Inverse Probability Weighting	2.75	2.75	1.55	2.91
	Predicted ATEs with trimmed populations (%)				
7	Unconditional	1.69	0.05	0.31	0.40
8	Linear (All variables)	1.64	0.27	0.42	0.51
9	Linear (Top-down variables)	1.65	0.27	0.42	0.52
10	Polynomial	1.65	0.29	0.39	0.48
11	Inverse Probability Weighting	1.67	0.30	0.44	0.54
12	Share of target in trimmed population	0.94	0.04	-	-
13	Share of sample in trimmed population	0.95	0.03	-	-
14	Prediction of control usage (kWh/day)	29.55	5.12	0.94	1.20
	Error from non-experimental estimates (%)				
15	Pre-post differences	3.15	2.28	2.08	2.36
16	Diff-in-diff against statewide control	3.75	3.51	2.98	3.65

Notes: This table presents results from extrapolating to each of the ten "target" sites in the microdata from a sample comprised of the other nine sites. The bottom panel uses non-experimental results presented in Allcott (2011).

D Testing Equality of Predicted and Actual ATEs

Section 6 predicts unweighted mean treatment effects for the 101 later sites and compares this to the true value. This appendix presents the test statistic of equality of these two estimates. Denote the 101-site target as $D = 0$ and the ten-site sample as $D = 1$, and further denote $\sum_s (1 - D_s) = N_0 = 101$ as the number of sites in the target.

The unweighted mean ATE is simply:

$$\overline{\hat{\tau}^{true}_{D=0}} = \frac{\sum_s (1 - D_s)\hat{\tau}^{true}_s}{N_0}. \tag{11}$$

Its variance is

$$\hat{Var}(\overline{\hat{\tau}^{true}_{D=0}}) = \frac{\sum_s (1 - D_s)\hat{Var}(\hat{\tau}^{true}_s)}{N_0^2}, \tag{12}$$

where $\hat{\tau}^{true}_s$ and $\hat{Var}(\hat{\tau}^{true}_s)$ are from the metadata, as estimated in Equation (9) in Appendix A.

The prediction of the re-weighting estimator $\hat{\tau}_{D=0}$ is the ATE from the reweighting estimator $\hat{\tau}^{rw}_{D=1}$ plus a frequency adjustment analogous to Equation (4):

$$\hat{\tau}_{D=0} = \hat{\tau}^{rw}_{D=1} + \hat{\phi}(\bar{F}_{D=0} - \bar{F}_{D=1}). \tag{13}$$

Its variance is

$$\hat{Var}(\hat{\tau}_{D=0}) = \hat{Var}(\hat{\tau}^{rw}_{D=1}) + \hat{Var}(\hat{\phi}) \cdot (\bar{F}_{D=0} - \bar{F}_{D=1})^2, \tag{14}$$

where $\hat{\phi}$ and $\hat{Var}(\hat{\phi})$ are as presented in column 3 of Appendix Table A10. In these equations, $\bar{F}_{D=0}$ is the unweighted mean frequency (reports/month) across target sites, while $\bar{F}_{D=1}$ is the sample mean frequency. Equation (14), like Equation (16) below, uses the assumption that the additive terms in the prediction have zero covariance.

The linear prediction is as in Equation (10):

$$\hat{\tau}_{D=0} = \hat{\tau}_{D=1} + \hat{\alpha}(\bar{X}^*_{D=0} - \bar{X}^*_{D=1}) + \hat{\phi}(\bar{F}_{D=0} - \bar{F}_{D=1}). \tag{15}$$

Its variance is

$$\hat{Var}(\hat{\tau}_{D=0}) = \hat{Var}(\hat{\tau}_{D=1}) + (\bar{X}^*_{D=0} - \bar{X}^*_{D=1})'\hat{Var}(\hat{\alpha})(\bar{X}^*_{D=0} - \bar{X}^*_{D=1}) + \hat{Var}(\hat{\phi}) \cdot (\bar{F}_{D=0} - \bar{F}_{D=1})^2. \tag{16}$$

In this equation, $\bar{X}^*_{D=0}$ is the column vector of unweighted means of $\bar{X}_s$ across target sites, while $\bar{X}^*_{D=1}$ is the vector of sample means of X^*.

For either the re-weighted or linear prediction, the prediction error is:

$$\hat{\Omega} = \overline{\hat{\tau}^{true}_{D=0}} - \hat{\tau}_{D=0}. \tag{17}$$

,

The variance of the prediction error is

$$\hat{Var}(\hat{\Omega}) = \hat{Var}(\overline{\hat{\tau}^{true}_{D=0}}) + \hat{Var}(\hat{\tau}_{D=0}). \tag{18}$$

The test statistic is:

$$t = \frac{\hat{\Omega}}{\sqrt{\hat{Var}(\hat{\Omega})}}. \tag{19}$$

In large samples, this test statistic has a standard normal distribution under the null hypothesis of no prediction error.

E Additional Tables and Figures Referenced in Paper

Table A8: **Site-Specific Heterogeneous Effects**

Site:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Treatment	1.874	1.214	2.184	3.658	3.538	2.076	1.300	0.201	0.626	2.206
	(0.272)***	(0.584)**	(0.874)**	(0.568)***	(2.147)*	(0.386)***	(0.196)***	(0.782)	(0.491)	(0.186)***
T x First Comparison	0.016	0.082	0.036	0.100	0.106	0.054	0.112	-0.022	0.283	0.128
	(0.020)	(0.015)***	(0.043)	(0.049)**	(0.023)***	(0.021)**	(0.015)***	(0.089)	(0.059)***	(0.015)***
T x Tract Mean Income	-0.009	0.029	0.030	0.069	-0.103	0.004	-0.002	0.021	0.015	0.018
	(0.007)	(0.014)**	(0.024)	(0.024)***	(0.068)	(0.013)	(0.008)	(0.016)	(0.015)	(0.015)
T x Tract College Share	-1.408	1.949	2.191	-8.542	8.674	-2.024	-0.376	-9.558	0.339	-4.293
	(1.701)	(2.911)	(4.018)	(4.938)*	(8.961)	(2.907)	(1.852)	(5.810)*	(2.179)	(1.861)**
T x Tract Hybrid Auto Share	79.481	-76.221	-103.228	-102.442	110.435	73.214	-9.128	26.078	12.382	25.797
	(30.310)***	(73.551)	(59.828)*	(62.835)	(199.330)	(47.422)	(26.378)	(63.565)	(28.710)	(24.532)
T x House Age	-0.001	0.007	0.015	0.002		-0.006	-0.006	-0.014	0.005	0.006
	(0.008)	(0.005)	(0.017)	(0.012)		(0.004)	(0.009)	(0.019)	(0.009)	(0.013)
T x Rent	0.331		0.035	-0.793	1.441	-1.139	0.380		-1.349	
	(0.761)		(0.714)	(0.648)	(1.705)	(0.773)	(0.918)		(1.102)	
T x Single Family	0.284	1.245	1.378	-0.202				0.216		
	(0.493)	(0.495)**	(0.792)*	(0.683)				(0.606)		
T x Square Feet	0.296	0.405	0.599	0.890	1.254	-0.000	0.591	0.438	0.410	0.890
	(0.344)	(0.370)	(0.431)	(0.521)*	(1.442)	(0.259)	(0.257)**	(0.796)	(0.335)	(0.343)**
T x Electric Heat		-2.608	1.652		2.014		0.925		1.197	1.974
		(0.749)***	(1.172)		(1.103)*		(0.564)		(0.658)*	(0.436)***
T x Has Pool				2.385		0.365		0.962		0.734
				(0.989)**		(0.921)		(1.207)		(0.372)*
T x Green Pricing										0.202
										(0.352)
T x EE Program Participant										0.171
										(0.420)
N	54,259	72,687	38,502	33,308	17,558	49,165	78,549	42,576	38,855	82,836

Notes: This table presents estimates of Equation (2) for each individual site in the microdata. The dependent variable is Y_{is}, household i's post-treatment electricity use normalized by the site s control group post-treatment average. Missing data are imputed by multiple imputation. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A9: **Empirical Likelihood Results for the Re-Weighting Estimator**

	(1)	(2)
Target:	National	Later Sites
First Comparison	0.005	0.003
	(0.001)***	(0.001)***
Electric Heat	-0.598	-0.542
	(0.057)***	(0.041)***
Has Pool	-0.008	0.038
	(0.049)	(0.020)*
Square Feet	0.022	0.026
	(0.024)	(0.019)
N	101,961	101,961

Notes: This table presents the empirical likelihood results used to re-weight the microdata. Column 1 presents the estimates to match national average characteristics, while column 2 presents estimates to match the average characteristics in the 101 later sites. To facilitate convergence, estimation was carried out on a randomly-selected 20 percent subsample. Missing data are imputed by multiple imputation. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A10: **Adjustment for Treatment Frequency**

	(1)	(2)	(3)
	Site 2	Site 7	Both
T x Reports/Month	0.489	0.554	0.517
	(0.283)*	(0.309)*	(0.209)**
Treatment x Site 2	1.465		1.445
	(0.248)***		(0.204)***
Treatment x Site 7		1.071	1.101
		(0.279)***	(0.209)***
R2	0.89	0.86	0.88
N	72,687	78,549	151,236
T x Site Indicators	N/A	N/A	Yes

Notes: This table presents estimates of the frequency adjustment parameter used in Equation (4). The estimating equation is Equation (2), using the number of reports per month as the only X covariate. The dependent variable is Y_{is}, household i's post-treatment electricity use normalized by the site s control group post-treatment average. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A11: **Cohort Trends Using Random Effects Meta-Regression**

Dependent Variable: Percent Effect

	(1)	(2)	(3)	(4)	(5)
Site Start Date (Years)	-0.194			-0.194	-0.202
	(0.035)***			(0.037)***	(0.038)***
Within-Utility Start Number		-0.090	-0.060	0.000	0.009
		(0.032)***	(0.028)**	(0.030)	(0.032)
Control Mean Usage (kWh/day)			0.016		0.002
			(0.004)***		(0.003)
I2	0.82	0.69	0.56	0.82	0.82
N	111	73	73	111	111

Dependent Variable: kWh/day Effect

	(1)	(2)	(3)	(4)	(5)
Site Start Date (Years)	-0.037			-0.021	-0.069
	(0.022)*			(0.023)	(0.015)***
Within-Utility Start Number		-0.046	-0.003	-0.042	0.014
		(0.022)**	(0.009)	(0.019)**	(0.012)
Control Mean Usage (kWh/day)			0.020		0.013
			(0.001)***		(0.001)***
I2	0.94	0.91	0.56	0.93	0.82
N	111	73	73	111	111
Utility Indicator Variables	No	Yes	Yes	No	No
Sample:	All Sites	Multi-Site Utilities	Multi-Site Utilities	All Sites	All Sites

Notes: This table parallels Table 6 using random effects meta-regression. Column 1 presents estimates of Equation (5), columns 2 and 3 present estimates of Equation (6), and columns 4 and 5 add controls to Equation (5). The average percent ATE is 1.31 percent, and the average ATE in levels is 0.47 kWh/day. Standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A12: **Cohort Trends: Additional Specifications**

	(1)	(2)	(3)	(4)
Dependent Variable:	Frequency- and X-Adjusted ATE (%)	Utility Mean Usage (kWh/day)	Control Mean Usage (kWh/day)	Control/ Utility Mean Usage
Site Start Date (Years)	-0.194	2.566	5.066	0.057
	(0.039)***	(0.742)***	(1.014)***	(0.029)*
R2	0.19	0.16	0.22	0.05
N	111	58	58	58
Sample:	All Sites	Utilities' 1st Sites	Utilities' 1st Sites	Utilities' 1st Sites

Notes: This table presents alternative estimates of Equation (5). The dependent variable in Column 4 is the ratio of control group mean usage to utility mean usage. Observations in column 1 are weighted by inverse variance. Standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A13: **Cohort Trends vs. Time Trends**

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Dependent Variable:	Frequency-Adjusted ATE	Frequency-Adjusted ATE	Frequency-Adjusted ATE	1(Dual-Fuel Utility)	Share Electric Heat	Average Reports/ Month	Frequency-Unadjusted ATE
Site Start Date (Years)	0.067	-0.178	-0.178	0.015	0.017	-0.022	-0.185
	(0.085)	(0.045)***	(0.032)***	(0.057)	(0.014)	(0.016)	(0.031)***
Within-Utility Start Number	-0.137						
	(0.064)**						
Total Wages Change		0.162					
		(0.941)					
Electricity Usage Change			1.089				
			(0.809)				
R2	0.65	0.22	0.26	0.00	0.01	0.06	0.24
N	73	111	107	111	111	66	111
Utility Indicator Variables	Yes	No	No	No	No	No	No
Sample:	Multi-Site Utilities	All Utilities	All Utilities	All Utilities	All Utilities	Utilities' 1st Sites	All Utilities

Notes: This table presents alternative estimates of Equations (5) and (6). Total Wages Change is the percent change in total wages in counties served by utility u between 2007 and the first post-treatment year for site s, using data from the Quarterly Census of Employment and Wages (BLS 2014). Electricity Usage Change is the percent change in residential electricity use for utility u between 2007 and the first post-treatment year for site s, using data from EIA (2014). Observations are weighted by inverse variance. Standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A14: **Utility-Level Outcome Equation Using Random Effects Meta-Regression**

	Outcomes			
	(1)	(2)	(3)	(4)
Dependent Variable:	Frequency-Adjusted ATE (%)	Frequency-Adjusted ATE (kWh/day)	Frequency-Adjusted ATE (%)	Frequency- and X-Adjusted ATE (%)
Utility Mean Usage (kWh/day)	-0.040	-0.014	-0.038	-0.061
	(0.007)***	(0.003)***	(0.007)***	(0.008)***
Normalized Population Preferences	0.151	0.055	0.124	0.104
	(0.046)***	(0.018)***	(0.043)***	(0.054)*
Normalized Other Programs	0.008	0.000	0.011	-0.011
	(0.014)	(0.005)	(0.013)	(0.017)
Municipally-Owned Utility	-0.442	-0.200	-0.387	-0.694
	(0.156)***	(0.060)***	(0.147)***	(0.181)***
Investor-Owned Utility	-0.471	-0.182	-0.372	-0.374
	(0.135)***	(0.054)***	(0.129)***	(0.160)**
ln(Residential Customers)	-0.070	-0.019	-0.079	-0.093
	(0.042)*	(0.016)	(0.039)**	(0.048)*
Within-Utility Start Number	-0.072	-0.013	-0.033	-0.018
	(0.027)***	(0.010)	(0.027)	(0.034)
Control Mean Usage (kWh/day)	0.015	0.018	0.016	0.013
	(0.003)***	(0.001)***	(0.003)***	(0.004)***
Site Start Date (Years)			-0.115	-0.132
			(0.031)***	(0.038)***
I2	0.75	0.75	0.70	0.75
N	111	111	111	111

Notes: This table presents estimates of Equation (8). The table parallels columns 3-6 of Table 7 using random effects meta-regression. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table A15: **Utility Size as a Proxy for Urban Households**

	(1)	(2)	(3)	(4)
Utility Mean Usage (kWh/day)	-0.040	-0.043	-0.040	-0.042
	(0.009)***	(0.009)***	(0.007)***	(0.007)***
Normalized Population Preferences	0.122	0.206	0.151	0.240
	(0.058)**	(0.060)***	(0.046)***	(0.051)***
Normalized Other Programs	-0.002	0.002	0.008	0.012
	(0.011)	(0.013)	(0.014)	(0.014)
Municipally-Owned Utility	-0.367	-0.278	-0.442	-0.336
	(0.171)**	(0.168)	(0.156)***	(0.153)**
Investor-Owned Utility	-0.485	-0.533	-0.471	-0.505
	(0.175)***	(0.165)***	(0.135)***	(0.130)***
ln(Residential Customers)	-0.043	-0.019	-0.070	-0.044
	(0.037)	(0.040)	(0.042)*	(0.041)
Within-Utility Start Number	-0.070	-0.060	-0.072	-0.060
	(0.016)***	(0.015)***	(0.027)***	(0.026)**
Control Mean Usage (kWh/day)	0.015	0.017	0.015	0.016
	(0.003)***	(0.003)***	(0.003)***	(0.003)***
Share Urban		-1.121		-1.228
		(0.418)***		(0.361)***
R2	0.47	0.51		
N	111	111	111	111
I2			0.75	0.73
Specification:	OLS	OLS	Random Effects Meta-Reg	Random Effects Meta-Reg

Notes: This table presents estimates of Equation (8), with the addition of Share Urban. Share Urban is the share of the population in counties in utility u's service territory that is in urbanized areas and urban places, using data from the 2010 Census (NHGIS 2013). In the "OLS" specifications, observations are weighted by inverse variance and standard errors are clustered by utility. Robust standard errors are in parenthesis. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Figure A2: Home Energy Report: Front Page

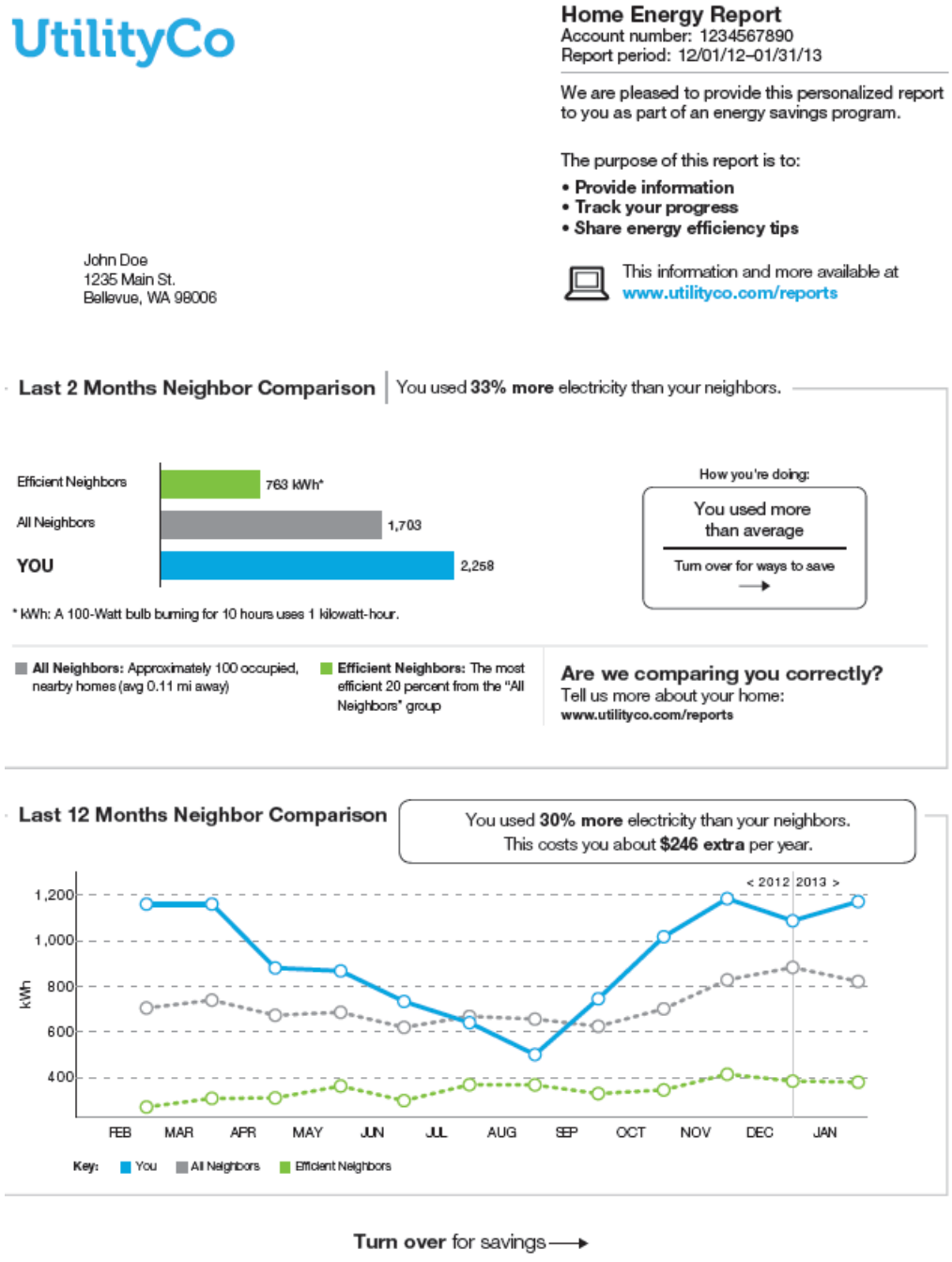
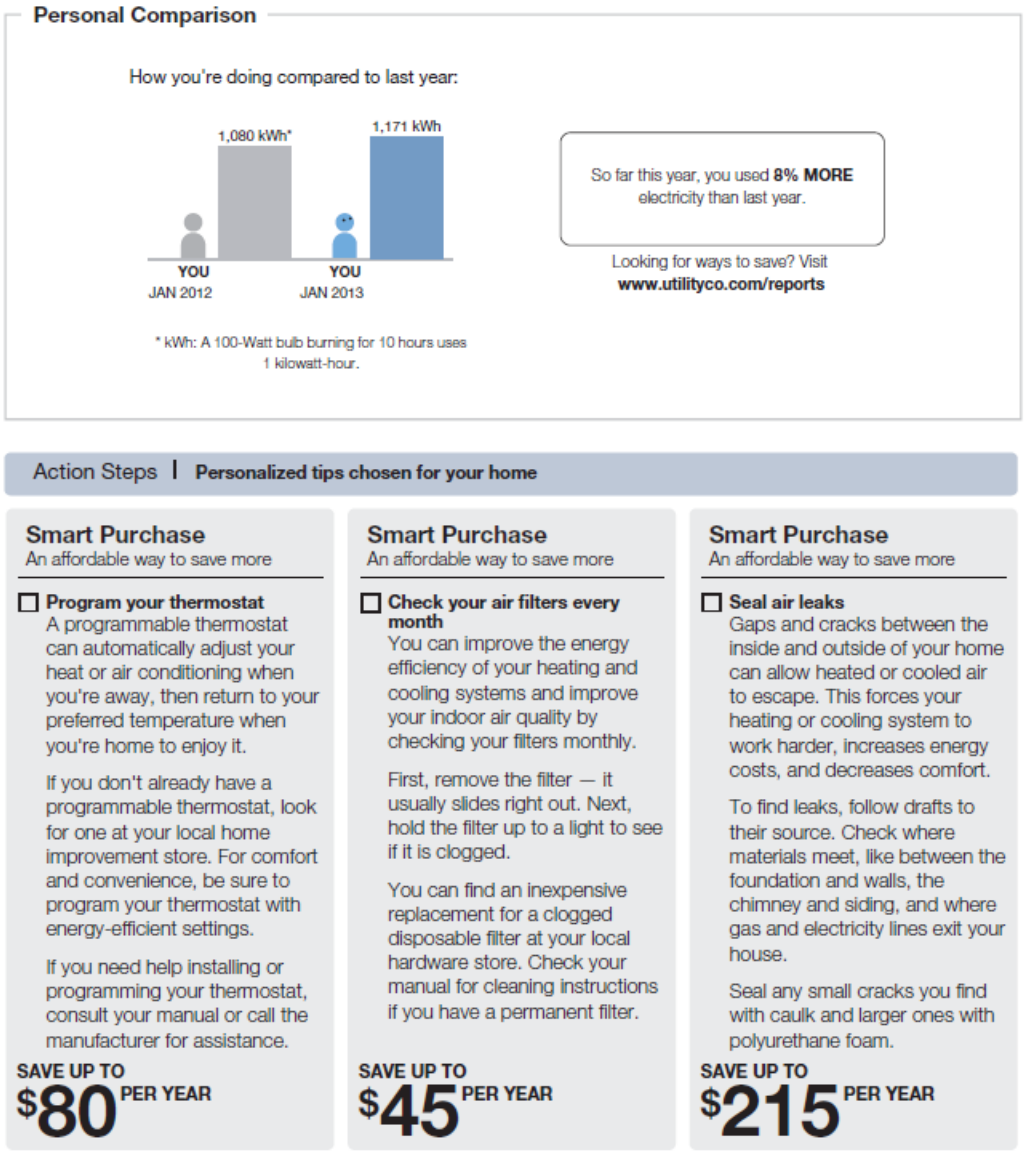


Figure A3: Home Energy Report: Back Page



References

[1] AHA (American Hospital Association) (2012). "AHA Annual Survey Database." See http://www.aha.org/research/rc/stat-studies/data-and-directories.shtml

[2] 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).

[3] BLS (Bureau of Labor Statistics) (2014). "Quarterly Census of Employment and Wages." http://www.bls.gov/cew/datatoc.htm

[4] CMS (Center for Medicare & Medicaid Services) (2013a). "CMS Medicare Provider of Services Files." Available from http://www.nber.org/data/provider-of-services.html

[5] CMS (Center for Medicare & Medicaid Services) (2013b). "Hospital Compare Data." Available from https://data.medicare.gov/data/hospital-compare

[6] Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik (2009). "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika*, Vol. 96, pages 187–199.

[7] CTTI (Clinical Trials Transformation Initiative) (2012). "Database for Aggregate Analysis of ClinicalTrials.gov." Available from http://www.trialstransformation.org/what-we-do/analysis-dissemination/state-clinical-trials/aact-database

[8] EIA (Energy Information Administration) (2014). "Form EIA-826 Detailed Data". http://www.eia.gov/electricity/data/eia826/

[9] 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).

[10] U.S. News (2013). "Methodology: U.S. News & World Report Best Hospitals 2013-2014." Available from http://www.usnews.com/pubfiles/BH_2013_Methodology_Report_Final_28August2013.pdf