Go for the silver? Comparing quasi-experimental methods to the gold standard

Patrick Baylis* Peter Cappers† Ling Jin† Anna Spurlock†
Annika Todd†

May 16, 2016

Abstract

Randomized controlled trials (RCTs) are widely viewed as the "gold standard" for evaluating the effectiveness of an intervention. However, because RCTs are perceived to be prohibitively expensive and challenging to successfully, they are not broadly executed in policy settings. Previous work examining the effect of electricity pricing has largely been conducted through a two commonly used quasi-experimental methodologies: difference-in-differences and propensity score matching. Using a rare set of large-scale randomized field evaluations of electricity pricing, we compare the estimates obtained from these quasi-experimental designs and from a regression discontinuity design to the true estimates from the experimental method.

We identify three facts in our context that highlight the importance of understanding selection bias and spillover effects. First, difference-in-differences and propensity-score methods misestimate the true effect by up to 5% of mean peak hour usage. Second, regression discontinuity methods can be heavily biased relative to the true average treatment effect, an effect which seems to be driven by the sample limitations of the RD. Finally, across nearly all designs we find that experimental groups with low compliances rates tend to exhibit larger biases.

*Department of Agricultural and Resource Economics, UC Berkeley. Corresponding author.

†Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720. The work described in this paper was funded in part by the Office of Electricity Delivery and Energy Reliability, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 through Lawrence Berkeley National Laboratory.

1 Introduction

In this paper we scrutinize the effectiveness of several methodologies commonly used in the evaluation of electricity Demand Response (DR) and pricing programs. We compare them to the gold standard randomized, controlled trial (RCT) experimental evaluation methodology and find systematic evidence of selection and spillover effects that bias the non-experimental estimates.

Most empirical analysis in economics is conducted using observational data. Because these data are collected from complex real-world processes, conducting causal inference using ordinary least squares requires the maintenance of untestable assumptions regarding the data generating process. To relax these assumptions and to provide more credible estimates of causal effects, empirical social scientists are turning with increasing regularity to RCTs, a method that has been more typically used in fields such as public health and psychology.

RCTs are considered to be the "gold standard" research design in empirical social science because the randomization process holds potential confounding factors equal across control and treatment groups, allowing the researcher to isolate the treatment effect of interest. For this reason, estimates obtained using properly implemented experimental designs are correctly viewed as reflecting the "true" estimate. However, the gold standard comes at a price: RCTs can be expensive, time-consuming, and challenging to implement correctly. They can be limited to settings where an experimental intervention is feasible, and are subject to concerns regarding external validity.

Meanwhile, a long history of empirical work has used an array of quasi-experimental research designs intended to simulate the experimental process, such as matching, propensity score weighting, regression discontinuity, and within-unit estimators. These research designs, because they rely on observational data, are less expensive and difficult to implement compared to RCTs, and can often be applied after a program has taken effect, which can disincentivize the need to plan for evaluation carefully at the program implementation stage. However, because they lack the randomized component of experimental designs, selection concerns and other forms of classical omitted variable bias can generate bias in the results obtained from these quasi-experimental methods. Without an experimental comparison, it is usually impossible to definitively ascertain whether the research design reflects the true estimate or these unobserved biases.

This paper builds on prior work in the peer-reviewed literature that compares results obtained using non-experimental research designs with experimental results. Much of the seminal work in this area was conducted in the labor literature. LaLonde (1986) conducts such a comparison in the context of an employment training program, finding that the non-experimental estimates frequently fail to align with the experimental results. Heckman, Ichimura, and Todd (1997) analyze a separate program and find that non-experimental estimates can perform well so long as the comparison samples are drawn from a similar sample. Dehejia and Wahba (2002) find that propensity score estimates can outperform traditional econometric estimators, although Smith and Todd (2001) note that the former finding may be due to the sample selection imposed.

Some recent work has extended this type of analysis to data from the electricity industry, which is the setting we use in this paper. A recent working paper by Jessoe, Miller, and Rapson (2015) examines the possibility of using high-frequency electricity data to recover causal effects without

an experimental comparison group.

An advantage of our approach beyond the previous work comparing experimental to non-experimental methodologies is that we use an experiment with multiple treatment arms to validate trends in our results. Whereas most prior work has relied on a single treatment arm within a single experimental setting, we use the multiple treatment arms to look for evidence of trends in the results[1]. The results from implementing the quasi-experimental estimators across all treatment arms allow us to ascertain in which cases there are consistent biases relative to the experimental estimates across all treatment arms, and quantify these biases on average. In particular, we provide strong suggestive evidence that selection biases and spillover effects drive the observed biases in the quasi-experimental results within the program we evaluate.

1.1 Empirical context: electricity pricing programs

In this paper we focus on a case from the electricity industry. Accurate evaluation of Demand Response (DR) and pricing programs in the electricity industry is important for several reasons. First, settlement and payment of incentives for incentive-based programs (such as peak time rebates) require an accurate evaluation of how consumption for a specific household changed on a single critical day relative to their baseline (or counterfactual) consumption. In these programs, customers are paid for the amount of electricity they saved on a given critical day relative to this baseline. Second, utilities often claim savings and recover costs from ratepayers as authorized by regulators, and these savings need to be accurately measured through a program impact evaluation. Third, an assessment of how well a program is working is crucial for future program and portfolio planning, so that ratepayer dollars are spent on programs that achieve the highest savings at the lowest cost. Fourth, accurate short- and long-term grid-level energy and capacity forecasts are necessary for maintaining reliability. These forecasts enter into resource planning efforts that inform the need for future infrastructure investment. Accurately predicting the effects of time-based rates and incentive-based programs on energy and peak demand can help with planning for that investment.

There are a variety of evaluation methods and protocols used by the electricity industry that differ by the type of rate or program being evaluated, budget constraints, and historical experience. Up to now RCTs have been met with substantial resistance. Concerns that have been raised include: they require substantial planning up front at the program implementation phase, rather than quasi-experimental techniques which typically require analysis only ex-post; they are seen as difficult to implement; and they are sometimes described as unfair because they restrict program participation to exclude the control group. As such, the majority of the evaluation methods used historically have been quasi-experimental. The specifics of these methods will be outlined later in the paper.

However, there has been a recent increase in interest in the application of randomized[2] evaluation methods in the electricity industry. This trend was spurred forward with the increased visibility and popularity of behavior-based programs, such as Opower's Home Energy Reports, *e.g.*, Allcott (2011b). The average effect sizes are quite small for behavior-based programs, and so regulators

1. Future versions of this work will incorporate multiple settings.
2. Either through RCTs or Randomized Encouragement Designs (REDs), which are similar but allow for selection into treatment within a randomized encouragement context.

required a higher bar for accurate and reliable evaluation to claim savings than had been applied to other types of programs historically. The discussion around RCTs in the context of behavior-based programs, however, facilitated the expansion of these methods beyond these programs alone.

In the context of time-based pricing programs, in 2009, the United States Department of Energy issued a funding opportunity announcement for its Smart Grid Investment Grant (SGIG) that requested proposals from utilities seeking funding to expand their smart meter infrastructure. It was required that these utilities include randomized pricing experiments to be enabled by this investment in advanced metering infrastructure in their proposal. Many in the industry were skeptical that utilities would be willing to propose such activities due in part to concerns that utilities would be unable to obtain local regulatory approval to implement pilots using randomization. However, ten utilities were ultimately funded under SGIG and undertook Consumer Behavior Studies (CBS) that utilized randomized evaluation methodology for their pricing pilots.

There is a long history from both inside and outside of economics documenting the effects of time-varying pricing on customer behavior. Academic researchers have typically focused on the fairly small set of experiments that have been conducted on time-varying pricing. Aigner (1984), Train and Mehrez (1994), and Jessoe, Rapson, Bushnell, et al. (n.d.) analyze the effect of separate time-of-use (TOU)[3] experiments. Allcott (2011a) analyses a real-time pricing (RTP)[4] experiment. Wolak (2007) examines the response to a critical peak pricing (CPP)[5] program. The fact that past instances of randomized experiments are relatively limited is indicative of the resistance we've mentioned to these methods in this industry historically.

We use the opportunity offered by the randomized time-based rate pilots under the SGIG CBS in order to assess the performance of the quasi-experimental designs most commonly employed to evaluate DR and pricing programs historically. Building on the pioneering work by LaLonde (1986), we take a set of electricity pricing experiments as the gold standard against which we compare our set of quasi-experimental estimates. Because electricity consumption is a data-rich context, we are able to implement a wide range of quasi-experimental techniques. Specifically, we estimate two difference-in-differences designs (DID), a propensity score estimator that reweights observations by their treatment likelihood, and a regression discontinuity (RD) design that discontinuously influences treatment likelihood. We compare the estimates of the average treatment effect obtained using these quasi-experimental techniques to the correct estimate obtained from the experimental methods.

We document empirical support for three general results. First, RD methods tend to overestimate the size of the true average treatment effect, underlining the limitation of RD to provide externally valid estimates. Second, difference-in-difference and propensity score methods tend to underestimate the effect, suggesting the presence of selection bias when using these methods. Third, biases in non-experimental research designs tend to be more pronounced in opt-in treatments relative to opt-out treatments, further confirming the selection effect interpretation[6].

3. With a TOU price structure the price for peak hours is higher than off-peak hours, and the definition of peak hours (e.g., 4-7pm on non-holiday weekdays) is fixed.

4. With an RTP price structure, the price varies continuously over time to better track variation in wholesale prices.

5. With a CPP price structure, the price is much higher during the pre-established peak hours of a finite set of event days which the utility calls in advance based on predicted grid conditions.

6. The opt-out experimental designs result in much higher enrollment (over 90%) compared to opt-in (around 20%), which

For policy-makers, this work contributes to our understanding of the usefulness of quasi-experimental designs as ex-post measurement of changes in consumption as a result of electricity rate design. Many utilities and public utilities commissions are considering a broader implementation of time-based pricing of electricity in the next decade. Policymakers may want to test the effects of these changes, but may not have the resources to implement a full RCT[7]. Our results suggest the following: first, difference-in-differences and propensity-score methods mis-estimate the true effect by up to 5% of mean peak hour usage. Second, propensity score estimates resemble difference-in-difference findings, but standard errors tend to be larger and point estimates are more biased for opt-out models. Third, regression discontinuity methods can be heavily biased relative to the true average treatment effect. Finally, we find strong evidence that biases are more pronounced in opt-in vs. opt-out designs.

The remainder of the paper is organized as follows. Section 2 describes the underlying econometric models and identifying assumptions required for the experimental and quasi-experimental designs we test in this paper. Section 3 describes examples from the evaluation community that use the previously described approaches, and 4 documents the experimental context in which we test these approaches. Section 5 presents results and stylized facts, and section 6 concludes.

2 Econometric background

Any estimation of a causal effect must contend with the fundamental problem of causal inference: it is impossible to simultaneously observe sample units in both treated and untreated states. In the context of estimating the effect of electricity pricing treatments, this means that researchers cannot observe how much electricity a control customer would have demanded had she been exposed to the treatment or how much a treatment customer would have demanded had she not been treated. Experimental methods circumvent this problem by randomizing, while quasi-experimental methods use a variety of techniques to claim that treatment is "as good as random." We formalize this relationship using the potential outcomes framework[8], writing the observed outcome for a given unit i as:

$$y_i = y_{0i} + (y_{1i} - y_{0i})D_i$$

D_i is a binary indicator of whether unit i is treated, y_{0i} is the outcome if i is not treated, and y_{1i} is the outcome if i is treated. Note that the expression $y_{1i} - y_{0i}$ captures the causal effect of treatment on unit i and is unobservable due to the fundamental problem of causal inference. Instead, researchers are often interested in estimating the average treatment effect (ATE), $E[y_i|D_i = 1] - E[y_i|D_i = 0]$, the difference between the average outcome if all units were treated and if all units were untreated.

In order to estimate the ATE, investigators must assume a set of conditions on the data generating process that will vary with the setting and research design. In a randomized experiment, assignment to treatment is random and the estimation of the ATE requires relatively few assumptions. In a

means there is more selection present with an opt-in design compared to an opt-out design.

7. We note that the existence of the present set of RCTs is due to a large DOE grant, which also funds this study.

8. Otherwise known as the Rubin Causal Model (Rubin 1974). The exposition that follows draws from Angrist and Pischke (2008).

quasi-experiment, assignment to treatment is non-random but may be plausibly random after conditioning on the appropriate covariates. We proceed by specifying the assumptions required for the randomized experiment and for each quasi-experimental design we compare to the experimental results.

Econometrically, the goal in any evaluation is to ensure that the error term (capturing any and all unobserved forces) is uncorrelated with the independent variable of interest. For example, in an electricity pricing setting, it must be assumed that households who participate in a new pricing program are not systematically different in ways that affect their electricity consumption compared to households that do not participate. In a randomized setting, this assumption is known to be true, by virtue of the randomization itself. In quasi-experimental settings, this assumption cannot be proved, but must be claimed. The following section provides an overview of the research designs we estimate, with an emphasis on the assumptions required to overcome the fundamental problem of causal inference[9].

2.1 Experimental design

The key feature of RCTs is that units are assigned randomly between control and treatment groups. Proper randomization and sufficient sample size should ensure that these two groups are similar across both observable and unobservable attributes. If this is the case, then any differences in the average outcome between the control and treatment groups should be entirely attributable to the treatment itself. Because the assignment mechanism is random, we know that the *potential* outcomes y_{0i} and y_{1i} are independent from the actual treatment assignment D_i. It is useful to proceed by characterizing the estimation procedure in a regression context:

$$y_i = \alpha + \beta D_i + \varepsilon_i$$

We can show that

$$\beta = \overbrace{\big(E[y_i|D_i = 1] - E[y_i|D_i = 0]\big)}^{\text{ATE}} - \overbrace{\big(E[\varepsilon_i|D_i = 1] - E[\varepsilon_i|D_i = 0]\big)}^{\text{Selection bias}} \tag{1}$$

If the potential outcomes y_{0i} and y_{1i} are uncorrelated with treatment status, it can be shown that the idiosyncratic error term ε_i is as well, which implies that the randomization has removed selection bias, as expected.

In our context, we randomly assign customers to treatment and control groups. To account for statistically insignificant but slight differences in the pre-treatment consumption of the two groups, we estimate the difference between the average change in electricity usage in the pre-treatment period and the post-treatment period between the treatment and control groups. Because not all customers from the treatment groups actually enrolled in the program, we are actually using a Randomized Encouragement Design (RED), which allows us to estimate the average effect of taking up the treatment. This design requires the additional assumption that treatment status (*i.e.*, being

9. For a detailed explanation of different types of impact evaluations for electricity experiments see Cappers et al. (2013).

encouraged to enroll) did not affect energy usage except by causing enrollment.

2.2 Non-experimental designs

If treatment assignment is nonrandom, we can't assume that the third and fourth terms in equation 1 are equal to zero, and the straightforward comparison of means may be biased due to differences between the type of customers who select into treatment.

Quasi-experimental techniques do not assume that treatment is unconditionally randomly assigned. Instead, they use different sources of identification to isolate the treatment effect from other determinants of the outcome variable. We discuss three common quasi-experimental approaches: difference-in-differences, propensity score matching, and regression discontinuity designs. In section 1.1 we discuss the implementations of these techniques, which can sometimes combine elements from more than one approach. We consider the base cases here to capture the range of possible approaches.

2.2.1 Difference-in-differences

Difference-in-differences estimators compare the difference in pre- and post-treatment electricity usage between treated and control customers. The identifying assumption is called the "parallel trends assumption", which is that the change in the control group is an appropriate counterfactual for the change the treatment group would have experienced. To see this, we extend the original setting to include two time periods, before and after treatment. Every customer i is in a group $g \in \{0, 1\}$ and is observed in each time period $t \in \{0, 1\}$. Again, using the regression context, researchers estimate the following model:

$$y_{igt} = \alpha + \beta_0 \text{POST}_t + \beta_1 \text{TREAT}_g + \tau D_{gt} + \overbrace{\mu_{gt} + \varepsilon_{igt}}^{\text{Error term}} \tag{2}$$

Note that $D_{gt} = \text{POST}_t \times \text{TREAT}_g$, where POST_t indicates a post-treatment observation and TREAT_g indicates a member of the treatment group. The coefficient of interest is τ[10]. The identifying assumption for τ is that differential changes between the two groups in the pre- and post-period are zero in expectation, or that $E[\mu_{11} - \mu_{10}] = E[\mu_{01} - \mu_{01}]$.

The validity of this assumption in our setting depends on the construction of the treatment and control groups. If the treatment group is composed of customers who selected into treatment and the control group is composed of the remaining customers, there is a strong possibility that the parallel trends assumption is violated. Suppose, for example, that treatment group customers are more energy conscious and are less likely to turn their air conditioners on during hot days. If the post-treatment period is warmer than the pre-treatment period, then τ will be biased away from zero.

One way to mitigate selection bias is to choose a control group without access to the treatment. Although the treatment group will remain selected, the control group is less likely to be substantially

10. Note that $\tau = \left(E[y_i^{\text{POST}}|D_i = 1] - E[y_i^{\text{PRE}}|D_i = 1]\right) - \left(E[y_i^{\text{POST}}|D_i = 0] - E[y_i^{\text{PRE}}|D_i = 0]\right)$.

different. If there does not appear to be substantial selection into treatment, then this could reduce the total bias. Here we note that there may be important differences between the opt-in and opt-out treatments. Since the opt-in treatments enrolled at most 20% of treated customers, it is likely that there is a substantial selection effect. However, since the opt-out treatments enrolled at least 90%, selection is likely to be more muted in this sample.

2.2.2 Propensity score matching

Our third quasi-experimental technique uses a standard propensity-score matching approach to account for selection into treatment. We construct estimates of each customer's enrollment likelihood based on their pre-treatment electricity usage. We then estimate a regression that adjusts for differences due to selection into treatment using the propensity score. There are a variety of ways to use the propensity score in a regression framework, but all rely the same conditional independence assumption: that treatment assignment is random after conditioning on the covariates. The propensity score simply provides a tractable way to condition.

The propensity score is a function that determines how likely a unit is to be treated based on their observables: $p(X_i) = E[D_i|X_i]$, typically estimated with a logit or a probit model to constrain $0 < p(X_i) < 1$. A straightforward way to use the propensity score is to simply include it in the regression:

$$y_i = \alpha + \beta_0 p(X_i) + \tau D_i + \varepsilon_i \tag{3}$$

The coefficient of interest here is τ, and the identification assumption is $y_{0i}, y_{1i} \perp\!\!\!\perp D_i|X_i$.

In practice, implementations of the propensity score vary widely and can incorporate other matching components as well as difference-in-difference techniques.

2.2.3 Regression discontinuity

Regression discontinuity (RD) designs take advantage of a cutoff c that alters the probability of treatment but not other factors which might affect the potential outcomes. Suppose there is some running variable X_i s.t. that when $X_i > c$, $D_i = 1$. If $X_i <= c \Rightarrow D_i = 0$. Researchers can exploit this threshold to estimate the effect of treatment by confining the sample to units with $c - h < X_i < c + h$, where h is some reasonable bandwidth.

$$y_i = \alpha + \beta_0(X_i - c) + \beta_1(X_i - c) \times D_i + \tau D_i + \varepsilon_i \tag{4}$$

The coefficient of interest is τ, and the identifying assumption is that $E[y_{0i}|X = x]$ and $E[y_{1i}|X]$ are continuous in x.

In the electricity context, a relevant cutoff might be generated if a program offers time-varying pricing to any customers with total pre-treatment period summer electricity usage above a given threshold but not to those below. The underlying assumption is that customers above and below the treatment threshold are similar except in their ability to join the pricing program. In essence

the assumption is that customers cannot anticipate the cutoff and manage their consumption such that they are able to orchestrate their qualification, or not, for treatment.

3 Use of quasi-experimental methods in the electricity pricing evaluation literature

The evaluation community has used these quasi-experimental approaches widely. For each approach used, there are many possible variations in the implementation, the details of which are determined on a per-evaluation basis and reflect the empirical context as well as the expertise of the evaluating team. However, the underlying identification techniques are identical across variations. It is worth noting that the potential biases associated with these approaches are generally recognized by evaluators, but because of the way the program was implemented there is no way to correct this after program implementation. The following are a few examples of their application.

3.1 Propensity-score methods

Because there are many ways to use propensity scores in evaluation, the approaches in the evaluation literature vary with context and available data. Propensity score matching techniques use $p(X_i)$ as a distance metric to construct matches. These approaches matches on some combination of load shapes, usage variables, and customer characteristics (George et al. 2014; Bell 2015; Savage and George 2015; Bell 2015). In particular, we modeled our application of the propensity score matching method off of the one employed in Savage and George (2015), which examined the effect of TOU pricing in PG&E.

3.2 Difference-in-differences

By contrast, the difference-in-differences techniques in the evaluation literature tend to be more standard: most studies employ a difference-in-differences approach with a selected treatment sample compared to a random control sample that was not offered the treatment (McAuliffe and Rosenfeld 2004; Violette, Erickson, and Klos 2007; Lutzenhiser et al. 2009)

3.3 Regression discontinuity

By contrast, regression discontinuity designs are not widely used in the evaluation. However, we include them here because we believe they represent a low-cost alternative to experimental designs. Rather than implementing a full randomized experiment, forward-thinking evaluators could implement treatment thresholds in advance in order to facilitate *ex post* evaluation. Jessoe, Rapson, and Smith (2014) offer one example from the academic literature.

4 Overview of field experiment

4.1 Random assignment

SMUD's customer base has approximately 530,000 residential households; some were excluded from the eligible experimental population. After these exclusions, approximately 174,000 households remained eligible[11].

There were two pricing treatments that differed from the standard rate: a time-of-use (TOU) program where customers faced higher prices 4pm to 7pm on non-holiday weekdays, and a Critical Peak Pricing (CPP) pricing program where they faced very high prices during the peak period of twelve critical event days called a day in advance over the course of each of two summers. Both programs were in effect between June 1 and September 30th for the two summers in the study (2012 and 2013)[12]. In addition, there was an enabling technology associated some of the treatment groups in which customers were offered in-home displays.

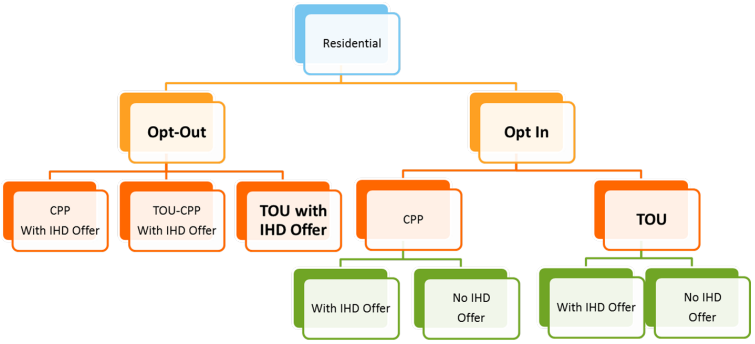
Households in the experimental population were randomly assigned into ten groups; for most of this paper, we examine seven of those groups, seven of which were encouraged to participate in a TOU or CPP treatment, while the seventh group was the control group, which received no encouragement and remained on the standard rate. There were two forms of encouragement: opt-in, where households were encouraged to enroll in the rate program; or opt-out, where households were notified that they were enrolled and were encouraged to stay in the rate program, but had the opportunity to leave the program if they wished[13]. Figure 1 displays the seven treatment arms we use in this paper.

11. Households were excluded from our experiment if: they did not have interval meters to capture hourly electricity usage installed prior to June 2011; they were participating in SMUD's Air Conditioning Load Management program, Summer Solutions study, PV solar programs, budget billing programs, or medical assistance programs; or if they had master metered accounts.

12. During the time period of our study, non-EAPR customers (EAPR refers to Energy Assistance Program Rate customers. This is SMUD's low-income rate) on SMUD's standard rate plan (i.e., customers in the control group) paid a $10 monthly fixed charge plus 0.0938 per kWh for the first 700 kWh of consumption and $0.1765 for consumption above 700 kWh. Under the TOU program, customers paid $0.2700 per kWh for electricity consumed from 4PM to 7PM on non-holiday weekdays, plus a monthly fixed charge and $0.0846 per kWh for the first 700 kWh and $0.1660 for consumption above 700 kWh, where on-peak consumption did not count towards the 700 kWh total. Customers on the CPP plan paid $0.7500 per kWh for consumption between 4PM and 7PM on twelve "event days" over the course of the summer. Customers were alerted about event days at least one day in advance. Consumption outside of the CPP event window was charged at a rate of $0.0851 per kWh up to 700 kWh and $0.1665 beyond.

13. Households who were encouraged to participate in an opt-in rate program were solicited through many channels, including direct mail letters, door hangers, and an outbound calling campaign. The messages listed generic benefits of participating in rate programs, including saving money, taking control, and helping the environment. Households who were encouraged to remain in an opt-out program were notified through a direct mail letter that they had been placed on the rate, and told to contact SMUD if they wished to drop-out. The TOU Opt-in group received encouragement messages that were slightly different than the other groups, because they were also part of a recruit-and-delay randomized controlled trial (which we are no incorporating into this paper). Their messages contained text that informed them that if they decided to opt-in to the rate program, they would be randomly assigned to a start date of either 2012 or 2014 (i.e., they may be delayed in experiencing treatment). The other three groups were told that their participation date would start in 2012 if they decided to opt-in or not opt-out. This means that while the CPP opt-in group can be directly compared to the CPP opt-out group, there is a caveat to the comparison between the TOU opt-out and opt-in groups given the slight different wording in the recruitment materials.

Figure 1: SMUD treatment arms



4.2 Data

We use hourly energy consumption data (in kW) for each household in our control group, as well as for each household in our seven treatment groups, regardless of whether or not they ended up enrolled on the treatment pricing, and whether or not they opted out at any point in the pilot period. This was collected for one year prior to the start of the pilot period (June 1st, 2011 âĂŞ May 31st, 2012) and two years during the pilot period (June 1st, 2012 - September 30th, 2013).

A comparison of pre-treatment energy usage, available in the appendix, documents no statistical difference between the control group and each of the seven experimental treatment groups (including average kWh per day, peak hours, and peak to off peak ratio).

We also use hourly weather data, including dry and wet bulb temperature as well as humidity. There is only one weather station in close proximity to all participants in the SMUD service area, so the weather data does not vary across households, only over time.

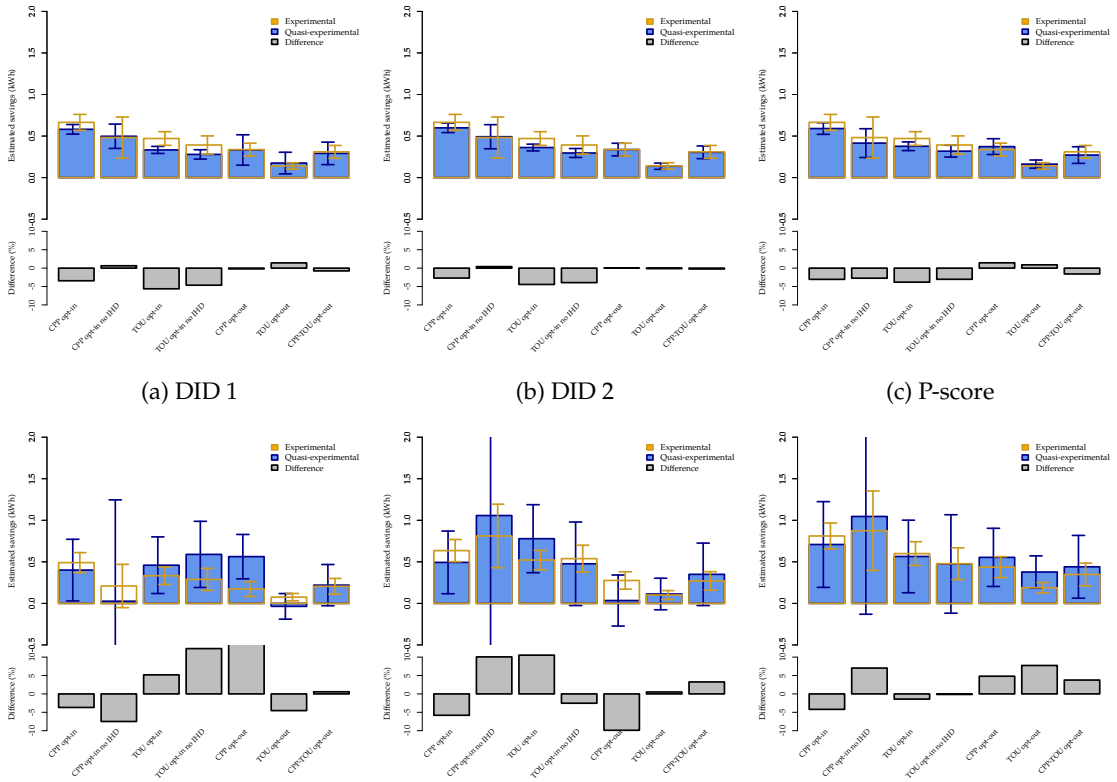
5 Results

Figure 2 summarizes the differences between the average treatment effect estimated using the field experiment and those obtained using the quasi-experimental approaches described in the previous section. For each quasi-experimental approach, the central dot represents the difference between the experimental estimates and the corresponding quasi-experimental estimates averaged across the treatment arms, and expressed as a percent of average hourly electricity consumption. The error bands document the average lower and upper 95 percent confidence interval of this value.

5.1 Difference-in-differences and propensity-score methods mis-estimate the true effect by up to 5% of mean peak hour usage

The difference-in-differences approaches and the propensity-score method mis-estimate the effect of the treatment relative to the randomized design in all of the opt-in treatment arms. To interpret

Figure 2: Comparing RCT and quasi-experimental estimates

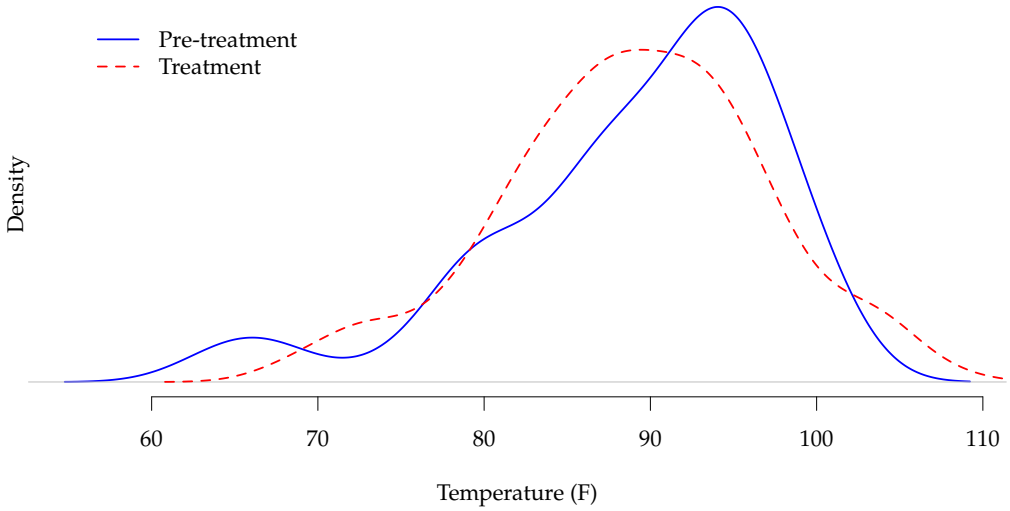


(a) DID 1 (b) DID 2 (c) P-score

(d) RD 1 (e) RD 2 (f) RD 3

Notes: Each plot compares the set of estimates obtained with a quasi-experimental technique to the RCT estimate across seven different treatment arms. The top subplot in each quadrant is absolute value of the treatment effect with standard errors and the bottom subplot is the difference between the RCT estimate and the quasi-experimental estimate. Blue bars are the quasi-experimental estimates, yellow bar outlines are the RCT estimates.

Figure 3: Temperature distribution by pre- and post-treatment periods



the result, we recall the design of the difference-in-difference estimators, which compare the change in average usage for each customer relative to his or her pre-treatment average across control and treatment groups. Importantly, the treatment group in this design consists entirely of customers who deliberately select into time-varying pricing. This group is observationally different from the control group and is likely to have different electricity usage patterns. We interpret the difference between the DID estimates and the RCT estimates as driven by this selection effect: customers who actively chose to participate in the time-varying pricing program are more energy conscious than those who did not and had different underlying trends, biasing the result downwards. We note that this bias could have been either towards or away from zero, depending on trends in weather. In the case of this study, weather in the pre-treatment period was warmer than weather in the post-treatment period; see figure 3.

5.2 Propensity score estimates resemble difference-in-difference results, more biased for opt-out

While the propensity-score results are similar for the opt-in groups, they are more biased for opt-out. To understand this result, it is useful to consider the construction of the propensity-score estimator: only control groups whose covariates match closely to a treatment unit are included in the analysis. In this case, the large size of the control group may have been an advantage.

5.3 Biases are more pronounced in opt-in vs. opt-out designs

In all designs, estimation of the average effect of the opt-out treatments is less biased than the opt-in treatments. We interpret this finding as strong evidence of a selection effect: because around 20% of individuals chose to opt-in to treatment when offered, the sample obtained using an opt-in enrollment method is likely to be more heavily selected than that obtained using an opt-out enrollment method, which achieved 90% enrollment. Because the difference-in-differences, propensity score, and RD approaches are potentially subject to sample selection biases, using a less-selected sample to begin with naturally improves the quality of the quasi-experimental estimate.

6 Discussion

Using a rich set of field experiments designed to test customer response to time-varying pricing, we estimate and compare a set of established quasi-experimental designs to their corresponding experimental estimates. By comparing across multiple treatment arms we are able to provide support for a set of stylized facts, each of which has important policy implications for ex post estimation of time-varying pricing programs.

First, we document that DID estimates which compare a self-selected treatment group with a control group who either did not choose or were not offered the opportunity to enroll in the program are likely to reflect bias even after including a rich set of fixed effects. In our setting, weather variation between the pre- and post- period likely caused the DID estimate to be biased towards zero. Second, we find that propensity-score matching techniques do not substantially reduce bias relative to the DID estimates, but increase standard errors due to the reduction in the effective size of the control group. Third, we show that even well constructed RD estimates can be biased away from the treatment estimate due to energy use level differences between the treatment and control groups. Finally, we observe that selection biases are more pronounced in all designs under opt-in treatments as compared to opt-out treatments. This finding strongly suggests that policy-makers should take this into account when designing the enrollment mechanism for a time-varying pricing program: in addition to being less costly and more effective at reducing total electricity usage, ex post estimation of opt-out designs using quasi-experimental designs are less likely to be unbiased. Our final two stylized facts related to the estimation of individual event day energy use reductions: we find that comparisons to high temperature non-event days (a common approach in incentive-based peak time rebate or critical peak pricing programs) tend to overestimate the actual reduction. We additionally find that estimates using a within-customer approach that compares the reduction during event hours to reductions during non-event hours tend to underestimate savings, likely as a result of spillover effects.

We caution that our results are limited to a set of treatment arms in a single experimental setting, and we emphasize that the direction of the biases in the quasi-experimental estimates is not necessarily likely to be stable in other contexts. Instead we suggest our results demonstrate the importance of careful consideration in research design: where possible, researchers and policy-makers should rely on true experiments. In other cases, attention should be given to underlying

trends in treatment and control groups when interpreting quasi-experimental results and when possible opt-out enrollment mechanisms should be implemented.

References

Aigner, Dennis J. 1984. "The welfare econometrics of peak-load pricing for electricity: Editor's Introduction." *Journal of Econometrics* 26 (1): 1–15.

Allcott, Hunt. 2011a. "Rethinking real-time electricity pricing." *Resource and Energy Economics,* Special section: Sustainable Resource Use and Economic Dynamics, 33, no. 4 (November): 820–842.

———. 2011b. "Social norms and energy conservation." *Journal of Public Economics* 95 (9-10): 1082–1095. arXiv: `arXiv:1011.1669v3`.

Angrist, Joshua D., and Jorn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion*. arXiv: `arXiv:1011.1669v3`.

Bell, Eric. 2015. *2014 Load Impact Evaluation of Southern California Edison's Peak Time Rebate Program.* Technical report.

Cappers, Peter, Annika Todd, Richard Boisver, and Michael Perry. 2013. "Quantifying the Impacts of Time-Based Rates, Enabling Technology, and Other Treatments in Consumer Behavior Studies." *Technical Report:* 142.

Dehejia, Rajeev H., and Sadek Wahba. 2002. "Propensity Score-Matching Methods for Nonexperimental Causal Studies." *Review of Economics and Statistics* 84, no. 1 (February): 151–161.

George, Stephen, Josh Schellenberg, Jeeheh Oh, and Marshall Blundell. 2014. *2013 Load Impact Evaluation of San Diego Gas & Electric Company's Opt-in Peak Time Rebate Program.* Technical report.

Heckman, James J., Hidehiko Ichimura, and Petra E. Todd. 1997. "Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme." *The review of economic studies* 64 (4): 605–654.

Jessoe, Katrina, Douglas Miller, and David Rapson. 2015. "Can high-frequency data and non-experimental research designs recover causal effects? Validation using an electricity usage experiment." *Working Paper.*

Jessoe, Katrina, David Rapson, Jim Bushnell, Colin Cameron, Scott Carrell, Michael Carter, Meredith Fowlie, et al. n.d. "Knowledge is (Less) Power: Experimental Evidence from Residential Energy Use." *Working Paper.*

Jessoe, Katrina, David Rapson, and Jeremy B. Smith. 2014. "Towards understanding the role of price in residential electricity choices: Evidence from a natural experiment." *Journal of Economic Behavior & Organization* 107:191–208.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *The American Economic Review* 76, no. 4 (September): 604–620. arXiv: arXiv:1011.1669v3.

Lutzenhiser, Susan, Jane Peters, Mithra Moezzi, and James Woods. 2009. *Beyond the Price Effect in Time-of-Use Programs: Results from a Municipal Utility Pilot , 2007-2008.* Technical report.

McAuliffe, Pat, and Arthur Rosenfeld. 2004. *Response of residential customers to critical peak pricing and time-of-use rates during the summer of 2003.* Technical report.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of Educational Psychology* 66 (5): 688–701.

Savage, Aimee, and Stephen George. 2015. *PG&E's Residential TOU Program.* Technical report.

Smith, Jeffrey A., and Petra E. Todd. 2001. "Reconciling conflicting evidence on the performance of propensity-score matching methods." *The American Economic Review:* 112–118.

Train, Kenneth, and Gil Mehrez. 1994. "Optional time-of-use prices for electricity: econometric analysis of surplus and Pareto impacts." *The RAND Journal of Economics:* 263–283.

Violette, Dan, Jeff Erickson, and Mary Klos. 2007. *Final Report for the MyPower Pricing Segments Evaluation.* Technical report.

Wolak, Frank A. 2007. "Residential Customer Response to Real-time Pricing: The Anaheim Critical Peak Pricing Experiment." *Center for the Study of Energy Markets.*