SITE SELECTION BIAS IN PROGRAM EVALUATION*

Hunt Allcott

"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 United States. Predictions based on rich microdata from the first 10 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. *JEL* Codes: C93, D12, L94, O12, Q41.

I. 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, implementation, and economic environments. As Angrist and Pischke (2010) write, "A constructive response to the specificity of a given

*This article replaces an earlier version that 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 Annual 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 article and related research on the economics of energy efficiency.

The Quarterly Journal of Economics (2015), 1117–1165. doi:10.1093/qje/qjv015.
Advance Access publication on March 18, 2015.

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, one 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 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 does 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, giving negative site selection bias. Site selection bias implies that even with a large number of internally valid replications, policy makers could still draw systematically biased inference about a program's impact at full scale.

Although 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: because 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 et al. (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 article's organizing question is, "how well can early Opower replications predict treatment effects in later sites?" Although the Opower program is only one case study of site selection bias, this particular out-of-sample prediction problem is highly 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 energy efficiency standards and subsidies. The Opower program has received substantial media coverage, and based on results from early sites, media reports such as Keim (2014) write that Opower has "consistently achieved energy savings of around 2 percent." Consultancy McKinsey & Co. recently released a study predicting "immense" potential for behavior-based conservation in the United States, with potential savings amounting to 16–20 percent of current residential energy consumption (Heck and Tai 2013). Policy makers use such predictions, as well as evaluations of early pilot RCTs, to help determine future program funding and the stringency of energy conservation mandates.[1]

I begin by using microdata from Opower's first 10 sites to predict effects in the next 101 sites. This is a relatively promising setting for extrapolation: there are large samples totaling 508,000 households, 10 replications spread throughout the country, and a useful set of individual-level covariates to adjust for differences between sample and target populations. 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 overpredict the mean average treatment effect (ATE) by 0.41 to 0.66 percentage point, which equals $560 to $920 million worth of retail electricity in the context of a nationally scaled program. This shows that even in a promising setting for extrapolation, estimates are not externally valid: early sites were strongly

[1] For examples of predictions used to make policy, see ENERNOC (2012) and Quackenbush (2013).

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 versus later customer subpopulations. 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 subpopulation, many utilities later expand it to additional subpopulations 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-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 article does not argue that site selection bias reflects suboptimal 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 article is that site-level selection can systematically bias inference and policy decisions, just as individual-level selection can. The Opower data also do not support an argument to "sacrifice internal validity for external validity," that is, to deemphasize RCTs in favor of less costly nonexperimental approaches that could perhaps be implemented in a more general sample of sites: in the Opower context, it is still more informative to extrapolate RCT results from other sites than to rely on nonexperimental 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 article builds on distinguished existing work on multisite program evaluation and external validity. The Job Training Partnership Act of 1982 (JTPA) evaluations are particularly closely related: 200 job training sites were approached to do RCTs, of which 16 eventually agreed. Hotz (1992), Heckman (1992), and others point out that these sites were not randomly selected and propose that this could lead experimental estimates to differ from the true nationwide effects. 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 JTPA treatment effects are homogeneous across sites. With much larger samples and many more sites, the Opower experiments allow a clearer analysis of these earlier ideas.

Also closely related are the academic studies of early Opower programs, including Allcott (2011), Ayres, Raseman, and Shih (2013), Costa and Kahn (2013), and Allcott and Rogers (2014). 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 10 Opower programs.

The article proceeds as follows. Section II presents case studies from microfinance and clinical trials of how RCT sites differ systematically from policy-relevant target sites. Section III formalizes a model of external validity and site selection bias. Section IV gives an overview of the Opower experiments, and

Section V presents the data. Section VI uses the Opower microdata for extrapolation, and Section VII uses the metadata to explain the site selection bias shown in Section VI. Section VIII concludes. All appendices are available online.

II. Motivation: Examples of Site Selection on Observables

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

II.A. Microfinance Institutions

In the past 10 years, there have been many RCTs with microfinance institutions (MFIs). Are MFIs 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 1,903 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 2 percent of MFIs in the database are RCT partners.

Microfinance RCTs study a variety of different treatments and consider both effects on borrowers and operational outcomes such as default rates. For example, the RCTs summarized by Banerjee, Karlan, and Zinman (2015) study various effects on borrowers, while Gine and Karlan (2014) study how group versus individual liability affects default rates, Field and Pande (2008) study how repayment frequency affects default rates, and Field et al. (2013) study how delayed repayment affects entrepreneurship. For this introductory table (Table I), I focus on eight MFI characteristics that are available in the MIX database and might theoretically be correlated with effects of some microfinance interventions. Average loan balance, percent of portfolio at risk of default, and the percent of borrowers who are female could be correlated with default rates. An MFI's structure (as measured

TABLE I

Microfinance Institution Characteristics: RCT Partners and Nonpartners

	(1) All	(2) Partners	(3) Nonpartners	(4) 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)***
Nonprofit	0.63	0.37	0.64	−0.27
	(0.48)	(0.49)	(0.48)	(0.08)***
Number of borrowers (millions)	0.06	0.85	0.05	0.80
	(0.40)	(1.84)	(0.27)	(0.31)***
Borrowers/staff ratio (000's)	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	1,903	35	1,868	
F-test *p*-value				.00002***

Notes. The first three columns present the mean characteristics for all global MFIs in the Microfinance Information Exchange database, field experiment partners, and field experiment nonpartners, respectively, with standard deviations in parentheses. The fourth column presents the difference in means between partners and nonpartners, with robust standard errors in parentheses. "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. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively. Currencies are in U.S. 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.

by age, nonprofit status, and size) could influence the strength of the MFI's relationship with its clients, which might in turn affect the MFI's ability to implement or monitor an intervention. Similarly, staff availability and expenditures per borrower could affect implementation or monitoring ability.

Table I presents means and standard deviations by partner status. Column (4) presents differences in means for partners versus nonpartners. Partners have smaller average loan balances, as well as a marginally insignificant 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 nonpartner 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 do not differ on observables.

II.B. 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 that treats hardening of the carotid artery in the neck. To be eligible, 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 nontrial hospitals, and for some classes of patients and hospitals, treatment with drugs instead of the surgical procedure might have been preferred.

Table II compares U.S. hospitals that have been the site of at least one clinical trial to those that have never hosted a registered trial. Clinical trial sites are from the ClinicalTrials.gov registry, and hospital characteristics are from Medicare and American Hospital Association databases; see Online Appendix A 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 with drugs. Of 4,653 U.S. hospitals, 1,722 have hosted a drug trial and 1,265 have hosted a procedure trial.

The first three rows of Table II 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.

TABLE II

Hospital Characteristics: Clinical Trial Sites and Nontrial 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	4,653		
F-test *p*-value		.0000***	.0000***

Notes. The first column presents the mean characteristic for all U.S. hospitals, with standard deviations in parentheses. The second and third columns present differences in means between clinical trial sites and nontrial sites, with robust standard errors in parenthesis. “Trial sites” are the hospitals listed as clinical trial sites on the ClinicalTrials.gov registry. “Drug” trials include drugs, biological interventions, and dietary supplements. “Procedure” trials include both surgical and radiation procedures. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively. 1,722 hospitals have hosted drug trials, and 1,265 have hosted procedure trials. *F*-test *p*-value is from a regression of a trial site indicator on all characteristics.

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 this may compromise the external validity of randomized control trials. The average trial site also offers 5 to 6 more of the 21 advanced technologies and 3 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 nontrial sites.

Clinical trial sites also have significantly different processes. They perform 0.33–0.35 standard deviation 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. Although suggestive, both examples are speculative and incomplete. Ideally, one 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.

III. A Model of External Validity and Site Selection Bias

III.A. 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 that 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 1 if individual i is in the sample.

The 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 analogs 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:

$$E[\tau_i \,|\, D_i = 0] = E[E[Y_i \,|\, T_i = 1, D_i = 1, X_i] - E[Y_i \,|\, T_i = 0, D_i = 1, X_i] \,|\, D_i = 0]. \tag{1}$$

This argument is closely comparable to Lemma 1 in Hotz, Imbens, and Mortimer (2005).[2]

[2] 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

III.B. Sites, Replication, and Site Selection Bias

External unconfoundedness requires conceptually different assumptions in single-site versus multisite 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).[3] Define F_s and V_s as vectors of characteristics of the treatment and economic environment, respectively, and define $\tau_s(x) = E[\tau_i \,|\, X_i = x, S_i = s]$ as the ATE 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. $D_s \in \{1, 0\}$ is an indicator that takes value 1 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''.

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 Online Appendix D.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.

[3] 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.

Assumption 3B. No site selection bias. $E[\tau_s(x) \mid D_s = 1] = E[\tau_s(x) \mid D_s = 0]$ 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 V_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.[4]

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 hold 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 multisite 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 (Manning et al. 1988; Kling, Liebman, and Katz 2007).

This discussion formalizes the appeal of replication and 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

[4] 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.

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 site selection bias.[5] 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.

IV. Opower: Overview and Site Selection Mechanisms

IV.A. *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 features a bar graph comparing 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. Online Appendix B presents an example report.

Except for a few utilities whose customer bases are too small for precise impact estimates, all Opower programs are implemented as RCTs, 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 10 programs, 2 randomized households between monthly and quarterly frequencies, while 3 others targeted heavier users with monthly reports

[5] 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—that is, that LATEs from different instruments were not only heterogeneous but were systematically different from the ATE in a target population.

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

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).

IV.B. Potential 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 with Opower's president and cofounder Alex Laskey (personal communication, August 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 they thought might be interested, and then moved forward with those partners that agreed. The founders initiated discussions with 50 to 100 utilities to land the first 10 (Laskey personal communication, August 2014); by now, 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) as part of 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, and others target specific local areas where reduced electricity demand could help 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, 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.[6]

[6] All of Opower's first 10 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

However, this does not make their site selection incentives qualitatively different: social programs and nonprofits 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. Although 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.

V. Data

This section provides a brief overview of the three main data sets: utility-level data, microdata from Opower's first 10 sites, and metadata from all 111 sites that began before February 2013. Online Appendix C provides substantial additional information on the Opower data.

I normalize the data in two ways to improve the ability to extrapolate across sites. First, I use the ATE only over each site's first year to average over seasonal variation and eliminate duration heterogeneity. Using ATEs over the first year (instead of first two or three years) allows the analysis to include more recent sites, and ATEs over the first year are highly predictive of ATEs over the first two years.

Second, I compare and extrapolate effects in percent terms, after normalizing electricity conserved by counterfactual usage (as measured by control group mean usage in site s over the first year posttreatment). Though one could also extrapolate ATEs measured in levels of electricity conserved, the main reason to extrapolate in percent is that it is more predictive: the coefficient of variation in ATEs across the 111 sites is 57 percent higher when measured in levels instead of percent, so the mean squared

contract structure, efficacy at previous sites affects subsequent utility adoption decisions.

error of predicted effects when extrapolating in levels is correspondingly higher than the mean squared error when extrapolating in percent.[7] The percent normalization is also commonly used in practice: Opower's website presents site-level impact estimates in percent terms, as do many media reports, academic evaluations, and consulting studies. After extrapolation is done in percent, predicted effects can be translated into economically important outcomes, such as consumer surplus, retail or wholesale electricity costs, or pollution emissions. In the Online Appendix materials, I show that empirical results are similar when measuring ATEs in levels.

V.A. Utility-Level Data

Table III shows characteristics of Opower's current and potential partner utilities. The 58 current partner utilities include all U.S. electric utilities that had started Home Energy Report RCTs by February 2013.[8] I define potential partner utilities to include all 882 large electric utilities in the United States.[9]

I consider variables that proxy for the five utility-level selection mechanisms proposed in Section IV.B—variables that might moderate both selection and treatment effects. Utility Mean Usage is daily average residential electricity usage. For context, 1 kilowatt-hour (kWh) is enough electricity to run either a typical new refrigerator or a standard 60-watt incandescent light bulb for about 17 hours. This variable and the bottom six in the table are available from EIA (2013) at the utility-by-year level; because Opower program adoption affects some of these variables, I use observations for 2007, the year before the first Opower programs began.

[7] As an analogy, if an educational intervention tended to increase test scores by around 2 percent, but some sites had tests out of 100 points while other sites had tests out of 500 points, the researcher would first want to normalize effects as a percent of the total possible points before predicting effects in different sites.

[8] 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.

[9] 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 5 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.

TABLE III

Utility Characteristics: Opower Partners and Nonpartners

	(1)	(2)	(3)	(4)
	All	Partners	Nonpartners	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				.0000***

Notes. The first three columns present the means of utility-level characteristics for all U.S. utilities, for Opower partners, and for Opower nonpartners, respectively. Standard deviations are in parentheses. The fourth column presents the difference in means between partners and nonpartners, with robust standard errors in parentheses. Utility mean usage and the last six variables are from EIA (2013) for calendar year 2007. The five variables from mean income to Green Party share are population-weighted means of county-level data for the counties in the utility's service territory. Mean income and share college grads are from the 2000 census, while hybrid auto share is the share of registered vehicles that were hybrid-electric as of 2013. Green Party share is the 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, from Pew Center (2011). Residential conservation/sales is the ratio of estimated electricity conserved by residential energy conservation programs to total residential electricity sold, while conservation cost/total revenues is the ratio of total spending on energy conservation programs to total revenues. *, **, ***: 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.

The next seven variables proxy for population preferences, measuring higher income and environmentalism. The sum of these seven variables (after normalizing each to mean 0, standard deviation 1) will be the "Normalized Population Preferences" variable in Section VII. 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.

Residential Conservation/Sales and Conservation Cost/Total Revenues measure complementary or substitute programs. The sum of these two (after similarly normalizing each to mean zero, standard deviation one) will be the Normalized Other Programs variable. 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 III shows that Opower's partner utilities are clearly different from nonpartners: 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 these 13 covariates are unbalanced with more than 90 percent confidence.

V.B. Microdata

I have household-level microdata through the end of 2010 for each of the 10 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 posttreatment. The data set includes household-level demographic data from public records and marketing data providers, as well as census tracts, which I use to merge in tract-level data. Columns (1), (2), and (3) of Table IV present observation counts, means, and standard deviations, respectively. Every variable has at least some missing observations; most 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 IV.B. First Comparison is the ratio of a household's usage to its mean neighbor's usage, as reported in the Social Comparison Module of the household's first report. (Opower also constructs this for control households). The mean of 1.08 implies that these first ten sites consisted of slightly above-mean usage households. The next four variables proxy for

TABLE IV

Household Covariates in Opower Early Site Microdata

	(1) Microdata sample size	(2) Microdata sample mean	(3) Microdata sample std. dev.	(4) Later sites mean
First comparison	475,278	1.08	0.54	1.09
Tract mean income ($000s)	508,082	73.8	28.2	59.3
Tract share college grads	508,082	0.35	0.17	0.27
Tract share hybrid autos	506,367	0.018	0.012	0.011
Green pricing participant	82,836	0.096	0.292	0.009
EE program participant	82,715	0.06	0.24	—
Electric heat	313,076	0.12	0.36	0.28
House age (years)	407,469	41.5	27.7	41.2
Has pool	207,885	0.19	0.35	0.17
Rent	272,308	0.09	0.32	0.33
Single family	241,332	0.77	0.40	0.64
Square feet (000s)	380,296	1.83	0.74	1.83

Notes. Columns (1), (2), and (3), respectively, present the observed sample sizes, means, and standard deviations of household characteristics in the microdata from the first ten Opower sites. Missing data are imputed by multiple imputation. Sample means and standard deviations are based on the imputed data, while sample sizes reflect only nonmissing data. The total microdata sample size is 508,295. First comparison is the ratio of the household's usage to the mean neighbor's usage on the first Home Energy Report. Mean income, share college grads, and share hybrid autos are census tract means from the same source as their utility-level analogs in Table III. Column (4) presents the unweighted mean of site-level average characteristics for the 101 later sites that are not included in the microdata. At each of the later sites, average first comparison is approximated based on the ratio of control group mean usage to utility mean usage, using the fitted values from a regression with data from the first 10 sites. The next four variables are utility-level averages from the data in Table III. There are no public data to approximate EE program participant outside the microdata. 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 five-year estimates for 2005–2009.

population preferences.[10] 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 energy efficiency program before the Opower program began. This and the Green Pricing Participant indicator are only available at one site, so the sample sizes are much lower in column (1).

[10] 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. Online Appendix D.D provides intuition for why this happens and presents results including Democratic vote share.

The final six variables measure characteristics of housing stock. Although I do not hypothesize that site-level variation in these factors directly affects site selection, there are clear theoretical reasons 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.

In the next section, I condition on these variables to predict the ATE for the 101 Opower programs that started after the 10 programs in the microdata. Because I do not have microdata for these later sites, I construct site-level average characteristics and predict the unweighted mean of later sites' ATEs by fitting treatment effects to the unweighted means of later sites' average characteristics. Column (4) of Table IV presents these unweighted means. Comparing columns (2) and (4) shows that the microdata sample differs from the later sites on observable proxies for population preferences: the sample has higher income, more college graduates, more hybrid autos, and is 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.

V.C. *Metadata*

Due to contractual restrictions, Opower cannot share microdata from many of their recent partners. Instead, they have provided site-level metadata, including ATEs and standard errors, control group mean usage, and number of reports sent for each posttreatment month of each RCT. Some utilities have multiple sites, typically because they began with one customer subpopulation and later added other subpopulations in separate RCTs. As of February 2014, there were 111 sites with at least one year of posttreatment data at 58 different utilities.

Opower's analysts estimated first-year ATEs using mutually agreed procedures and code; see Online Appendix C 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 1 in every 12 in the United States. First-year treatment effects average 1.31 percent, or 0.47 kWh/day.

V.D. 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. Formally, if assumption (3A) holds across all sites, this is sufficient for assumption (3B). In reality, the 111 ATEs vary substantially, from 0.50 to 2.63 percent, or from 0.10 to 1.47 kWh/day. 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 percent ATEs are homogeneous with a p-value of less than .001. The percent ATEs have standard deviation of 0.45 percentage point, while the average standard error is only 0.18 percentage point. If measuring ATEs in kWh/day levels, ATEs appear even more dispersed: the ratio of maximum to minimum ATEs is much larger, Cochran's Q test rejects with even higher confidence, and the coefficient of variation is larger.

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 I presents a 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.

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?

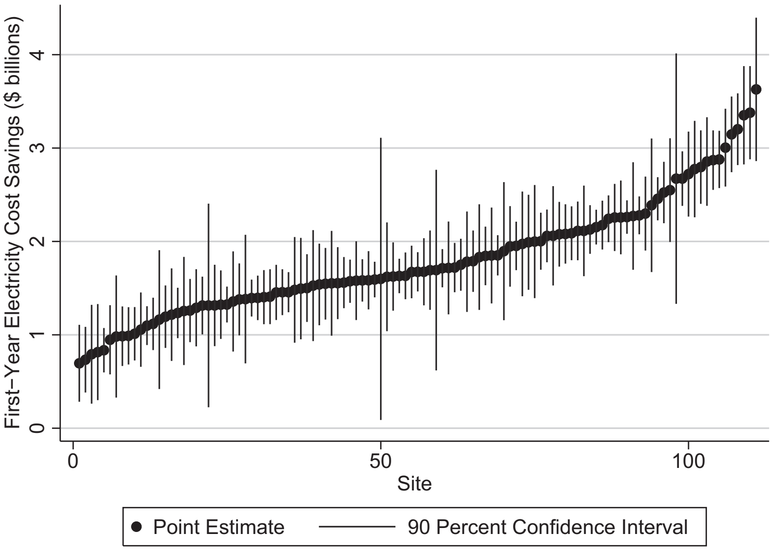

FIGURE I

Opower Program: Nationwide Savings Predicted by Effect at Each Site

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 households at all 882 potential Opower partner utilities. Sites are ordered on the *x*-axis by effect size.

VI. MICRODATA: EXTRAPOLATION UNDER EXTERNAL UNCONFOUNDEDNESS

VI.A. *Empirical Strategy*

Under assumptions (1)–(4) in Section III, 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. This section uses the microdata to test external unconfoundedness "in sample" by extrapolating from the microdata to the remainder of the sites in the metadata. Online Appendix D contains supporting materials for this section.

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).[11]

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 sites. However, I can still use two simple off-the-shelf procedures commonly used in applied work: linear prediction and reweighting 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 reweighting estimator, because it reduces extreme weights that match samples on Xs that don't actually moderate the treatment effect.

1. Determining the Set of Conditioning Variables. Y_{is} is household i's mean daily electricity use (in kWh/day) over the first year posttreatment, C_s is the control group mean usage in site s over that same year, and $y_{is} = \frac{100Y_{is}}{C_s}$. As discussed in Section V, ATEs are more naturally extrapolated in percent terms, so I use y_{is} as the dependent variable. $\overline{X}_{D=1}$ is the vector of sample means of the covariates reported in column (2) of Table IV, where the mean is taken across all 25 imputations. $\tilde{X}_{is} = X_{is} - \overline{X}_{D=1}$ 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).[12] Heterogeneous treatment effects are estimated using the following equation:

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

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

[11] 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.

[12] Given that the γ_s coefficients are site-specific, normalizing the Y_0 vector by control mean usage C_s would only rescale γ_s and would not affect the α coefficients used for extrapolation.

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 1 percent effect. 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 et al. (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^*.

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. $E[\tau_i \mid X_i = x] = \alpha x + \alpha_0$.

I denote sample and target ATEs 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.

3. Reweighting. A second approach to extrapolation is to reweight 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. $f_{D=1}(x) = f_{D=0}(x) \cdot (1 + \lambda(x - \overline{X}^*_{D=0}))$.

Under this assumption, observation weights $w_i = \frac{1}{1+\lambda(X^*_i - \overline{X}^*_{D=0})}$ reweight 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 $\sum_i w_i X^*_i = \overline{X}^*_{D=0}$. In words, the second constraint is that the reweighted 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.

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" $\tilde{\tau}$ is adjusted to match the mean frequency $\overline{F}$ across all 111 sites in the metadata, which is 0.58 report per month. Denoting the frequency at site s as F_s, the adjustment is:

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

Standard errors are calculated using the delta method.

VI.B. Results

1. Heterogeneous Treatment Effects. Table V presents heterogeneous treatment effects using combined microdata from the first 10 sites. Columns (1)–(3) have y_{is} (electricity usage as a percent of control) as the dependent variable. Column (1) shows that the ATE $\hat{\tau}_{D=1}$ across the first 10 sites is 1.707 percent. Because column (1) excludes the X covariates, this is the only column in

TABLE V

Opower Program: Heterogeneous Treatment Effects in Early Site Microdata

	(1)	(2)	(3)	(4)
				Usage
Dependent variable:	Usage (percent of control)			(kWh/day)
Treatment	1.707 (0.056)***	1.790 (0.055)***	1.785 (0.058)***	0.533 (0.018)***
T × first comparison		2.672 (0.260)***	2.707 (0.264)***	0.840 (0.084)***
T × tract mean income		0.001 (0.003)		
T × tract share college grads		−0.641 (0.675)		
T × tract share hybrid autos		7.052 (7.448)		
T × green pricing participant		0.024 (0.242)		
T × EE program participant		0.005 (0.266)		
T × electric heat		0.960 (0.222)***	0.984 (0.212)***	0.308 (0.065)***
T × house age		−0.002 (0.002)		
T × has pool		0.569 (0.223)**	0.646 (0.221)***	0.193 (0.066)***
T × rent		−0.252 (0.257)		
T × single family		0.205 (0.232)		
T × square feet		0.417 (0.131)***	0.460 (0.110)***	0.137 (0.034)***
N	508,295	508,295	508,295	508,295

Notes. This table presents estimates of equation (2) with different X characteristics. The dependent variable is household i's posttreatment electricity use normalized by the site s control group posttreatment average. First comparison is the ratio of the household's usage to the mean neighbor's usage on the first Home Energy Report. Missing data are imputed by multiple imputation. Robust standard errors, clustered at the level of randomization (household or block batch), are in parentheses. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

Table V that does not use multiple imputation. The $R2$ is 0.86, reflecting that fact that lagged usage Y_{0i} explains much of the variation in current usage 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 et al. (2008) top-down procedure, including only the $\tilde{X}^*$ that statistically significantly moderate the treatment effect. The $\hat{\alpha}$ coefficients are very similar between columns (2) and (3).

The signs and magnitudes are also sensible. The First Comparison interaction is strongly positive: informing a household that it uses 10 percentage points more relative to its mean neighbor (meaning that the First Comparison variable increases by 0.1) is associated with a 0.27 percentage point larger treatment effect. Homes with electric heat conserve about 1 percentage point more, suggesting that reduced heating energy use is an important effect of the program. Homes that have pools or are 1,000 square feet larger both have approximately 0.5 percentage point larger effects.

For comparison, column (4) repeats column (3) using Y_{is} (unnormalized usage in kWh/day) as the dependent variable. The average usage across all control group households in the sample is 30.0 kWh/day. Thus, the ratios of α coefficients in column (4) to column (3) should be approximately $\frac{30.0}{100}$. Although this is not exact because control group usage C_s varies across the 10 sites, the ratios are all between 0.299 and 0.313.

The empirical likelihood estimates for the reweighting estimator are in Online Appendix D. As suggested by comparing sample and target means in Table IV, they imply higher weights for households with electric heat while keeping other variables' means approximately the same. Online Appendix D also presents the estimated frequency adjustment; the estimated $\hat{\phi}$ is 0.517 percent of electricity use per report/month. This point estimate implies that a 1 standard deviation change in reports per month across the 111 sites (0.11 reports/month) would change the ATE by $0.517 \times 0.11 \approx 0.057$ percentage point. Frequency adjustment does not meaningfully affect the analyses, both because frequency is uncorrelated with other factors and because the adjustment is small relative to the variation in effect sizes.

2. Predicting Target Treatment Effects. Table VI presents estimates and extrapolation results. The left three columns present frequency-adjusted ATEs for the first 10 sites. Column (1) presents the frequency-adjusted sample ATE $\tilde{\tau}_{D=1}$. This is simply the estimate in column (1) of Table V adjusted by 0.04 percent to match the 111-site mean reports/month using equation (4).

Columns (4)–(6) present estimates for the 101 later sites. Column (4) shows that the mean of the 101 true frequency-adjusted ATEs is 1.26 percent. Per equation (3), the linear prediction in column (5) is simply the frequency-adjusted sample

TABLE VI

Opower Program: Predicted Effects Using Microdata

	(1)	(2)	(3)	(4)	(5)	(6)
	First 10 Sites			Later 101 Sites		
		Nonexperimental estimates			Prediction from first ten sites	
	True ATE	Pre-post	W/ state control	True ATE	Linear	weighted
Frequency-adjusted ATE (percent)	1.67	2.92	2.88	1.26	1.92	1.66
Standard error	0.06	0.08	0.08	0.04	0.08	0.06
Difference from true value (percent)	—	1.25	1.20	—	0.66	0.41
Value of difference in a nationally-scaled program (billion)	—	\$1.72	\$1.66	—	\$0.92	\$0.56

Notes. This table presents estimated and predicted effects of the first year of the Opower program using microdata. ATEs are "frequency adjusted" to match the average number of Home Energy Reports per month across all 111 sites in the metadata. Columns (1)–(3) present estimated effects for the first 10 sites. Columns (2) and (3) present nonexperimental results. Column (2) is a pre-post comparison using treatment group observations only, controlling for household fixed effects and weather differences. Column (3) adds a control for average usage in untreated utilities in the same state. Columns (4)–(6) present effects for the later 101 sites. Column (4) contains the true unweighted mean of the frequency-adjusted ATEs across the later 101 sites. Columns (5) and (6) present effects predicted by the microdata from the first ten sites. Column (5) contains the predicted ATE using equation (3), assuming that effects are linear in covariates. Column (6) contains the ATE predicted by reweighting to match target mean observables. "Value of difference in a nationally-scaled program" multiplies the difference from true value by total annual retail electricity expenditures for residential consumers at all 882 potential partner utilities, which equals \$138 billion.

ATE $\tilde{\tau}_{D=1}$ adjusted by the differences in sample and target mean X^* (the fourth minus the second column of Table IV) multiplied by the $\hat{\alpha}$ estimates (column (3) of Table V). This linear adjustment primarily consists of an increase of $(0.36 - 0.12) \times 0.984\% \approx 0.24$ % predicted by the higher proportion of electric heat households in the target. The weighted prediction in column (6) is closer to the unconditional sample ATE.

The linear and weighted predictions, respectively, are 0.66 and 0.41 percentage point larger than the true ATE. As suggested by the standard errors, the overpredictions are highly statistically significant, with p-values $< .0001$.[13] Across all consumers at all 882 potential partner utilities nationwide, annual retail electricity expenditures are approximately \$138 billion. Thus, in the context of a nationally scaled program, these mispredictions would cause first-year retail electricity cost savings to be overstated by \$560–920 million. If the only goal were to predict

[13] See Online Appendix D.B for formal details on this test.

the target ATE, this $560–920 million illustrates the improved inference from randomly sampling a sufficiently large number of replication sites instead of allowing nonrandom site selection.

As a benchmark, Table VI also includes nonexperimental estimates of the sample ATE. Column (2) is a pre-post comparison using treatment group microdata only, controlling for household fixed effects and weather differences. Column (3) adds a control for average usage at untreated utilities in the same state. Both nonexperimental estimates are substantially different than the true sample ATE, which underscores the importance of using randomized control trials in this context. Online Appendix D.C.4 more formally details these estimators and also shows that it is more informative (lower mean squared error) to extrapolate RCT results from other sites than to rely on nonexperimental estimates from the same site. These results show that in the Opower context, it would be less predictive to "sacrifice internal validity for external validity," that is, to deemphasize RCTs in favor of less costly nonexperimental approaches.

Predictions can also be made for each of the 101 later sites. Figure II 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 versus gray circles distinguish predictions that are versus are not statistically different from the true $\tilde{\tau}_s$ with 90 percent confidence; the 24 nonsignificant 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 uncorrelated or perhaps negatively correlated with the site-specific heterogeneity. This echoes the result from Table VI that observables are not very informative about unobservables in this context.[14] Thus, the logic of inferring the direction and magnitude of bias from

[14] 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."

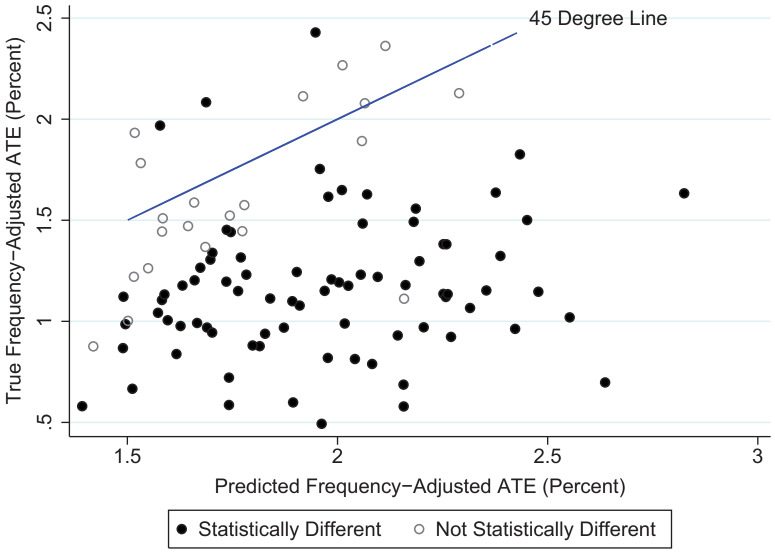

FIGURE II

Opower Program: Site-Specific Predictions Using Microdata from Early Sites

This figure plots the actual ATE for each of the 101 later Opower 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.

unobservables (Altonji, Elder, and Taber 2005) would not work well here.

3. *Explaining the Prediction Failure.* So far, the results show a systematic failure of two simple approaches to predict effects in later sites. Does this happen due to a violation of external unconfoundedness, lack of external 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 $\frac{1}{4}$. 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 4 of the 101 individual target sites in Figure II 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 IV 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). Online Appendix D.C tests whether prediction can be improved when $f_{D=0}(x)$ is known by predicting the ATE for each of the 10 sites in the microdata, using the other 9 sites as the "sample." Results show that predictions from the linear approach can be marginally improved (reducing root mean squared prediction error by around 5 percent) by using a polynomial in X that also includes squares and interactions, and/or by predicting effects only for the target subpopulation with improved overlap.

This marginal improvement should be interpreted with three caveats. First, while the within-sample tests in Online Appendix D.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 improved conditioning on X had substantially improved prediction between the sample sites, it might still be difficult 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 effects in later sites. As suggested by Heckman et al. (1998) and Smith and Todd (2005) in the context of individual-level selection bias, such adjustments might be possible, but only under particular conditions. Overall, these results suggest that external unconfoundedness does not hold in this context, despite 10 replications.

VII. Metadata: Explaining Site Selection Bias

Why were Opower's first 10 sites positively selected from the full set of 111 sites? 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 IV.B, I first separate within-utility versus between-utility selection and then use utility-level covariates to test the hypothesized utility-level mechanisms. Online Appendix E presents robustness checks and additional regressions that may be of interest for readers particularly interested in the Opower program.

VII.A. Empirical Strategy

1. Cohort Trends and Within- vs. Between-Utility Selection. The microdata analysis compared the 10 initial sites to all later sites, which exploits only a coarse binary measure of early versus 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. Of course, the results of Section VI suggest that $\hat{\eta}$ will be negative. In all regressions with a treatment effect τ as the dependent variable, I weight observations by analytic weights $\frac{1}{\widehat{Var(\tau)}}$; this improves precision by weighting more heavily the τ, which are more precisely estimated.

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 IV.B 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.

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 IV.B.

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 10 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 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.

VII.B. Results

1. Cohort Trends and Within- vs. Between-Utility Selection. Table VII presents trends for earlier versus later sites. 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 point smaller ATEs.

Figure III illustrates this regression. 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. The 46 sites that started after January 2012 have particularly low ATEs, averaging 1.05 percent. This further corroborates the results from Section VI that extrapolating from early sites would overstate efficacy in later sites.

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 point 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 0, implying that much 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. Negative coefficients in both columns suggest that earlier utilities were positively selected from later utilities.

2. Testing Utility-Level Selection Mechanisms. Table VIII 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. Larger utilities with higher-income and more environmentalist populations are more likely partners.

TABLE VII

Opower Program: Cohort Trends and Within- versus Between-Utility Selection

	(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)
R^2	0.22	0.65	0.76	0.22	0.22
N	111	73	73	111	111
Utility indicator variables	No	Yes	Yes	No	No
Sample	All sites	Multisite utilities	Multisite 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 the reported covariates to equation (5). The dependent variable is frequency-adjusted ATE, and the mean of this variable is 1.31 percent. Within-utility start number takes value L if the site is the Lth site within the same utility. Site start date is the date (in Stata's "td" format) that the program began, divided by 365. Observations are weighted by inverse variance. Robust standard errors are in parentheses. *, **, ***: Statistically significant with 90, 95, and 99 percent confidence, respectively.

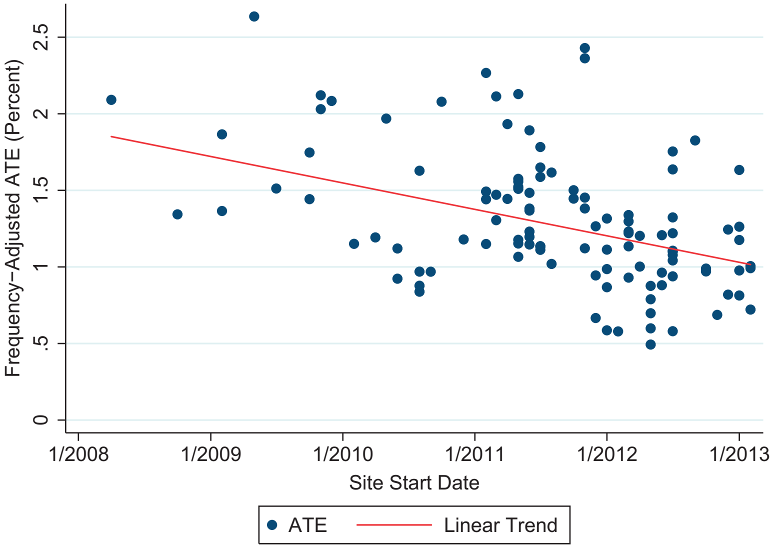

Figure III

Opower Program: Efficacy Trend across Sites

This figure plots the data and fitted regression line for column (1) of Table VII. In this regression, observations are weighted by inverse variance. All ATEs are "frequency adjusted" to match the average number of Home Energy Reports per month across the 111 sites in the metadata.

TABLE VIII

Opower Program: Utility-Level Selection

	(1)	(2)	(3)	(4)
	Selection		Outcomes	
Dependent variable:	1 (early partner)	1 (partner)	Frequency-adjusted ATE (%)	Frequency-adjusted ATE (%)
Utility mean usage (kWh/day)	−0.068	−0.007	−0.040	−0.038
	(0.046)	(0.012)	(0.009)***	(0.007)***
Normalized population preferences	0.808	0.337	0.122	0.094
	(0.359)**	(0.091)***	(0.058)**	(0.043)**
Normalized other programs	0.093	0.071	−0.002	0.008
	(0.144)	(0.059)	(0.011)	(0.011)
Municipally owned utility	−2.145	0.405	−0.367	−0.331
	(1.109)*	(0.264)	(0.171)**	(0.129)**
Investor-owned utility	−3.695	0.557	−0.485	−0.382
	(1.111)***	(0.302)*	(0.175)***	(0.149)**
ln(residential customers)	0.541	0.494	−0.043	−0.060
	(0.440)	(0.078)***	(0.037)	(0.042)
Within-utility start number			−0.070	−0.031
			(0.016)***	(0.018)*
Control mean usage (kWh/day)			0.015	0.016
			(0.003)***	(0.003)***
Site start date (Years)				−0.122
				(0.036)***
Pseudo-R^2	0.43	0.44		
N	58	882	111	111
R^2			0.47	0.56
Estimator	Probit	Probit	OLS	OLS
Sample	Partner utilities	All utilities	All sites	All sites

Notes. Columns (1) and (2) present estimates of equation (7), while columns (3) and (4) present estimates of equation (8). Normalized population preferences 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 0, standard deviation 1. Normalized other programs is the sum of residential conservation/sales and conservation cost/total revenues after normalizing each to mean 0, standard deviation 1. Within-utility start number takes value L if the site is the Lth site within the same utility. Site start date is the date (in Stata's "td" format) that the program began, divided by 365. In columns (3) and (4), 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.

Point estimates suggest that preexisting energy efficiency programs are positively associated with selection, although this is not statistically significant. Ownership structure, however, has different associations early (column (1)) versus overall (column (2)). This is consistent with anecdotal evidence (Laskey personal communication, August 2014): initially, the program was unproven at scale, and the company relied on innovative and nonprofit utilities for business. As initial RCTs gave positive results, and as EERS policies expanded, the more conservative and heavily regulated IOUs increasingly adopted the program.

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 personal communication, August 2014). Instead, this between-utility selection seems to have been based on other "unintentional" mechanisms.

Columns (3) and (4) present outcomes estimates from equation (8). Because the specifications also condition on control mean usage C_s, 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. Normalized Population Preferences is strongly positive and significant, meaning that higher income and environmentalist populations have larger treatment effects. IOUs have smaller effects, perhaps due to lack of customer engagement. Municipally owned utilities also have smaller effects than the omitted ownership category (coops and other nonprofits), but point estimates suggest larger effects than IOUs. Point estimates suggest larger utilities have smaller effects, and Online Appendix E presents additional regressions that support this result and suggest that it acts through urban areas where people are less likely to know their neighbors.

How much site selection is explained by utility-level observables? Column (4) adds site start date M_s to equation (8), which is also equivalent to adding the Z_u variables to column (5) of Table VII. Adding the Z_u variables attenuates the η coefficient on site start date M_s from -0.175 to -0.122 percentage point a year, suggesting that site-level observables explain just under $\frac{1}{3}$ of the decline in efficacy between earlier and later sites.

Including utility-level covariates explains more of site selection than individual level covariates 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 capture some population characteristics, as suggested by a case study of the Democrat vote share variable in Online Appendix D.D. In other words, part of the prediction failure with *individual-level* observables in Section VI is due to extrapolating to sites with different *site-level* observables.[15]

Variables that moderate both selection and outcomes suggest mechanisms of site selection bias, and recall that if ρ and θ have the same sign (opposite signs) for a given Z variable, this suggests that Z causes positive (negative) selection. Utility ownership structure and population preferences are the two that are consistently statistically significant both selection and outcome estimates. Ownership structure is associated with positive selection of early partners from current partners: in Table VIII, the Municipally Owned and Investor-Owned Utility variables both have $\hat{\theta} < 0$ in column (3) and $\hat{\rho} < 0$ in column (1). However, the $\hat{\rho}$ point estimates for both variables are positive in column (2), meaning that ownership structure is now associated with negative selection of current partners from all potential partners.

Figure IV shows the unconditional relationship between the Normalized Population Preferences variable and the frequency-adjusted percent ATE at each utility's first site. While Population Preferences is normalized to mean 0 across the 882 potential partner utilities, the mean for current partners is approximately 1, implying strong selection of current partners. Empty squares versus solid circles denote the first 10 versus later 101 sites; the fact that the squares tend toward the right of the figure illustrates that early partners were also positively selected from current partners. The best fit line slopes upward, illustrating larger treatment effects at higher-income and environmentalist utilities. The figure thus illustrates that $\hat{\rho}$ and $\hat{\theta}$ are both positive, suggesting that population preferences have caused positive selection of both early partners and current partners.

What do these metadata results predict would be the first-year effects of a nationwide scaled program? To answer this, I use the fitted values of the outcome equation in column (3) to predict total effects across all consumers at all 882 potential partner utilities nationwide.[16] This predicts national first-year retail

[15] 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 VIII. Sixty-one of the 101 later sites are outside the support of the scores for the first 10 sites, meaning that they are different on site-level observables. When predicting site-specific ATEs from the 10-site microdata as in Figure II, these 61 sites have larger absolute prediction errors.

[16] More specifically, 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}

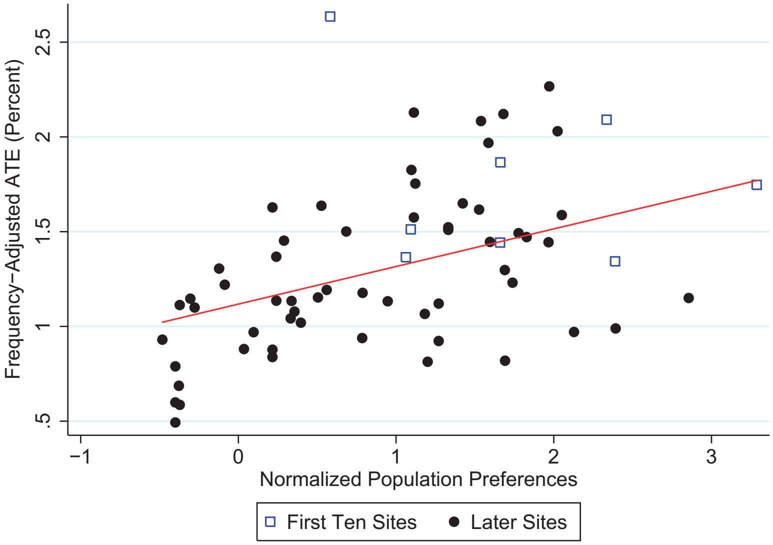

FIGURE IV

Opower Program: Site Selection on Population Preferences

This figure plots the unconditional 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 0, standard deviation 1. In estimating the best fit line, observations are weighted by inverse variance.

electricity cost savings of $1.45 billion. Of course, these predictions rely on the assumption that $\upsilon \perp \xi$, that is, that no unobserved factors that moderate treatment effects affected selection of the 58 current partners from the set of 882 potential partners. Column (4) of Table VIII 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 predictions with a sample of 111 sites may be biased due to unobserved selection mechanisms.

By comparing this prediction to a prediction that does not adjust for site-level observables, it is possible to quantify the

equal to the sample mean, predict fitted values of $\tilde{\tau}_s$ for each of the 882 utilities using coefficients from column (3), calculate the average fitted $\tilde{\tau}_s$ weighted by each utility's number of residential consumers, and multiply by total annual electricity expenditures ($138 billion).

extent to which current sites are selected on observables. Multiplying the mean ATE from the 111 sample sites (1.31 percent) by retail electricity expenditures at the 882 potential partner utilities ($138 billion) predicts first-year savings of $1.80 billion. This overprediction of $350 million shows that current 111 sample sites are still positively selected on observables from the nationwide population.

VII.C. Prediction with Random Sampling versus Self-Selected Sites

In some cases, such as the JTPA job training program or Mexico's PROGRESA conditional cash transfer program, a program implementer has a target population and the choice of whether to evaluate the program in the entire target, a randomly selected subset of sites, or a self-selected subset. Opower is different, in the sense that these 111 sites were not envisioned as a final target population during the program's early expansion. However, the 111-site metadata can be used as an example to illustrate the predictive gains of random sampling from the 111 sites versus prediction using sites that self-select into early evaluations.

Figure V considers four different approaches to predicting the mean ATE across the 111 sites, which is 1.31 percent. The graph presents the root mean squared error (RMSE) of these predictions as the number of sites used for prediction N_s increases from 1 to the full 111. The solid line presents predictions using the mean frequency-adjusted ATE from N_s randomly-selected sites, where the RMSE is over 1,000 random draws.[17] With a large number of sites, external unconfoundedness is fulfilled under assumption (3B), and the figure correspondingly shows that the RMSE approaches 0 as the number of sites approaches the full 111.

The dot-dashed line presents predictions using stratified random sampling. As an example, I stratify on Normalized Population Preferences, randomly sampling $\frac{N_s}{2}$ above-median sites and $\frac{N_s}{2}$ below-median sites. (I drop all odd values of N_s.) Of course, the gains from stratification are decreasing in N_s and

[17] To be precise, the RMSE for N_s sites is $\sqrt{\frac{\sum_{j=1}^{1000}\left[\frac{\sum_{\mathcal{R}_j}\tilde{\tau}_s}{N_s} - 1.31\right]^2}{1000}}$, where j indexes draws and $\mathcal{R}_j$ represents the set of N_s randomly selected sites in draw j.

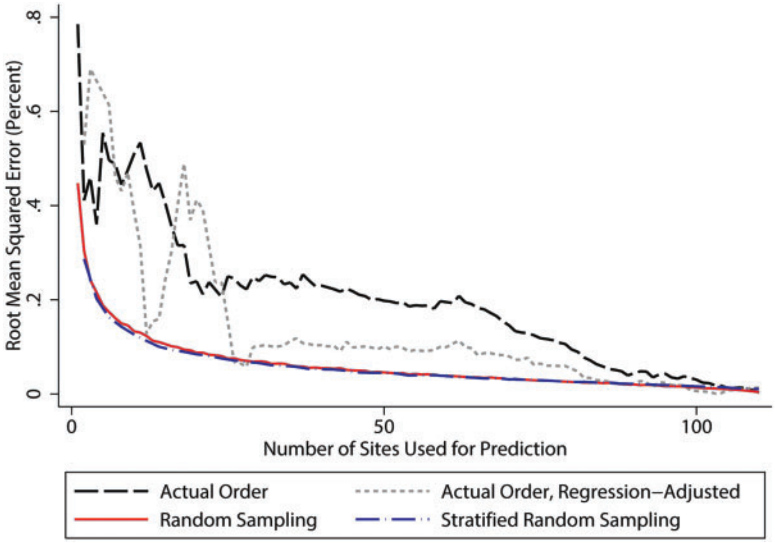

FIGURE V

Opower Program: Predicting Mean ATE across All 111 Sites

This figure plots the root mean squared error (RMSE) of predictions of the mean ATE across the 111 sites in the metadata, which is 1.31 percent. The number of sites used for prediction, denoted N_s, is on the x-axis. All ATEs are “frequency adjusted” to match the average number of Home Energy Reports per month across the 111 sites. “Actual Order” is the RMSE when predicting with the mean ATE from the first N_s sites that implemented the program. “Actual Order, Regression-Adjusted” estimates equation (8) with the first N_s sites and constructs fitted values of the ATE for all 111 sites; the prediction is the mean of the 111 fitted values. “Random Sampling” is the RMSE across 1,000 draws where in each draw, the mean of N_s randomly selected ATEs is used to predict the 111-site mean ATE. “Stratified Random Sampling” stratifies the random sample by above- versus below-mean 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 0, standard deviation 1.

depend on the number of strata and the predictiveness of the variable used to stratify. In this example, there are small gains of approximately 6.5 percent (i.e., a 0.01 percentage point decrease in the RMSE) with fewer than 14 sites.

The dashed line labeled “Actual Order” plots predictions using the mean frequency-adjusted ATE from the first N_s sites that actually started the Opower program. This RMSE also equals the absolute value of the error using the first

N_s frequency-adjusted ATEs graphed in Figure III.[18] As that earlier figure showed, the early sites overestimate the 111-site average ATE, so this prediction includes bias, and the RMSE is larger. Furthermore, Figure III showed that the 46 sites that started after January 2012 have particularly low ATEs. This relatively late decrease in efficacy means that the RMSE on Figure V doesn't approach the lower RMSE from random selection until the final 10–30 sites.

One benefit of having more replications, even if not randomly selected, is that they allow researchers to learn how site-level factors moderate treatment effects. As an example of this learning process, I estimate equation (8) using the first N_s sites and construct fitted values of $\tilde{\tau}_{su}$ for all 111 sites; the "regression-adjusted" prediction is the mean of these fitted values.[19] Of course, the regression will overfit with small N_s, and the gains from regression-adjustment depend on the variables used. In this example, the regression adjustment is unreliable with N_s less than about 25, and adding sites can change the predictions substantially. For larger N_s, the regression adjustment improves the RMSE substantially relative to the unadjusted prediction, but random site selection still performs better.

Figure V has three main implications. First, site selection bias can persist even when the number of sites is very large, both in absolute terms and relative to the number of target sites. Second, using site-level observables to informally or econometrically control for site selection bias may not be helpful until there are a large number of sites. Third, however, random sampling can generate meaningful improvements in inference with a relatively small number of sites.[20]

[18] This RMSE is $\sqrt{\left[\frac{\sum_{\mathcal{A}_N}\tilde{\tau}_s}{N_s}-1.31\right]^2}=\left|\frac{\sum_{\mathcal{A}_N}\tilde{\tau}_s}{N_s}-1.31\right|$, where $\mathcal{A}_N$ denotes the set of the first N_s sites that started the program.

[19] When implementing the regression, I allow Stata to drop regressors when the number of observations N_s is insufficient to identify coefficients.

[20] When it is possible to randomly sample sites from the target population subject to a research budget constraint, the optimal number of sites depends on two factors other than the cost per site. First, larger expected variance in site-level unobserved heterogeneity implies that additional sites are more valuable. Second, additional sites are useful to the extent that they can more precisely identify how site-level covariates moderate the treatment effect, which reduces the site-level unobserved heterogeneity.

VIII. 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 what they would be in target sites. This site selection bias could arise with both RCTs and non-experimental evaluations.

The Opower energy conservation programs are a remarkable opportunity to study these issues, given a large sample of microdata 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). While the within-utility positive selection could have been predicted qualitatively with the knowledge that utilities initially target high-usage consumers, 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, several standard econometric approaches are unhelpful, and any econometric approach can be biased by selection on unobservables. Second, one might propose to “sacrifice internal validity for external validity” by running less costly nonexperimental evaluations in a more general sample of sites. In the Opower example, however, nonexperimental estimators perform relatively poorly. Third, when reporting results, researchers can clearly define policy-relevant target populations and compare sample and target on observables, as in Tables I, II, and III. Although this can help to detect site selection bias, it does not solve the problem. Fourth, 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 10 replications in sites that did differ on potentially relevant site-level factors. Fifth, researchers can devote extra effort to recruiting especially reluctant evaluation partners. This can improve representativeness on unknown or econometrically unobservable moderators, analogous to how additional follow-up can help reduce bias from individual-level attrition (Dinardo, McCrary, and Sanbonmatsu 2006).

Sixth, researchers can consider randomly or fully sampling from the target population. With a very large budget, programs could be evaluated in the entire population of sites to which they 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. Random assignment (whether between treatment and control or between sample and target) is costly, so it is certainly not always worthwhile. But just as researchers have increasingly considered randomized and quasi-randomized research designs to address individual-level selection bias, we may wish to further consider such strategies to address site-level selection bias.

New York University, Nber, Jpal, and E2E

Supplementary Material

An Online Appendix for this article can be found at QJE online (qje.oxfordjournal.org).

References

Allcott, Hunt, “Social Norms and Energy Conservation,” *Journal of Public Economics*, 95 (2011), 1082–1095.

Allcott, Hunt, and Todd Rogers, “The Short-Run and Long-Run Effects of Behavioral Interventions: Experimental Evidence from Energy Conservation,” *American Economic Review*, 104 (2014), 3003–3037.

Altonji, Joseph, Todd Elder, and Christopher Taber, “Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools,” *Journal of Political Economy*, 113 (2005), 151–184.

Angrist, Joshua, and Guido Imbens, “Identification and Estimation of Local Average Treatment Effects,” *Econometrica*, 62 (1994), 467–475.

Angrist, Joshua, and Jorn-Steffen Pischke, “The Credibility Revolution in Empirical Economics: How Better Research Design Is Taking the Con out of Econometrics,” *Journal of Economic Perspectives*, 24 (2010), 3–30.

Ayres, Ian, Sophie Raseman, and Alice Shih, “Evidence from Two Large Field Experiments that Peer Comparison Feedback Can Reduce Residential Energy Usage,” *Journal of Law, Economics, and Organization*, 29 (2013), 992–1022.

Banerjee, Abhijit, Dean Karlan, and Jonathan Zinman, “Six Randomized Evaluations of Microcredit: Introduction and Further Steps,” *American Economic Journal: Applied Economics*, 7 (2015), 1–21.

Bold, Tessa, Mwangi Kimenyi, Germano Mwabu, Alice Ng’ang’a, and Justin Sandefur, “Scaling Up What Works: Experimental Evidence on External Validity in Kenyan Education,” Mimeo, Goethe University Frankfurt, 2013.

Chandra, Amitabh, and Douglas Staiger, “Productivity Spillovers in Health Care: Evidence from the Treatment of Heart Attacks,” *Journal of Political Economy*, 115 (2007), 103–140.

Costa, Dora, and Matthew Kahn, “Energy Conservation ‘Nudges’ and Environmentalist Ideology: Evidence from a Randomized Residential Electricity Field Experiment,” *Journal of the European Economic Association*, 11 (2013), 680–702.

Crepon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora, “Do Labor Market Policies Have Displacement Effects? Evidence from a Clustered Randomized Experiment,” *Quarterly Journal of Economics*, 128 (2013), 531–580.

Crump, Richard, Joseph Hotz, Guido Imbens, and Oscar Mitnik, “Nonparametric Tests for Treatment Effect Heterogeneity,” *Review of Economics and Statistics*, 90 (2008), 389–405.

Dehejia, Rajeev, “Was There a Riverside Miracle? A Hierarchical Framework for Evaluating Programs with Grouped Data,” *Journal of Business and Economic Statistics*, 21 (2003), 1–11.

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

Dinardo, John, Justin McCrary, and Lisa Sanbonmatsu, “Constructive Proposals to Dealing with Attrition: An Empirical Example,” Mimeo, University of California at Berkeley, 2006.

Donabedian, Avedis, “The Quality of Care: How Can It Be Assessed?,” *Journal of the American Medical Association*, 260 (1988), 1743–1748.

EIA (U.S. Energy Information Administration), “Form EIA-861 Data Files,” 2013, available at http://www.eia.gov/electricity/data/eia861/.

ENERNOC, “New Jersey Energy Efficiency Market Potential Assessment,” Technical Report, 2012, available at http://www.njcleanenergy.com/files/file/Library/NJ_Potential_Final_Report-Vol_1-Exec-Summary_2012-10-17.pdf.

Field, Erica, and Rohini Pande, “Repayment Frequency and Default in Microfinance: Evidence from India,” *Journal of European Economic Association Papers and Proceedings*, 6 (2008), 501–509.

Field, Erica, Rohini Pande, John Papp, and Natalia Rigol, “Does the Classic Microfinance Model Discourage Entrepreneurship among the Poor? Experimental Evidence from India,” *American Economic Review*, 103 (2013), 2196–2226.

Gine, Xavier, and Dean Karlan, “Group versus Individual Liability: Short and Long Term Evidence from Philippine Microcredit Lending Groups,” *Journal of Development Economics*, 107 (2014), 65–83.

Harrison, Glenn, and John List, “Field Experiments,” *Journal of Economic Literature*, 42 (2004), 1009–1055.

Heck, Stefan, and Humayun Tai, “Sizing the Potential of Behavioral Energy-Efficiency Initiatives in the U.S. Residential Market,” Technical Report, McKinsey & Company, 2013.

Heckman, James, “Randomization and Social Policy Evaluation,” in *Evaluating Welfare and Training Programs*, Charles Manski, and Irwin Garfinkel, eds. (Cambridge, MA: Harvard University Press, 1992).

Heckman, James, Hidehiko Ichimura, Jeffrey Smith, and Petra Todd, “Characterizing Selection Bias Using Experimental Data,” *Econometrica*, 66 (1998), 1017–1098.

Heckman, James, and Jeffrey Smith, “The Sensitivity of Experimental Impact Estimates: Evidence from the National JTPA Study,” NBER Working Paper 6105, 1997.

Heckman, James, Sergio Urzua, and Edward Vytlacil, “Understanding Instrumental Variables in Models with Essential Heterogeneity,” *Review of Economics and Statistics*, 88 (2006), 389–432.

Heckman, James, and Edward Vytlacil, “Policy-Relevant Treatment Effects,” *American Economic Review*, 91 (2001), 107–111.

———, “Structural Equations, Treatment Effects, and Econometric Policy Evaluation,” *Econometrica*, 73 (2005), 669–738.

Hellerstein, Judith, and Guido Imbens, “Imposing Moment Restrictions from Auxiliary Data by Weighting,” *Review of Economics and Statistics*, 81 (1999), 1–14.

Hotz, Joseph, “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, 1992.

Hotz, Joseph, Guido Imbens, and Jacob Klerman, “Evaluating the Differential Effects of Alternative Welfare-to-Work Training Components: A Reanalysis of the California GAIN Program,” *Journal of Labor Economics*, 24 (2006), 521–566.

Hotz, Joseph, Guido Imbens, and Julie Mortimer, “Predicting the Efficacy of Future Training Programs Using Past Experiences at Other Locations,” *Journal of Econometrics*, 125 (2005), 241–270.

Imbens, Guido, and Donald Rubin, *Causal Inference in Statistics and the Social Sciences* (Cambridge: Cambridge University Press, 2014).

JD Power, “Electric Utility Residential Customer Satisfaction Study,” Technical Report, 2014, available at http://www.jdpower.com/press-releases/2014-electric-utility-residential-customer-satisfaction-study.

Jones, Michael, “Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression,” *Journal of the American Statistical Association*, 91 (1996), 222–230.

Keim, Brandon, “The Hot New Frontier of Energy Research Is Human Behavior,” *Wired*, June 9, 2014, available at http://www.wired.com/2014/06/energy-and-human-behavior/.

Kling, Jeffrey, Jeffrey Liebman, and Lawrence Katz, “Experimental Analysis of Neighborhood Effects,” *Econometrica*, 1 (2007), 83–119.

LaLonde, Robert, “Evaluating the Econometric Evaluations of Training Programs with Experimental Data,” *American Economic Review*, 76 (1986), 604–620.

Leip, David, “Dave Leip’s Atlas of U.S. Presidential Elections,” 2013, available at http://uselectionatlas.org/.

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

Nolan, Jessica, Wesley Schultz, Robert Cialdini, Noah Goldstein, and Vladas Griskevicius, “Normative Influence Is Underdetected,” *Personality and Social Psychology Bulletin*, 34 (2008), 913–923.

Pew Center, “Energy Efficiency Standards and Targets,” available at http://www.c2es.org/us-states-regions/policy-maps/energy-efficiency-standards (accessed May 2011).

Pritchett, Lant, “It Pays to Be Ignorant: A Simple Political Economy of Rigorous Program Evaluation,” Mimeo, Kennedy School of Government, 2002.

Quackenbush, John, “Readying Michigan to Make Good Energy Decisions: Energy Efficiency,” Technical Report, Michigan Public Service Commission, 2013.

Rubin, Donald, “Estimating Causal Effects of Treatments in Randomized and Non-Randomized Studies,” *Journal of Educational Psychology*, 66 (1974), 688–701.

———, *Multiple Imputation for Nonresponse in Surveys* (New York: Wiley, 1987).

Schultz, Wesley, Jessica Nolan, Robert Cialdini, Noah Goldstein, and Vladas Griskevicius, “The Constructive, Destructive, and Reconstructive Power of Social Norms,” *Psychological Science*, 18 (2007), 429–434.

Smith, Jeffrey, and Petra Todd, “Does Matching Address LaLonde’s Critique of Nonexperimental Estimators?,” *Journal of Econometrics*, 125 (2004), 305–353.

Wennberg, David, F.L. Lucas, John Birkmeyer, Carl Bredenberg, and Elliott Fisher, “Variation in Carotid Endarterectomy Mortality in the Medicare Population,” *Journal of the American Medical Association*, 279 (1998), 1278–1281.

This page intentionally left blank