Machine Learning from Schools about Energy Efficiency

Fiona Burlig
UC Berkeley

Christopher Knittel
MIT

David Rapson
UC Davis

Mar Reguant
Northwestern University

Catherine Wolfram*
UC Berkeley

March 16, 2017

PRELIMINARY, COMMENTS WELCOME.
Click here for latest version

Abstract

We study the impacts of energy efficiency investments at public K-12 schools in California. We leverage high frequency data – electricity use every 15 minutes – to develop several approaches to estimating counterfactual energy consumption in the absence of the efficiency investments. In particular, we use difference-in-differences approaches with rich sets of fixed effects. We show, however, that these estimates are sensitive to the set of fixed effects included and to the set of schools included as controls. To address these concerns, we develop and implement a novel machine-learning approach to predict counterfactual energy consumption at treated schools and validate the approach with non-treated schools. We find that the energy efficiency projects in our sample reduce electricity consumption between 2 to 5% on average, which can result in substantial savings to schools. We also compare our estimates of the energy savings to ex ante engineering estimates. Realized savings are generally less than 50% of ex ante forecasts and quite low for measures other than heating and air-conditioning systems or lighting.

*Burlig: Agricultural and Resource Economics and Energy Institute at Haas, UC Berkeley, fiona.burlig@berkeley.edu. Knittel: William Barton Rogers Professor of Energy Economics, Sloan School of Management, and Director of the Center for Energy and Environmental Policy Research, MIT, and NBER, knittel@mit.edu. Rapson: Associate Professor of Economics, UC Davis, dsrapson@ucdavis.edu. Reguant: Assistant Professor of Economics, Northwestern University and NBER, mar.reguant@northwestern.edu. Wolfram: Cora Jane Floor Professor and Director of the Energy Institute at Haas, Haas School of Business, UC Berkeley, and NBER, cwolfram@berkeley.edu. We thank Dan Buch, Arik Levinson, and Ignacia Mercadal, as well as seminar participants at the Energy Institute @ Haas Summer Energy Camp, MIT, the Colorado School of Mines, the University of Arizona, Arizona State University, Texas A & M, Iowa State University, Boston College, the University of Maryland, and the NBER Summer Institute for useful comments. We thank Joshua Blonz and Kat Redoglio for excellent research assistance. We gratefully acknowledge financial support from The California Public Utilities Commission.

1 Introduction

Energy efficiency is a cornerstone of global greenhouse gas (GHG) abatement efforts. For example, worldwide proposed climate mitigation plans rely on energy efficiency to deliver 42 percent of emissions reductions (IEA, 2015). The appeal of energy efficiency investments is straightforward: they may pay for themselves by lowering future energy bills. At the same time, lower energy consumption reduces reliance on fossil fuel energy sources, providing the desired GHG reductions. A number of public policies – including efficiency standards, utility-sponsored rebate programs, and information provision requirements – aim to encourage more investment in energy efficiency.

Policymakers are likely drawn to energy efficiency because a number of analyses point to substantial unexploited opportunities for cost-effective investments (see, e.g., McKinsey & Company (2009)). These analyses are almost universally based on engineering estimates of the potential energy savings over time rather than field evidence of actual savings. One strand of the economics literature has attempted to explain why consumers might fail to avail themselves of profitable investment opportunities (see, e.g., Allcott and Greenstone (2012); Gillingham and Palmer (2014); Gerarden et al. (2015)). The most popular explanations have emphasized the possibility of market failures, such as imperfect information, capital market failures, split incentive problems, and behavioral biases, including myopia, inattentiveness, prospect theory and reference-point phenomena. The literature also points to the possibility that engineering estimates understate the costs consumers face.

Economists have also pointed out that accurately measuring the returns to energy efficiency investments is difficult as it requires constructing a counterfactual energy consumption path from which reductions caused by the efficiency investments can be measured (Joskow and Marron, 1992). Recent studies use both experimental (e.g., Fowlie et al. (2015)) and quasiexperimental (e.g., Levinson (2016), Myers (2014), Davis et al. (2014)) approaches to developing this counterfactual. Many of these papers suggest that the energy savings from efficiency investments are lower than expected, and the results cast doubt on the extent to which emissions reductions can be achieved through energy efficiency. At the same time, energy efficiency proponents have criticized these findings as only pertaining to specific contexts (Kushler, 2015), and are skeptical that experimental and quasi-experimental estimation approaches can be applied broadly. In general, there is a clear need for techniques to estimate returns to energy efficiency programs that can be applied in a wide set of contexts.

The proliferation of high-frequency data in electricity markets provides a promising opportunity to estimate treatment effects associated with energy efficiency investments wherever advanced metering infrastructure (AMI, or "smart metering") is installed. In this paper, we use high-frequency data to implement several approaches to estimating counterfactual energy consumption absent the energy efficiency investments. We begin with difference-in-difference approaches that employ rich sets of fixed effects. We show, however, that the resulting estimates are sensitive to the set of observations included as controls as well as to the fixed effects included in the specification. To address these concerns, we develop and implement a novel machine-learning approach to predict counterfactual energy consumption at treated schools and validate the approach with non-treated schools.

Specifically, we match hourly electricity consumption data from public K-12 schools in California to energy efficiency installation records, and exploit temporal and cross-sectional variation to

estimate the causal effect of the energy efficiency investments on energy use. Our data span 2008 to 2014, although only 20 percent of schools had smart meter data in 2008 and half have entered our data set by 2011. Our methodology includes traditional regression-based approaches, which allow us to non-parametrically control for a wide array of potential confounders. We implement a novel machine learning approach, using model selection and forecasting methods to construct school-specific counterfactual electricity usage using only pre-treatment data. We compare the results of this approach to the traditional difference-in-difference methods, and find, consistent with Monte Carlo simulations, that our machine learning approach yields slightly larger treatment effects.

Our contributions to the literature are both policy-relevant and methodological. From a policy perspective, this paper departs from much of the previous academic literature on energy efficiency by examining energy efficiency outside the residential sector. 37 percent of electricity use in the United States in 2014 was residential, and over half is attributable to commercial and industrial uses (EIA, 2015). A more complete view of what energy efficiency opportunities are cost-effective requires more evidence from a variety of settings, which, in turn, requires an informed understanding of the costs and benefits of investment in settings that have traditionally been difficult to study.

Our results demonstrate that energy efficiency investments can lead to substantial energy savings in schools. Across all types of investments, energy efficiency appears to deliver between 2 and 5 percent reductions in electricity use. We also look at the two most prevalent upgrade categories in our sample, lighting, which makes up 22 percent of upgrades; and heating, ventilation, and air conditioning (HVAC), which makes up 51 percent of upgrades. During daytime hours, investments in energy efficient lighting lead to 3 to 7 percent reductions in electricity use and smaller reductions at night. Investments in energy efficient HVAC systems produce a 2-4 percent reduction in the daytime hours, when temperatures are highest. These results translate into a significant amount of overall energy savings, representing about 60 KWh in daily electricity savings per school, which we estimate using both regression and machine learning techniques. When we compare these ex post effects to ex ante savings, however, we find that these ex-post saving estimates appear to deliver than 50 percent of expected ex-ante savings.

From a methodological perspective, high frequency data provides large benefits, but also presents new challenges. Using 15-minute interval electricity consumption data allows us to incorporate a rich set of controls and fixed effects in order to non-parametrically separate the causal effect of energy efficiency upgrades from other confounding factors. However, over-saturation is a concern; fixed effects estimators that absorb too much identifying variation can spuriously detect "treatment effects" that are simply artifacts of measurement problems in the data (Fisher et al., 2012). The machine learning method that we develop in this paper uses LASSO, a form of regularized regression, to generate a model of counterfactual electricity consumption that leverages our high-frequency data while avoiding overfitting. This allows us to optimally saturate the regression (trading off precision with bias). Furthermore, we only use pre-treatment data to train the LASSO model, meaning that the school-specific prediction models we generate do not risk absorbing part of the treatment, which is not present in the data used to build the model.

To our knowledge, this is the first paper in economics to incorporate machine learning methods into a selection-on-unobservables design in order to conduct causal inference.[1] We validate the

[1] In a recent NBER working paper, Cicala (2017) implements a variant on this methodology, using random forests rather than LASSO, in the context of electricity market integration.

machine learning predictions at our control schools, finding tightly-estimated zero effects for non-treated schools. We further embed our machine learning predictions into a difference-in-difference framework to account for the possibility of systematic trends in the prediction errors. Using a series of Monte Carlo simulations, we demonstrate that the machine learning approach performs as expected relatively to the regressions. We believe our methodology may be of interest to empiricists in a broad class of settings.

The remainder of the study proceeds by describing our empirical setting and datasets (Section 2). We then describe the baseline difference-in-differences methodology and estimate results using these standard tools (Section 3). Section 4 introduces our machine learning methodology, and presents the results. In Section 5, we compare our savings estimates to the ex ante engineering projections to calculate realization rates. Section 6 concludes.

2 Context and Data

Existing ex ante estimates suggest that commercial buildings, including schools, may present important opportunities to increase energy efficiency. For example, McKinsey & Company, who developed the iconic global abatement cost curve (see McKinsey & Company (2009)), note that buildings account for 18 percent of global emissions and as much as 30 percent in many developed countries. In turn, commercial buildings account for 32 percent of building emissions, with residential buildings making up the balance. Opportunities to improve commercial building efficiency primarily revolve around lighting, office equipment, and HVAC systems.

Commercial buildings such as schools, which are not operated by profit-maximizing agents, may be less likely to take advantage of cost-effective investments in energy efficiency, so they may yield particularly high returns to targeted programs that encourage energy efficiency investments. On the other hand, schools are open fewer hours than many commercial buildings, so the returns may be lower. Energy efficiency retrofits for schools gained prominence in California with Proposition 39, which voters passed in November 2012. The proposition closed a corporate tax loophole and devoted half of the revenues to reducing the amount public schools spend on energy, largely through energy-efficiency retrofits. Over the first three fiscal years of the program, the California legislature appropriated $1 billion to the program (CEC, 2017). To put this in perspective, it represents about one-third of what California is currently spending on *all* utility-funded energy efficiency programs (ranging from low-interest financing to light bulb subsidies to complex industrial programs) and about 5 percent of what utilities nationwide spend on energy efficiency (Barbose et al., 2013). Though our time period precedes most investments financed through Proposition 39, our results are relevant to expected energy savings from this large public program.

Methodologically, schools provide a convenient laboratory to isolate the impacts of energy efficiency. School buildings are all engaged in nearly the same activities, subject to relatively similar trends in education, and are clustered within distinct neighborhoods and towns. Other commercial buildings, by contrast, can house anything from an energy intensive data center that operates around the clock to a church that operates very few hours per week. Finally, given the public nature of schools, we are able to assemble relatively detailed data on building characteristics and recent investments.

Most of the existing empirical work on energy efficiency focuses on the residential sector. There

is little existing work on energy efficiency in commercial buildings. Kahn et al. (2014) provide descriptive evidence on differences in energy consumption across one utility's commercial buildings as a function of various observables, including incentives embedded in the occupants' leases, age and other physical attributes of the buildings. In other work, Kok and co-authors analyze the financial returns to energy efficiency attributes, though many of the attributes were part of the building's original construction and not part of deliberate retrofits, which are the focus of our work (Kok and Jennen, 2012; Eichholtz et al., 2013).

2.1 Data Sources

This project merges data from several sources. We combine high-frequency electricity consumption and account information with data on energy efficiency upgrades, school characteristics, community demographics, and weather.

We obtained 15-minute interval electricity metering data for the universe of public K-12 schools in Northern California served by Pacific Gas and Electric Company (PG&E). The data begin in January 2008, or the first month after the school's smart meter was installed, whichever comes later. 20 percent of the schools in the sample appear in 2008; the median year schools enter the sample is 2011. The data series runs through 2014. To speed computation time, we aggregate these 15-minute observations to three-hourly "blocks."[2]

In general, PG&E's databases link meters to customers for billing purposes. For schools, this creates a unique challenge: in general, school bills are paid at the district, rather than individual school site, level. In order to estimate the effect of energy efficiency investments on electricity consumption, we required a concordance between meters and school sites. We developed a meter matching process in parallel with PG&E. The final algorithm that was used to match meters to schools was implemented as follows: first, PG&E retrieved all meters associated with "education" customers.[3] Next, they used GPS coordinates attached to each meter to match meters from this universe to school sites, using school location data from the California Department of Education. This results in a good but imperfect match between meters and schools. In some cases, multiple school sites match to one or more meters. This can often be resolved by hand, and was wherever possible, but several "clusters" remain. We use only school-meter matches that did not need to be aggregated. Robustness tests suggest that the results presented here do not change substantively when we include these "clusters." Our final sample includes 2,094 schools.

The PG&E data also describe energy efficiency upgrades at the schools as long as the school applied for rebates from the utility.[4] A total of 6,971 upgrades occurred at 1,039 schools between January 2008 and December 2014. For each energy efficiency measure installed, our data include the measure code, the measure description[5], a technology family (e.g., "HVAC", "Lighting", "Food service technology"), the number of units installed, the installation date, the expected lifetime of

[2]Robustness checks suggest that our results are similar if we aggregate to hourly blocks.

[3]PG&E records a NAICS code for most customers in its system; this list of education customers was based on the customer NAICS code.

[4]Anecdotally, the upgrades in our database are likely to make up a large share of energy efficiency upgrades undertaken by schools. PG&E reports making concerted marketing efforts to reach out to schools to induce them to make these investments; schools often lack funds to devote to energy efficiency upgrades in the absence of such rebates.

[5]One example lighting measure description from our data: "PREMIUM T-8/T-5 28W ELEC BALLAST REPLACE T12 40W MAGN BALLAST-4 FT 2 LAMP"

the project, the engineering-estimate-based expected annual kWh savings, the incremental measure cost, and the PG&E upgrade incentive received by the school.[6]

Many schools undertake multiple upgrades, either within or across categories. We include all upgrades in our analysis, and break out results for the two most common upgrade categories: HVAC and lighting. Together these two categories make up over 70 percent of the total upgrades, and over 60 percent of the total projected savings.

We obtained school and school-by-year information from the California Department of Education on academic performance, number of students, the demographic composition of each school's students, the type of school (i.e., elementary, middle school, high school or other) and location. We also matched schools and school districts to Census blocks in order to incorporate additional neighborhood demographic information, such as racial composition and income. Finally, we obtained information on whether school district voters had approved facilities bonds in the two to five years before retrofits began at treated schools.[7]

We obtained hourly temperature data from 2008 to 2014 from over 4,500 weather stations across California from MesoWest, a weather data aggregation project hosted by the University of Utah.[8] We matched school GPS coordinates provided by the Department of Education with weather station locations from MesoWest to pair each school with its closest weather station to create a school-specific hourly temperature record.

2.2 Summary Statistics

Table 1 displays summary statistics for the data described above, across schools with and without energy efficiency projects. Of the 2,094 schools in the sample, 1,039 received some type of energy efficiency upgrade. 628 received only HVAC upgrades and 493 received only lighting upgrades. There are 1,055 schools that received no upgrade. Our main variable of interest is electricity consumption, which we observe every 15 minutes but summarize in Table 1 at the 3-hourly "block" level for practical purposes. We observe electricity consumption data for the average school for a three-year period.

For schools that are treated, expected energy savings are almost 30,000 kWh, which is approximately 5 percent of average electricity consumption. Savings are a slightly larger share of consumption for schools with lighting interventions.[9]

[Table 1 and Figure 1 about here]

The first three columns of Table 1 highlight measurable differences between treated and non-treated schools. Treated schools use over 50 percent more electricity, have 30 percent more enrolled students, different student demographics and are generally further south and further east. Figure 1 shows the spatial distribution of treatment and control schools. Schools that received HVAC and/or lighting upgrades look different across an array of observable characteristics from schools that did

[6] We have opted not to use the cost data as we were unable to obtain a consistent definition of the variables.

[7] Bond data are from EdSource (edsource.org).

[8] We performed our own sample cleaning procedure on the data from these stations, dropping observations with unreasonably large fluctuations in temperature, and dropping stations with more than 10% missing or bad observations.

[9] We do not summarize projected savings since it is zero for the control schools.

not receive these upgrades. Schools receiving lighting upgrades perform less well academically than non-upgrading schools, but this difference disappears when comparing schools that did and did not receive HVAC upgrades. Because these schools are different on a range of observable characteristics, and because these indicators may be correlated with electricity usage, it is important that we consider selection into treatment as a possible threat to econometric identification in this setting.

3 Regression Analysis

We first present regression results that estimate the treatment effect of the energy efficiency interventions using a difference-in-difference framework. Specifically, we compare schools with and without energy efficiency interventions, using data that spans the period before and after schools invested.

3.1 Difference-in-Difference Approach

In order for a difference-in-difference analysis to be valid, the identifying assumption that treated and untreated schools were trending similarly prior to the treatment must hold. While this assumption is fundamentally untestable, we provide evidence in support of the assumption by using an event study framework. In this event study, we compare treated and untreated schools in event time (i.e., in the years prior to and after treatment), and demonstrate that these groups of schools were not trending differently before the treated schools installed energy efficiency upgrades. Because we observe upgrades, including HVAC investments, that may be expected to deliver savings during some parts of the year and not others, we estimate our event studies at the annual level. Figure 2 displays the results of this event study using a parsimonious specification, including only school and block fixed effects.[10]

[Figure 2 about here]

In the years prior to treatment, we are unable to statistically distinguish the difference between treated group and untreated group outcomes from zero. However, after treatment, we observe a statistically significant reduction in electricity consumption. This suggests that the difference-in-difference design is likely to be valid in this context.

Moving into the full difference-in-difference design, as our base specifications, we estimate regressions similar to the following:

$$Y_{ith} = \beta_h D_{it} + \alpha_i + \kappa_h + \gamma_t + \varepsilon_{ith}, \tag{3.1}$$

where Y_{ith} represents the log of electricity consumption in kWh at school i on date t during hour-block h. D_{it} is a treatment variable equal to the cumulative fraction of upgrades installed in

[10] To generate the coefficients displayed in Figure 2, we randomly assign a treatment date to untreated schools, and regress the log of energy consumption on a set of years-to-treatment dummies (excluding the year before treatment), a set of years-to-treatment × treated-school dummies, school fixed effects, and block fixed effects. Figure 2 reports the coefficients on the years-to-treat × treated-school dummies.

school i by date t, as measured by ex ante expected kWh savings.[11] Therefore, this coefficient can be interpreted as the effect of going from no upgrades to 100% of a school's upgrades. This parameterization allows us to estimate treatment effects in a setting where schools undergo multiple energy efficiency upgrades. We allow the coefficient on the treatment, β_h, to be different for each block, as we expect upgrades to have heterogeneous effects on energy consumption throughout the day. We also include school fixed effects, α_i, to absorb school-specific average energy consumption, for instance related to its physical characteristics or average enrollment. We include hour-block fixed effects, κ_h, to absorb variation over the daily cycle in energy use, and γ_t, date or month-of-sample fixed effects, to capture common shocks across the sample. ε_{ith} is an idiosyncratic error term. We cluster our standard errors at the school level, allowing for arbitrary dependence between any two observations within the same school.

We present results for several specifications, each with a different combination of fixed effects, which range from fairly parsimonious specifications, including only school and block fixed effects, to much more flexible ones, allowing for, for instance, school-specific hour-block by month-of-year effects. Under the assumption that, conditional on fixed effects, treatment is as good as randomly assigned, or, that conditional on fixed effects, there are no remaining time-varying differences between treated and untreated schools, we can use this approach to identify the causal effects of energy efficiency upgrades on electricity consumption.

3.2 Difference-in-Difference Results

Table 2 reports the results from a series of difference-in-difference specifications. As we move across columns, we progressively include richer fixed effects and controls. The top panel reports estimates of the aggregate treatment effect of the retrofits across all hours. Below it, we report differential treatment effects for each three-hour block.

[Table 2 about here]

Looking at the aggregate effects, we find that the average project implemented at the schools in our sample delivered a reduction in electricity consumption between 2 and 5%. The estimated effect is sensitive to the presence of month-of-sample fixed effects and controls, which are included in columns (4) and (5), respectively. This could reflect common trends across schools. Examining the block patterns, we find that the largest reductions accrue during the hours of school operation, which seems intuitive. Note that consumption of electricity (in levels) is larger in those hours (often twice as high), so the difference in electricity savings in levels across hours is substantial.[12]

Table 3 presents the results for two subsample: treated schools that installed HVAC upgrades and treated schools that installed lighting upgrades (both compared to untreated schools that installed no upgrades over our sample period). The results for HVAC and lighting differ substantially. We estimate reductions of between 2 and 5 % for HVAC interventions, but the HVAC estimates are relatively noisy. Lighting interventions appear to drive larger reductions in electricity consumption:

[11] Specifically, an untreated school will have $D_{it} = 0$ in all periods. At a treated school that undergoes two upgrades with projected savings of 10 kWh each, D_{it} will be zero prior to the first upgrade; after the first upgrade, D_{it} will be 0.5; and after the second upgrade, D_{it} will be 1.

[12] Boomhower and Davis (2016) measure the benefits of efficiency investments by time of day and show that reductions in the middle of the day are worth significantly more in California.

between 4 and 7 %. Furthermore, the HVAC effects appear to be relatively consistent throughout the day, while the lighting effects are stronger during the main school hours.

[Table 3 about here]

One potential explanation for this difference is that the effects of lighting interventions are relatively homogeneous across schools, while the effects of HVAC interventions are much more sensitive to the local climate, current weather conditions, and other factors. Therefore, it might be harder to control for all confounding factors in a parsimonious way. For example, the inclusion of school-specific month effects appears to affect some of the estimates for the HVAC specifications (e.g., columns (3) and (5)), whereas they are much less important for the lighting interventions.[13]

3.3 Robustness

Despite the promising event study shown above, several features of our data suggest that the treatment effects estimated using the difference-in-difference approach may be biased. For one, the coefficient estimates appear sensitive to the sets of fixed effects we include. Also, the number and composition of schools in the sample (and the treatment group) changes over time, which means it is important to flexibly control for time-varying, confounding factors. Since electricity consumption follows seasonal, weekly and diurnal patterns, controlling for all potential changes can be difficult.

Here, we present two sets of robustness checks on the difference-in-difference approach. First, we show a placebo exercise, in which we randomly assign "treatments" to schools in the period prior to actual treatment, to test the robustness of our fixed effects approaches. We observe patterns in these placebo estimates that suggest that our specifications do not adequately control for time-varying unobservable characteristics that may bias our difference-in-difference results. We next present results from a series of nearest-neighbor matching estimators, in an effort to avoid issues with selection bias that may arise given the differences between treated and untreated schools in our sample. We find that these results are highly unstable, and sensitive to different matching specifications.

3.3.1 Placebo Tests

We first conduct a series of placebo tests using the same specifications summarized in Tables 2 and 3 to gauge the extent to which our difference-in-difference approach is appropriately controlling for time-varying unobservable characteristics which may threaten our identification strategy. To do this, we drop all post-treatment observations and randomly assign approximately 50% of schools into a placebo "treated" group, to match the proportion of schools actually in treatment in the real sample. We then randomly assign these "treated" schools a "treatment date" by taking a uniform date draw between their first appearance in the sample and their last appearance in the sample.[14] We then estimate specifications (1) through (5) and store the block-wise coefficients. We repeat this process 25 times.

Figure 3 reports the results of this exercise. Panels 1 to 5 match the specifications in columns (1) to (5) of Tables 2 and 3. We plot treatment effects for each placebo run in gray, and overlay

[13] In ongoing work, we are exploring heterogeneity in more detail.

[14] We do not allow schools to be "treated" in either their first or last month in the sample.

our estimated treatment effects for any upgrades, HVAC upgrades and lighting upgrades in shades of blue. We also plot the average coefficient across all placebo estimates (solid thick gray line). Notably, the placebo coefficient estimates display a systematic pattern by block across all of the simulations. Specifically, estimates appear negative in the early part of the day and positive after noon. These results suggest changes in the temporal pattern of consumption over time that are not captured by the rich fixed effects we include. This calls into question the validity of these difference-in-difference specifications.

3.3.2 Matching

Given the systematic differences between treated and untreated schools summarized in Table 1, we also explore the sensitivity of the coefficient estimates to the set of untreated schools included in the estimation sample. In particular, we implement nearest neighbor matching based on several candidate choice sets, with the goal of choosing matches that are most closely aligned on observable characteristics, in order to reduce selection bias.

Matching presents a challenge in the school setting. An unrestricted nearest neighbor match will tend to select untreated schools from the same district as the treated schools, since schools within the same district tend to be demographically similar and have similar weather and electricity consumption. However, school infrastructure decisions are often made at the *district* level, meaning that unrestricted matches and matches that are restricted to be within the same school district may be problematic. In particular, if a district selects one of its schools to receive an energy efficiency upgrade and not others for reasons that are unobservable to the econometrician, this type of nearest-neighbor matching strategy may induce selection bias. On the other hand, restricting matches to schools that are in different districts makes it more difficult to find untreated schools that are comparable on observable characteristics.

We present results from three types of matches: unrestricted, restricted to schools in the same district, and restricted to schools in other districts. Our treatment effect estimates exhibit strong sensitivity to the matching criteria that we choose. Table 4 presents aggregate treatment effect estimates, analogous to the top row of Table 2. In each panel, matches are drawn from an unrestricted set in the first row, the set of schools in the same district in the second row and the set of schools in outside districts in the third row.[15] Examining the top panel ("Any interventions"), it is clear that estimates are sensitive to modeling choices. Matching on "Any district" yields larger estimates than when candidate matches are restricted, although the differences are greater when matches are drawn from the same district. Looking across panels, which isolate HVAC and lighting interventions, a similar narrative holds. The sensitivity to matching criteria and specification reflects a tension between quality of candidate matches and selection bias on imperfect matches. These results suggest that it is difficult to use matching methods to construct an appropriate counterfactual in this setting.

[15] The matching variables are summarized in the table notes and include both demographic variables and electricity usage patterns. Results using other sets of matching variables are similarly sensitive to the specification and match issues summarized in Table 4.

4 Machine Learning Analysis

Our machine learning approach deploys forecasting methods to predict electricity consumption at each school and uses this prediction as a counterfactual to estimate energy savings. Broadly, these methods use algorithms to construct prediction models that are designed to perform well out of sample. The researcher provides the machine learning tool with a dataset and the candidate variables for model selection. The machine learning algorithm will then subset the data into "training" and "test" subsamples, fit models using the training data only, and test the out-of-sample fit of these models using the test data.[16] These methods are designed to trade off bias and variance, and penalize overfitting, such that the resulting model provides the best out-of-sample fit. There are a variety of machine learning algorithms that can used to do this, including LASSO, trees, random trees, a combination of models (bagging), and others. Ultimately, these methods return a forecasting model, which we use to generate a counterfactual data series.

4.1 Machine Learning Approach

The goal of the predictive model in our setting is to provide a counterfactual during the treated period that describes what would have happened to energy consumption at treated schools if they had not implemented an energy efficiency project. Machine learning tools are particularly well-suited to constructing counterfactuals, since the goal of building the counterfactual is not to individually isolate the effect of any particular variable, but rather to provide a good overall prediction fit. With the goal of achieving best predictive power, machine learning techniques give substantial flexibility to the algorithms to build a statistical model that considers many potential regressors. Machine learning models tend to out-perform models that are chosen by the researcher in a more idiosyncratic fashion when it comes to predictive power, since they use an algorithm and penalty rule that trades off fit and variance. This enables the econometrician to select models from a much wider space than would be possible with trial-and-error. Importantly, in our empirical setting, we generate a prediction model separately for each school This allows for a great deal of flexibility in the control variables used to create our counterfactual and makes it feasible to retrieve estimates of heterogeneous treatment effects.

Relative to the difference-in-differences estimators, a key difference in our machine learning approach is that we only use pre-treatment data to generate the counterfactual model. By providing only pre-treatment data to the algorithm, we generates a predictive model that is a function of co-variates and, by construction, is uncontaminated by the treatment effect itself.[17]

4.1.1 Methodological Contribution

Until now, the intersection of machine learning and causal inference has focused on two applications: randomized controlled trials and selection-on-observables designs. Our approach combines machine learning with selection on unobservables. We leverage high-frequency data to construct unit-specific counterfactuals in a panel fixed-effects framework.

[16] Often, this subsampling procedure is repeated several times to improve the performance of predictions.

[17] Note that the pre-treatment data is further subsetted by the algorithm into a "training" and "test" sample, as described above. None of the post-treatment data is used to "test" the predictive model. The method is described in detail below.

The first strand of the existing literature to combine machine learning and causal inference focuses on randomized trials. Under random (or as good as random) assignment, there is little need for machine learning to identify average treatment effects. In many randomized settings, however, researchers are often also interested in retrieving heterogeneous effects while minimizing concerns about "cherry-picking". Athey and Imbens (2015) propose a recursive partitioning methodology which uses random forests to estimate heterogeneous treatment effects with proper inference. While these methods are useful when units are randomly assigned to treatment, the non-random selection present in our setting makes this method undesirable, as the partitioning itself may introduce greater selection bias.

Machine learning methods have also been developed for settings without random assignment into treatment. The existing work in this area focuses on identifying average treatment effects in selection-on-observables designs, where, conditional on observable characteristics, the identifying assumption is that the outcome of interest would be equal across groups in the absence of treatment (i.e., no selection on unobservables). Three broad classes of methods have arisen to use machine learning to improve selection on observables methods. The first approach is akin to existing propensity score methods (cross-sectional units are assigned PS weights according to their ability to predict treatment status), but uses machine learning tools for model selection (McCaffrey et al. (2004)). A closely-related method forces covariate "balance" by directly including a covariate balancing constraints in the machine learning algorithm (e.g., Wyss et al. (2014)). Finally, Belloni et al. (2014) proposes a "double selection" approach, using machine learning in the form of LASSO, to both predict selection into treatment as well as to predict an outcome, using both the covariates that predict treatment assignment and the outcome in the final step.

These double selection methods improve upon existing selection-on-observables machine learning designs, but all three methods are subject to the same identifying assumptions as traditional selection-on-observables methods. Furthermore, these methods perform best with large N, and when the degree of heterogeneity is small relative to the sample size (i.e., it is possible to find control units that are suitable matches on observable characteristics). In our setting, it is challenging to find good matches across the treated and untreated groups, as we observe a limited number of units and the nature of selection in our setting makes matching difficult. In particular, it seems attractive to match treated schools to untreated schools in the same district, but it is likely that selection into treatment occurs within district. Alternatively, restricting our matches to occur across districts means that our matches will often remain relatively different on observables.

Though our setting features non-random assignment to treatment and small N, we have large T: our electricity consumption data consists of tens- or hundreds-of-thousands of observations per cross-sectional unit. As a result, we propose a new method for combining machine learning and causal inference. Rather than using machine learning to predict selection into treatment, we use untreated time periods to create unit-specific predictions of the outcome in the absence of treatment. For each school, we generate a LASSO-based counterfactual electricity consumption time series for the treated periods. We can then use these data in conjunction with a within estimator to recover the causal effect of treatment. While our method has some particular data requirements that limit its suitability to settings with large-T samples, it also offers many benefits. There is an intuitive appeal to using untreated observations within a subsequently treated unit as a counterfactual. Machine learning allows the researcher to take an agnostic approach to model selection, where

outcomes of different cross-sectional units may be optimally predicted using different variables. Furthermore, the machine learning approach protects against oversaturation. Our method also provides benefits when some features of the data-generating process are exposed to measurement error, as we will demonstrate.

Our method proceeds in two steps. In a first step, we build a statistical model that explains electricity consumption at a given school, as a function of a set of observable variables, only using pre-treatment data. We repeat this step for each school separately, generating an individual prediction model tailored to match past consumption patterns. In a second step, we compare actual energy consumption to the prediction from the model during the treatment period. Exploiting the same structure as in the difference-in-difference specifications, we take advantage of untreated schools by comparing prediction errors both for treated and untreated schools.

4.1.2 Step 1

We begin by forecasting electricity consumption for each school using machine learning. The machine learning algorithms pick and/or create explanatory variables and generate a prediction model based on a LASSO regression.

A key feature of our approach is that we only use data from the pre-period to "train" and "test" the model. For treated schools, the pre-period is defined as the period before any intervention occurs. For untreated schools, we select a subset of the data to be the pre-treatment period by randomly assigning a treatment date between the 20th percentile and 80th percentile calendar dates available.[18] Data after the treatment or pseudo-treatment date is set aside and not used for the purposes of estimating the model.[19] Once the pre-treatment period is defined, we use a cross-validation method for each school separately to algorithmically select the best predictive model.

The resulting output of this first step is a school-specific prediction model for electricity consumption that we apply to each observation in our sample. Using the coefficients from the prediction model in this first step, and observations on the covariates during our entire sample, we predict energy consumption for the whole sample period for each school.

4.1.3 Step 2

Armed with this prediction, we can apply a variety of estimators. The simplest treatment effect estimate is to take the difference between actual electricity consumption at treatment schools and our prediction from the machine learning model resulting from the first step in the post-treatment period. The resulting estimator of the average effect across treated schools becomes

$$\hat{\beta}^T = \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}],$$

[18] We set the threshold to be between the 20th and 80th percentile to have a more balanced number of observations in the pre- and post-sample.

[19] Imagine an untreated school that we observe between 2009 and 2013. We randomly select a cutoff date for this particular school, e.g., March 3, 2011, and only use data prior to this cutoff date when generating our prediction model.

where $Y^T_{i,post}$ represents actual average consumption of electricity during the treatment period, and $\hat{Y}^T_{i,post}$ is the predicted electricity consumption during the same treatment period, based on the predictive model built using pre-treatment data only. If an energy efficiency project achieved savings, we would expect $\hat{\beta}^T$ to be negative, as the prediction $\hat{Y}^T_{i,post}$ would overstate expected electricity consumption.

To account for any systematic biases in the prediction, we also use pre-treatment data in a differenced framework, and obtain,

$$\hat{\beta}^{TD} = \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^T_{i,pre} - \hat{Y}^T_{i,pre}],$$

where $Y^T_{i,pre}$ and $\hat{Y}^T_{i,pre}$ are actual and predicted log electricity consumption during the pre-period, respectively. In the context of prediction models, this correction will tend to be very minor on average, as by construction the average prediction in the pre-treatment period should be close to average actual consumption. For example, these two averages cancel each other by construction if the prediction is based on OLS.[20]

These estimators do not leverage the presence of untreated schools. As with any case study, a before-and-after comparison can suffer from substantial pitfalls. Whereas a richer predictive model can help control for nonlinear effects of observable variables, there is still the concern that there might be trends and other shocks that are not properly captured by the model.

To partially address these concerns, we can use untreated schools as in the previous section, as long as trends are common across schools. We can check the performance of the prediction during the treatment period by examining prediction errors at untreated schools, i.e.,

$$\hat{\beta}^U = \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}].$$

If the model has good performance out of sample, $\hat{\beta}^U_i$ should be zero in expectation. In practice, however, the predictive model might fail at capturing behavioral changes over time or other trends in unobservables, leading to some systematic differences that would imply $\hat{\beta}^U \neq 0$.

We can also test the model's performance in a differenced approach similar to estimating $\hat{\beta}^{TD}$, where we examine the differences between real and predicted consumption in the post and pre period for untreated schools only, and obtain,

$$\hat{\beta}^{UD} = \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] - \mathbb{E}[Y^U_{i,pre} - \hat{Y}^U_{i,pre}],$$

Again, $\hat{\beta}^{UD}$ should be zero in expectation if the model is successful.

To the extent that there are systematic differences between the prediction and the observed outcomes for untreated schools, and to the extent that these differences reflect trends and biases in the predictive model that are common across schools, we can use these differences as a bias

[20] When using more sophisticated prediction methods that trade-off overfitting, or when allowing for heterogeneous treatment effects within a school (e.g., across hours, for different temperature ranges, etc.), the average prediction might not exactly match average actual consumption at a given cell, making this correction potentially more relevant.

correction for the treated schools. We have,

$$\begin{aligned}\hat{\beta}^{DD} &= \mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] \\ &= \hat{\beta}^T - \hat{\beta}^U,\end{aligned}$$

which will provide a corrected treatment effect estimate under the assumption of common trends and shocks between treated and untreated schools. As before, we can also rely on a triple difference that exploits the differences in predictions between treated and untreated schools during the pre- and post-period,

$$\begin{aligned}\hat{\beta}^{3D} &= \Big(\mathbb{E}[Y^T_{i,post} - \hat{Y}^T_{i,post}] - \mathbb{E}[Y^T_{i,pre} - \hat{Y}^T_{i,pre}]\Big) \\ &\quad - \Big(\mathbb{E}[Y^U_{i,post} - \hat{Y}^U_{i,post}] - \mathbb{E}[(Y^U_{i,pre} - \hat{Y}^U_{i,pre}]\Big) \\ &= \hat{\beta}^{TD} - \hat{\beta}^{UD}.\end{aligned}$$

As mentioned above, this triple difference will tend to provide very similar results to those only using post data when looking at average effects.

Additionally, we consider embedding our machine learning predictions into a regression framework to account for potential additional confounding factors, such as the composition of schools over time. In particular, we consider a regression in which we regress the differences between actual energy consumption and the prediction on the treatment dummy, school-hour fixed effects, month-of-sample fixed effects, and other controls. The approach parallels a standard difference-in-differences methodology, with the major difference that the dependent variable is the error from the prediction model, instead of electricity consumption itself. The only difference with the regression approach is that we include an additional control, "post-period", which is a dummy to account for the fact that there is a break between the predictions in- and out-of-sample. Including this control is an additional way to take into account that predictions are not as accurate during the post-period, which could introduce an artificial structural break in the data.

4.2 Methods Comparison: Monte Carlo

The regression and machine learning methodologies use different sources of variation to generate their respective counterfactuals, and as a result they may retrieve different estimates of the average treatment effect (ATE). Without ex ante knowledge of the true ATE and the true underlying data generating process (DGP), it is not possible to evaluate the relative performance of the estimators as they are applied to our observational data. Monte Carlo simulations, on the other hand, provide a way to control the DGP and thereby compare the relative merits of various approaches against a known benchmark. In this subsection, we describe a series of Monte Carlo simulations, which allow us to understand how well the regression and machine learning approaches perform when applied to settings with known potential confounders.

The baseline DGP is chosen to reflect the main features of our observational data on school energy use. Cross-sectional units (call them "schools") arrive in the dataset at different dates and are observed at high frequency for different durations. The outcome variable of interest at these schools ("energy use") varies according to school size and temperature, which in turn varies by

geographic location (inland versus coastal), season, and hour-of-day. At some schools, a treatment of known magnitude and timing is superimposed on this DGP; the remaining schools are untreated.

There are many potential confounders that one may want to consider. The ones that we highlight in this Monte Carlo exercise are motivated by actual uncertainty that we encounter when attempting to correctly specify our regression approaches in the context of school energy efficiency investments: 1) the treatment effect may be nonlinear with respect to temperature; and, 2) the treatment date may be observed with error.

There are several intuitive reasons why the machine learning approach described above can offer benefits relative to a regression approach that seeks to retrieve the ATE. In order to recover a consistent estimate of the ATE, the regression model must be correctly specified. In a setting with a complex DGP, properly selecting controls can be extremely difficult, and there are often not clear rules governing variable selection. In contrast, machine learning uses an algorithm, including cross-validation, to select control variables, imposing discipline on variable selection. Furthermore, because we can implement school-specific machine learning predictions, we are able to much more flexibly specify the counterfactual-generating model than would be computationally feasible in a regression context. Regression models are also subject to over-fitting. While omitting variables from a regression may lead to bias, including variables that are correlated with treatment can also introduce bias, causing controls to soak up some of the treatment effect. Because our machine learning predictions are constructed using pre-treatment data only, the counterfactual model will not suffer from these problems.

Such overfitting would not occur with the machine learning approach. The prediction model is tested and trained on pre-treatment data only, and will capture the true relationship between Y_{ith} and covariates in the absence of the intervention. This allows the researcher to retrieve the correct ATE by comparing the prediction to the true data. There is an important potential benefit, therefore, from using only pre-intervention data to estimate control parameters. This is potentially feasible in a regression approach as well, but it is rarely the way regressions are implemented. In addition, measurement error in the timing of treatment may cause bias in a regression model, as mistakenly attributing post-treatment data to the pre-treatment period and vice versa will attenuate estimated treatment effects. Machine learning will be subject to less bias than regression if the treatment date is perceived to be earlier than it actually is, since the counterfactual will remain uncontaminated. In the case where the treatment date is thought to be later than it actually is, we may have concerns that the testing and training dataset used to build the model includes treated observations, but in cross-validating the model, these observations should be given less weight, as they will appear to be outliers.

Ultimately, the machine learning algorithm results in a parsimonious prediction model that allows the researcher to remain more agnostic about variable selection, prevents model overfitting, and reduces concerns about measurement error. Our Monte Carlo results provide evidence that there are real benefits to implementing such an approach.

Table 5 presents summary statistics from the Monte Carlo simulations. We test three DGPs, each of which includes a treatment effect that is nonlinearly increasing in temperature. The first DGP ("No Measurement Error") assumes that the effective treatment date used by the researcher is accurate. Two alternate DGPs assume that the researcher is using a mismeasured treatment date that is either before ("early") or after ("late") the true date. We attempt to retrieve the true

treatment effect using the regression and machine learning approaches under four different levels of controls. Specification 0 is reported only for the machine learning method, and it controls for nothing in the second stage. Specifications 1-3 deploy controls that sequentially increase saturation.[21] The table reports the percentage distance between the estimated ATE and the truth, and the standard error.

[Table 5 about here]

Several qualitative conclusions emerge. The regression estimates are systematically attenuated (i.e., the statistics in Table 5 are negative), and appear to be sensitive to the specification. When there is no measurement error in the treatment date, the fully-specified regression retrieves the true ATE and is precise. This is consistent with the efficiency of OLS (best linear unbiased estimator). Regression also performs well when slightly underspecified (Specification 2), although this feature may not be generalizable since correlations between observed and unobserved variables may be stronger in the real world. Machine learning without controls (Specification 0) performs significantly better than regression under Specification 1, when the regression model is poorly specified. Including additional controls with the machine learning model seems to have little effect on its overall performance under this scenario. This suggests, as expected, that machine learning may have substantial benefits relative to a misspecified regression.

Measurement error in the treatment date introduces significant attenuation bias into the regression estimates. While machine learning does not solve these problems entirely, it appears to yield results with superior properties relative to the regression estimates. First, machine learning estimates are less sensitive to the presence of controls in the second stage. This is a direct result of the prediction itself having been generated using LASSO-selected variables at the school level, which already embody the most important sources of variation. This can be seen most clearly in Specification 0, which performs remarkably well in the absence of any second-stage control variables.

Second, while regression estimates are attenuated across the board, machine learning estimates are far less so. This is particularly true in the presence of measurement error. When the treatment date is mismeasured "early", the counterfactual will be uncontaminated but the "treatment period" will include untreated periods, leading to attenuation. One might be concerned when mismeasurement of the treatment date is late, since this may expose the training/testing procedure to some contamination of the treatment effect. However, the LASSO procedure may perceive these as outliers. Recall that it penalizes the inclusion of additional variables, and thus trades off parsimony against goodness of fit. The large reduction in bias that we observe from machine learning is encouraging, and aligns with our intuition about the channels of potential bias.

There is some question as to the necessity of implementing the second stage at all. A clear tradeoff exists between bias and precision, which can be seen most clearly in "Wrong treatment date (Late)". In that case, the second stage biases the estimates, presumably due to measurement error; but there are clear gains in precision. The extent to which this trade-off favors implementing the second stage or not will depend on the true DGP, and it is impossible to draw a general conclusion from the single DGP that we simulate here.

[21] Specification 1 includes school fixed effects. Specification 2 adds month of year and year fixed effects. Finally, specification 3 include month-of-sample fixed effects, school interacted with time-of-day fixed effects, and school-specific temperature controls.

In summary, machine learning appears to offer some meaningful advantages over the regression approach. While we have tested it on a limited class of potential confounding factors, our intuition is that some of the benefits will extend to a broad class of DGPs. Parsimonious variable selection, robustness to second-stage (researcher) control variable selection, and training/testing on pre-treatment period data all have appealing features in a general sense.

4.3 Machine Learning Results

4.3.1 Step 1

We use a LASSO estimator with pre-treatment data to estimate each school-specific electricity consumption model. We are flexible in the allowed regressors that are considered by the estimator, including: block fixed effects, day of the week, a holiday dummy, a seasonal spline, a temperature spline, and several multiple combinations of these variables (e.g., day of the week interacted with holiday, the seasonal spline interacted with temperature and the blocks, etc). In total, this generates over 3,000 candidate variables.[22]

The machine learning estimator tries to balance the number of explanatory variables and the prediction errors of the model, using cross-validation. Naturally, given that we are estimating these coefficients for each school separately, the optimal models do not include all regressors. In fact, we find that the optimal models usually include between 50 and 100 variables. Importantly, however, the variables are not the same for each school. In fact, we find that the joint set of variables across all schools covers all of the more than 3,000 candidate variables that we consider.

The richness of the selected model depends substantially on the available data. Figure 4 shows the relationship between the amount of data available to "train" the model and the number of non-zero coefficients in the prediction model. One can see that when data availability is limited, the LASSO coefficient will try to avoid overfitting and limit the number of coefficients that produce the predictive model. One can also see that, even with larger data sets, at some point the inclusion of additional variables is limited.

[Figure 4 about here]

The variables related to the holiday dummy provide a good example to understand the power of the machine learning approach applied to each school *individually*.[23] Holidays dramatically affect electricity consumption in general, and certainly at K-12 schools. Indeed, almost all the school-specific models pick up at least one variable containing the holiday dummy, potentially interacted with other regressors (e.g., interacted with the time of day or the day of week). We allow for over fifty such interactions. The median model picks up eight of them, but some school models pick up over twenty holiday variables, depending on data availability and the importance of holidays to that school. The LASSO methodology tries to avoid including too many variables that could lead to collinearity issues and poor performance out of sample. The coefficient estimates on those

[22] Note that if we were to run all schools together, we would have over 6,000,000 candidate variables, as each variable is allowed to have a school-specific coefficient!

[23] We define holiday to include major national holidays, as well as the Thanksgiving and Winter break common to most schools. Unfortunately, we do not have school-level data for the summer break, although the seasonal splines should help account for long spells of inactivity at the schools.

variables are mostly negative, and we do not observe many large outliers, as seen in Figure 5.

[Figure 5 about here]

4.3.2 Step 2

Table 6 presents the baseline estimates comparing prediction errors across schools. In column (1), one can see that the average prediction error for untreated schools during the post-treatment period (i.e., out of sample) is zero. The algorithm appears to predict average consumption well on average, although the prediction errors are not zero at the block level, suggesting that the prediction model performs less well when predicting particular times of the day. This is because the prediction methodology was built to minimize average errors, but was not stratified to also balance errors across hours.[24]

Column (2) also reports average prediction errors, but in this case for treated schools. One can see that the differences for this set of schools are more systematic, as implied by the average effect. At the block level, all of the estimates are negative, implying energy consumption reductions during the treatment period for most blocks of hours. Overall, one can see that the prediction errors for treated and untreated schools are systematically different, with prediction errors for untreated schools precisely estimated around zero, and prediction errors for treated schools significantly negative, i.e., implying energy savings.

Column (3) presents the difference between prediction errors for treated and untreated schools. To the extent that both predictive models have common biases, $\hat{\beta}^{DD}$ should provide a better estimate of the treatment effect. One can see that the estimated effects between column (2) and (3) are not very different on average, although the effects by block appear to be more balanced after performing the correction. Columns (4) and (5) exploit the pre-treatment data to correct for potential prediction biases already present in the baseline. The results are very similar to column (3), suggesting that the estimated treatment effect is not an artifact of the predictive model performing poorly for treated schools in-sample.

[Table 6 about here]

The previous results present differences-in-differences across treated and untreated schools, without controlling for any other factors such as school fixed-effects or seasonality and time trends. However, the prediction model has limitations, and it is thus important to further control for confounding factors. Table 7 presents post-period average effects including additional controls, which parallels Table 2 in the difference-in-differences results. In column (1), one can see that the effects for treated schools remain negative and significant, albeit smaller, as we include school fixed effects that take away some persistent heterogeneity. We obtain similar results if we also include school-specific block effects, as seen in column (2). Column (3) includes school-specific block times month effects, to control for seasonality. One can see that the results are not very different. As in the regression section, to control for trends, we include month-of-sample fixed effects in column (4),

[24] We also consider a predictive model for each school-block, to correct some of these biases. We find that the main findings in this section remain, with most of the biases at the block level that we observe being corrected even without controls.

and a month-of sample time trend in column (5). We find that controlling for seasonality, trends, and changes in the composition of the panel is important and reduces the treatment effect, but the main patterns remain.

[Table 7 about here]

As shown in the regression results, our data include both HVAC and lighting interventions. Using the alternative machine learning approach, we also explore whether the treatment effects across interventions appear to be different. Table 8 presents the results for schools with only HVAC interventions and those with only lighting interventions. We find results that are broadly consistent with the findings in Table 3, although substantially larger in some specifications.

For HVAC, we find that there is no effect at night, and most of the effects are highest during sunlight hours, between 9am and 3pm, and are statistically significant between 3pm and 6pm. We also find some positive (not significant) results in the early hours of the morning, which can be explained by potential ramping up of new HVAC systems, since these ramps tend to be better timed in newer systems. The results appear to be less sensitive across specifications than in the case of regression. For lighting, we find that there is a treatment effect both during the day and at night, although the effect is largest during the day and only significant during school hours, suggesting that the interventions appear to be most useful during active hours of operation. The results appear to be stable and large, in the order of 3-7% during the day.

[Table 8 about here]

4.4 Methods Comparison: Results

The results from the machine learning approach are broadly consistent with our findings using the difference-in-differences approach. However, there are some differences. One major difference is that the traditional regressions are more sensitive to the inclusion of fixed effects and controls. An extreme version of this difference is the fact that regression results without including school fixed effects are severely biased, whereas the prediction model has already removed substantial differences across schools, by being a de-facto triple-difference estimator. This means that fixed effects in the prediction model do not need to absorb as many remaining differences across schools. One can also see that estimates from the machine learning approach appear to be larger, which is consistent with our initial hypothesis and Monte Carlo results, and could be due to the presence of measurement error and/or specification bias.

One way to see the power of the machine learning methodology is by running a series of placebo simulations akin to the ones we implemented for the difference-in-differences methodology. As above, we drop all post-treatment observations, thereby eliminating usage data that are exposed to treatment, and randomly assign approximately 40 percent of schools to be "treated". For the prediction placebos, and for computational reasons, we keep the same random treatment date that we used to generate the prediction.

Figure 6 presents a series of placebo tests with 25 placebo realizations for the prediction regressions. As was the case for the difference-in-differences placebos, the average placebo effects are centered around zero for all the simulations. Comparing the two sets of placebo results (Figures

3 and Figures 6), one can see that the hourly patterns that previously emerged in the difference-in-differences placebos are no longer as systematic for the machine learning estimates. Thus, the machine learning methodology appears to more successfully control for compositional changes over time.

[Figures 3-6 about here]

5 Realized versus Ex-Ante Predicted Energy Savings

Both the regression and machine learning approach provide estimates that suggest average energy savings at schools that range between 2 to 5%, depending on the specification. While these effects provide an average sense of the treatment effect, the interventions across schools can be quite different, ranging from replacing a few fixtures, to upgrading a whole HVAC system. Therefore, it might be difficult to interpret such measures in terms of success or failure of these interventions.

In this section, we take advantage of the fact that each intervention has an engineering estimate of expected savings as a first step to understand the effectiveness of these heterogeneous measures. We estimate a treatment effect that is proportional to these measures of expected savings, so that interventions that are quite different in scope can be easily compared.

5.1 Expected savings in the data

Before turning to the regression analysis on electricity savings, it is useful to understand the variation in expected savings present in the data. Figure 7 shows the distribution of expected savings in the population of schools that implement energy efficiency upgrades. Most interventions represents savings of less than 10% of a school's average electricity consumption. However, there is a long tail of interventions that represent much larger savings.

[Figure 7 about here]

The distribution of expected savings also highlights some of the challenges with measuring the effectiveness of these interventions, given the vast heterogeneity across measures. Furthermore, there are potential measurement issues with the data. For example, we find that some schools have reported expected energy savings that exceed their average electricity consumption in our data, which cannot be possibly true. The measurement issues could be due to some measurement error in expected savings, or due to a mismatch between schools and interventions.

5.2 Assessing realized vs. expected savings

To estimate the effectiveness of the interventions when compared to ex-ante expected savings, we consider the following set of regressions (and variants):

$$Y_{ith} = -\beta S_{it} + \alpha_i + \kappa_h + \gamma_t + \epsilon_{ith},$$

where Y represents electricity consumption for the difference-in-differences approach, and prediction error for the machine learning methodology. The variable S_i represents expected ex-ante energy

savings for a given intervention, which are estimated at the annual level. In particular, we define the expected energy savings at any point in time to only include the energy savings that belong to efficiency projects that have already taken place.

The main coefficient of interest is β. Note that these regressions are in levels, as opposed to logs, so that a coefficient of one can be interpreted as ex-post expected savings matching on average ex-ante energy savings. A coefficient larger than one would suggest that the observed energy savings are larger than those estimated ex-ante. On the contrary, an estimate smaller than one would suggest that the ex-post realized savings are not as large as anticipated.

[Table 9 about here]

Table 9 presents the results for all interventions (Panel A), HVAC interventions (Panel B) and lighting interventions (Panel C). In each panel, we present the estimates for the regression approach, in which the dependent variable is electricity consumption followed by estimates based on the machine learning approach, in which the dependent variable is prediction error from the machine learning model.

The results suggest that realized savings are consistently lower than predicted ex ante. None of the estimated coefficients, which are akin to realization rates, are above 50 percent and some are closer to 20 percent. Across interventions, the results suggest that expected energy savings from HVAC and lighting interventions match ex ante expected savings more closely than other interventions. These highlights some of the difficulties in assessing energy efficiency investment in the presence of very heterogeneous measures.

Energy efficiency policy discussions sometimes distinguish between "net" and "gross" savings, where the former subtracts energy efficiency investments customers would make absent a utility program. Because the machine learning approach provides school-specific counterfactuals, it allows us to disentangle the extent to which untreated schools (i.e., schools who are not receiving rebates from the utility) are also reducing their consumption over the time period. For example, comparing estimates of $\hat{\beta}^U$ and $\hat{\beta}^T$ in Table 6 suggests that untreated schools are not reducing consumption over time. This suggests that the low realized savings are not driven by unmeasured efficiency upgrades at untreated schools, and are more likely driven by overly optimistic ex ante predictions or rebound.

6 Conclusions

We study the incidence and impacts of energy efficiency investments at public K-12 schools in California. Using high-frequency hourly electricity data, we estimate the treatment effect of energy efficiency projects at treated schools, leveraging the presence of control schools that did not participate in these programs. We explore two complementary methodologies: a regression approach and a machine learning approach. We find that results are comparable across methodologies and a battery of controls, although machine learning estimates are systematically larger. We discuss in the methodological section how these differences can be explained by potential measurement error and specification bias.

Focusing on HVAC and lighting interventions, we find that these energy efficiency investments delivered about 2 to 4% electricity consumption savings on average, when compared to control

schools. These energy savings appear to be a substantial share of ex-ante expected savings for the lighting and HVAC interventions, with actual savings predicted to be around 70-90% of projected savings. However, realized savings appear to be noisily measured and small when we consider a wider battery of measures, with estimates of at most 15% of ex-ante expected savings. Heterogeneity in how expected savings are defined, as well as other sources of measurement error (e.g., in treatment date), could be driving some of these results.

In future work, we plan to further analyze the relative merits of each of these methodologies, as well as to extend the cost-benefit analysis. Furthermore, we plan to extend the empirical assessment of both regression and prediction methods in the estimation of energy efficiency savings.

References

Allcott, H. and Greenstone, M. (2012). Is there an energy efficiency gap? *The Journal of Economic Perspectives*, 6(1):3–28. 1

Athey, S. and Imbens, G. (2015). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Science*, 113(27):7353–7360. 4.1.1

Barbose, G. L., Goldman, C. A., Hoffman, I. M., and Billingsley, M. A. (2013). The future of utility customer-funded energy efficiency programs in the United States: projected spending and savings to 2025. *Energy Efficiency Journal*, 6(3):475–493. 2

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection amongst high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650. 4.1.1

Boomhower, J. and Davis, L. (2016). Do energy efficiency investments deliver at the right time? *Energy Institute at Haas Working Paper*, (WP 271). 12

CEC (2017). Proposition 39: California clean energy jobs act, k-12 program and energy conservation assistance act 2015-2016 progress report. Technical report, California Energy Commission. 2

Cicala, S. (2017). Imperfect markets versus imperfect regulation in u.s. electricity generation. *NBER Working Paper*. 1

Davis, L., Fuchs, A., and Gertler, P. (2014). Cash for coolers: evaluating a large-scale appliance replacement program in Mexico. *American Economic Journal: Economic Policy*, 6(4):207–238. 1

EIA (2015). Electric power monthly. Technical report, Energy Information Administration. 1

Eichholtz, P., Kok, N., and Quigley, J. M. (2013). The economics of green building. *Review of Economics and Statistics*, 95(1):50–63. 2

Fisher, A., Haneman, M., Roberts, M., and Schlenker, W. (2012). The economic impacts of climate change: Evidence from agricultural output and random fluctuations in weather: Comment. *American Economic Review*, 102(7):1749–1760. 1

Fowlie, M., Greenstone, M., and Wolfram, C. (2015). Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program. *Mimeograph*. 1

Gerarden, T. D., Newell, R. G., and Stavins, R. N. (2015). Assessing the energy-efficiency gap. Technical report, Harvard Environmental Economics Program. 1

Gillingham, K. and Palmer, K. (2014). Bridging the energy efficiency gap: policy insights from economic theory and empirical evidence. *Review of Environmental Economics and Policy*, 8(1):18–38. 1

IEA (2015). World energy outlook. Technical report, International Energy Agency. 1

Joskow, P. L. and Marron, D. B. (1992). What does a negawatt really cost? Evidence from utility conservation programs. *The Energy Journal*, 13(4):41–74. 1

Kahn, M., Kok, N., and Quigley, J. (2014). Carbon emissions from the commercial building sector: The role of climate, quality, and incentives. *Journal of Public Economics*, 113:1–12. 2

Kok, N. and Jennen, M. (2012). The impact of energy labels and accessibility on office rents. *Energy Policy*, 46(C):489–497. 2

Kushler, M. (2015). Residential energy efficiency works: Don't make a mountain out of the E2e molehill. *American Council for an Energy-Efficient Economy Blog*. 1

Levinson, A. (2016). How much energy do building energy codes save? evidence from california houses. *American Economic Review*, 106(10):2867–2894. 1

McCaffrey, D., Ridgeway, G., and Morral, A. (2004). Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *RAND Journal of Economics*, 9(4):403–425. 4.1.1

McKinsey & Company (2009). Unlocking energy efficiency in the U.S. economy. Technical report, McKinsey Global Energy and Materials. 1, 2

Myers, E. (2014). Asymmetric information in residential rental markets: implications for the energy efficiency gap. *Job Market Paper*. 1

Wyss, R., Ellis, A., Brookhart, A., Girman, C., Funk, M., LoCasale, R., and Sturmer, T. (2014). The role of prediction modeling in propensity score estimation: An evaluation of logistic regression, bcart, and the covariate-balancing propensity score. *American Journal of Epidemiology*, 180(6):645–655. 4.1.1

Table 1: Average characteristics of schools in the sample

Category	Untreated	Any Interventions		HVAC Interventions		Lighting Interventions	
		Treated	T-U	Treated	T-U	Treated	T-U
Hourly energy consumption (kWh)	34.2	57.1	22.9	63.5	29.4	61.6	27.5
	[15.3, 39.1]	[24.7, 59.5]	(2.5)	[28.7, 63.9]	(2.9)	[27.3, 60.8]	(3.1)
First year in sample	2012	2009.9	-2.0	2009.5	-2.3	2009.8	-2.1
	[2011.0, 2013.0]	[2008.0, 2012.0]	(0.1)	[2008.0, 2011.0]	(0.1)	[2008.0, 2011.0]	(0.1)
Total enrollment	553.5	722.3	168.8	765.9	212.4	747.7	194.2
	[355.4, 658.9]	[462.5, 808.3]	(19.9)	[496.8, 832.0]	(22.8)	[479.8, 812.4]	(24.5)
Academic Performance Index	789.4	793.7	4.3	792.8	3.4	786.5	-2.9
	[734.8, 855.3]	[737.9, 854.6]	(4.2)	[734.0, 856.0]	(4.9)	[736.5, 836.0]	(5.1)
Bond passed – 2 yrs	0.3	0.3	0.0	0.3	-0.0	0.3	-0.1
	[0.0, 1.0]	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)
Bond passed – 5 yrs	0.4	0.4	0.0	0.4	-0.0	0.4	-0.1
	[0.0, 1.0]	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)	[0.0, 1.0]	(0.0)
High School Graduates	23.5	23.3	-0.2	23.7	0.1	24.1	0.6
	[14.4, 32.0]	[15.3, 31.8]	(0.5)	[15.3, 32.5]	(0.6)	[17.5, 31.5]	(0.7)
College Graduates	20.1	20.3	0.2	19.4	-0.6	19.7	-0.4
	[9.5, 29.8]	[9.8, 29.4]	(0.5)	[9.0, 28.8]	(0.6)	[9.3, 28.5]	(0.7)
Percent single mothers	20.1	19.4	-0.7	20.0	-0.1	19.9	-0.2
	[4.8, 30.4]	[4.9, 28.8]	(0.9)	[4.9, 30.4]	(1.0)	[4.8, 31.1]	(1.1)
Percent African-American	5.7	6.2	0.5	5.5	-0.3	5.9	0.1
	[1.0, 6.3]	[1.3, 7.8]	(0.4)	[1.3, 7.3]	(0.4)	[1.3, 8.1]	(0.5)
Percent Asian	9.5	11.3	1.9	12.3	2.8	9.6	0.1
	[1.0, 11.0]	[1.8, 13.3]	(0.7)	[2.0, 13.8]	(0.8)	[1.8, 12.3]	(0.7)
Percent Hispanic	42.0	43.7	1.7	45.6	3.6	46.2	4.2
	[16.0, 65.5]	[21.0, 65.0]	(1.2)	[22.3, 68.8]	(1.4)	[25.0, 65.1]	(1.5)
Percent White	34.3	30.7	-3.6	29.8	-4.5	29.7	-4.6
	[8.5, 56.5]	[8.8, 51.9]	(1.1)	[8.3, 49.3]	(1.3)	[8.5, 48.0]	(1.4)
Percent 2+ races	2.4	2.0	-0.4	1.6	-0.8	1.9	-0.5
	[0.3, 3.5]	[0.3, 2.8]	(0.1)	[0.0, 2.0]	(0.1)	[0.3, 2.8]	(0.2)
Temperature (F)	60.2	60.7	0.5	61.2	1.0	60.9	0.7
	[57.5, 62.6]	[58.1, 63.0]	(0.2)	[58.4, 63.7]	(0.2)	[58.2, 63.2]	(0.2)
Latitude	37.7	37.5	-0.2	37.4	-0.3	37.4	-0.2
	[37.2, 38.2]	[36.9, 38.0]	(0.0)	[36.8, 38.0]	(0.1)	[36.8, 38.0]	(0.1)
Longitude	-121.6	-121.3	0.4	-121.0	0.6	-121.2	0.4
	[-122.2, -121.1]	[-122.0, -120.4]	(0.0)	[-122.0, -119.8]	(0.1)	[-122.0, -120.0]	(0.1)
Number of schools	1055	1039		628		493	

Notes: This table displays characteristics of the treated and untreated schools in our sample, by type of intervention. Interquantile range in brackets; p-value on differences between treated and untreated schools in parentheses.

Table 2: Diff-in-Diff Results by Hour-Block

	(1)	(2)	(3)	(4)	(5)
Treatment (aggregate)	-0.0454	-0.0450	-0.0488	-0.0191	-0.0174
	(0.0068)	(0.0068)	(0.0068)	(0.0082)	(0.0082)
Midn. to 3 AM x Treat	-0.0560	-0.0505	-0.0538	-0.0242	-0.0222
	(0.0096)	(0.0081)	(0.0082)	(0.0092)	(0.0092)
3 AM to 6 AM x Treat	-0.0571	-0.0487	-0.0519	-0.0223	-0.0204
	(0.0093)	(0.0082)	(0.0083)	(0.0092)	(0.0093)
6 AM to 9 AM x Treat	-0.0229	-0.0264	-0.0306	-0.0007	0.0008
	(0.0079)	(0.0066)	(0.0064)	(0.0084)	(0.0082)
9 AM to Noon x Treat	-0.0416	-0.0490	-0.0518	-0.0237	-0.0205
	(0.0084)	(0.0061)	(0.0062)	(0.0082)	(0.0081)
Noon to 3 PM x Treat	-0.0362	-0.0454	-0.0481	-0.0201	-0.0167
	(0.0088)	(0.0067)	(0.0068)	(0.0085)	(0.0084)
3 PM to 6 PM x Treat	-0.0454	-0.0438	-0.0503	-0.0180	-0.0188
	(0.0090)	(0.0082)	(0.0082)	(0.0092)	(0.0092)
6 PM to 9 PM x Treat	-0.0514	-0.0453	-0.0506	-0.0192	-0.0191
	(0.0089)	(0.0080)	(0.0081)	(0.0091)	(0.0091)
9 PM to Midn. x Treat	-0.0524	-0.0511	-0.0535	-0.0247	-0.0220
	(0.0089)	(0.0078)	(0.0079)	(0.0089)	(0.0089)
Observations	19,253,018	19,252,986	19,251,960	19,252,986	19,251,960
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is log hourly electricity consumption, averaged by blocks of three hours.

Table 3: Diff-in-Diff Results by Type of Intervention

	(1)	(2)	(3)	(4)	(5)
HVAC Interventions: (aggregate)	-0.0489	-0.0484	-0.0523	-0.0238	-0.0213
	(0.0089)	(0.0089)	(0.0089)	(0.0108)	(0.0103)
Midn. to 3 AM x Treat	-0.0654	-0.0538	-0.0582	-0.0304	-0.0293
	(0.0120)	(0.0100)	(0.0101)	(0.0111)	(0.0108)
3 AM to 6 AM x Treat	-0.0671	-0.0521	-0.0560	-0.0286	-0.0271
	(0.0115)	(0.0100)	(0.0101)	(0.0111)	(0.0109)
6 AM to 9 AM x Treat	-0.0193	-0.0195	-0.0271	0.0034	0.0018
	(0.0100)	(0.0082)	(0.0080)	(0.0098)	(0.0093)
9 AM to Noon x Treat	-0.0408	-0.0457	-0.0496	-0.0230	-0.0209
	(0.0103)	(0.0076)	(0.0075)	(0.0094)	(0.0090)
Noon to 3 PM x Treat	-0.0302	-0.0416	-0.0444	-0.0190	-0.0157
	(0.0107)	(0.0083)	(0.0082)	(0.0100)	(0.0095)
3 PM to 6 PM x Treat	-0.0320	-0.0406	-0.0477	-0.0177	-0.0188
	(0.0110)	(0.0103)	(0.0102)	(0.0113)	(0.0108)
6 PM to 9 PM x Treat	-0.0526	-0.0498	-0.0561	-0.0265	-0.0273
	(0.0112)	(0.0101)	(0.0102)	(0.0111)	(0.0108)
9 PM to Midn. x Treat	-0.0526	-0.0557	-0.0591	-0.0323	-0.0302
	(0.0111)	(0.0094)	(0.0094)	(0.0105)	(0.0102)
Observations	14,939,790	14,939,760	14,938,895	14,939,760	14,938,895
Light Interventions: (aggregate)	-0.0674	-0.0669	-0.0705	-0.0421	-0.0427
	(0.0095)	(0.0095)	(0.0097)	(0.0124)	(0.0121)
Midn. to 3 AM x Treat	-0.0575	-0.0486	-0.0510	-0.0223	-0.0196
	(0.0128)	(0.0110)	(0.0112)	(0.0122)	(0.0122)
3 AM to 6 AM x Treat	-0.0593	-0.0489	-0.0537	-0.0227	-0.0223
	(0.0126)	(0.0111)	(0.0113)	(0.0122)	(0.0124)
6 AM to 9 AM x Treat	-0.0274	-0.0370	-0.0443	-0.0112	-0.0131
	(0.0105)	(0.0084)	(0.0083)	(0.0105)	(0.0103)
9 AM to Noon x Treat	-0.0515	-0.0629	-0.0668	-0.0376	-0.0356
	(0.0110)	(0.0080)	(0.0081)	(0.0103)	(0.0105)
Noon to 3 PM x Treat	-0.0416	-0.0606	-0.0622	-0.0354	-0.0311
	(0.0114)	(0.0089)	(0.0089)	(0.0109)	(0.0110)
3 PM to 6 PM x Treat	-0.0582	-0.0559	-0.0608	-0.0303	-0.0295
	(0.0116)	(0.0109)	(0.0110)	(0.0122)	(0.0123)
6 PM to 9 PM x Treat	-0.0647	-0.0531	-0.0581	-0.0271	-0.0268
	(0.0116)	(0.0104)	(0.0106)	(0.0116)	(0.0118)
9 PM to Midn. x Treat	-0.0659	-0.0553	-0.0561	-0.0290	-0.0247
	(0.0119)	(0.0108)	(0.0110)	(0.0120)	(0.0121)
Observations	12,940,064	12,940,036	12,939,239	12,940,036	12,939,239
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is log hourly electricity consumption, averaged by blocks of three hours. Only schools that experienced either an HVAC or light upgrade are included in the treated sample, respectively.

Table 4: Matching Results

	(1)	(2)	(3)	(4)	(5)
Any :nterventions:					
Any district	-0.0504	-0.0506	-0.0541	-0.0289	-0.0304
	(0.0124)	(0.0124)	(0.0131)	(0.0123)	(0.0130)
Same district	-0.0149	-0.0147	-0.0096	0.0008	0.0100
	(0.0141)	(0.0141)	(0.0149)	(0.0143)	(0.0147)
Opposite district	-0.0471	-0.0471	-0.0506	-0.0215	-0.0258
	(0.0119)	(0.0120)	(0.0127)	(0.0123)	(0.0130)
Observations	4,828,122	4,828,108	4,826,977	4,828,108	4,826,977
HVAC Interventions:					
Any district	-0.0098	-0.0100	-0.0054	-0.0113	-0.0068
	(0.0263)	(0.0263)	(0.0273)	(0.0265)	(0.0265)
Same district	0.0079	0.0081	0.0132	0.0136	0.0157
	(0.0222)	(0.0222)	(0.0236)	(0.0213)	(0.0215)
Opposite district	-0.0665	-0.0668	-0.0679	-0.0394	-0.0335
	(0.0151)	(0.0151)	(0.0156)	(0.0155)	(0.0166)
Observations	2,379,037	2,379,033	2,378,466	2,379,033	2,378,466
Lighting Interventions:					
Any district	-0.0602	-0.0599	-0.0562	-0.0339	-0.0251
	(0.0212)	(0.0212)	(0.0220)	(0.0205)	(0.0235)
Same district	-0.0459	-0.0455	-0.0451	-0.0270	-0.0287
	(0.0158)	(0.0158)	(0.0168)	(0.0168)	(0.0193)
Opposite district	-0.0461	-0.0462	-0.0492	-0.0040	-0.0051
	(0.0122)	(0.0122)	(0.0128)	(0.0190)	(0.0203)
Observations	1,914,567	1,914,563	1,914,147	1,914,563	1,914,147
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: This table displays regression results where the untreated group is chosen via nearest-neighbor matching. We match one untreated school to each treatment school. Each row in the table employs a different restriction on which schools are allowed to be matched to any given treatment school. "Any district" matches allow any control school to be matched to a treatment school; "same district" matches are restricted to untreated schools in the same school district, and "opposite district" matches are restricted to untreated schools from different districts. In each case, the matching variables are the mean, maximum, and standard deviation of electricity consumption in each three-hour block (e.g. 9 AM-Noon) from the pre-treatment period; demographic indicators at the census block level, including the poverty rate, log of per capita income, school-level indicators (enrollment; age of the school; grades taught; an academic performance index; and climate).

Table 5: Monte Carlo Results, Percent Deviations from "True" Effect

DGP	Specification	Regression	Machine Learning
No Measurement Error	0		0.0213
			(0.0200)
	1	-0.1656	-0.0029
		(0.0071)	(0.0107)
	2	-0.0032	0.0057
		(0.0027)	(0.0068)
	3	0.0005	0.0130
		(0.0014)	(0.0018)
Wrong treatment date (Early)	0		-0.1870
			(0.0212)
	1	-0.4259	-0.2301
		(0.0107)	(0.0149)
	2	-0.3002	-0.2216
		(0.0089)	(0.0094)
	3	-0.3102	-0.2113
		(0.0106)	(0.0093)
Wrong treatment date (Late)	0		0.0050
			(0.0197)
	1	-0.3538	-0.1569
		(0.0148)	(0.0111)
	2	-0.2818	-0.1516
		(0.0133)	(0.0090)
	3	-0.2827	-0.1438
		(0.0126)	(0.0083)

Notes: This table reports percentage deviations from the "true" treatment effect using regression and machine learning approaches under different data generating processes. "No Measurement Error" indicates that the reported date of treatment is accurate. "Wrong treatment date (Early)" and "Wrong treatment date (Late)" refer to reported treatment dates (used by the researcher) that are before and after the true treatment date, respectively. Specification 0 has no second-stage control variables, Specification 1 includes school FEs, Specification 2 includes school, month, and year FEs, and Specification 3 includes school-by-block, month-of-sample, and school-specific temperature controls. Standard errors, clustered by school, in parentheses.

Table 6: Prediction Results - Average prediction errors

	(1) $\hat{\beta}^{U}$	(2) $\hat{\beta}^{T}$	(3) $\hat{\beta}^{DD}$	(4) $\hat{\beta}^{UD}$	(5) $\hat{\beta}^{TD}$	(6) $\hat{\beta}^{3D}$
Aggregate	-0.0042	-0.0465	-0.0423	-0.0113	-0.0529	-0.0501
	(0.0050)	(0.0061)	(0.0079)	(0.0050)	(0.0060)	(0.0063)
Midn. to 3 AM	-0.0149	-0.0510	-0.0361	-0.0165	-0.0513	-0.0361
	(0.0073)	(0.0081)	(0.0109)	(0.0074)	(0.0081)	(0.0109)
3 AM to 6 AM	-0.0146	-0.0467	-0.0321	-0.0169	-0.0472	-0.0321
	(0.0074)	(0.0081)	(0.0110)	(0.0074)	(0.0081)	(0.0110)
6 AM to 9 AM	-0.0051	-0.0272	-0.0221	-0.0126	-0.0335	-0.0221
	(0.0055)	(0.0058)	(0.0080)	(0.0054)	(0.0057)	(0.0080)
9 AM to Noon	0.0024	-0.0360	-0.0384	-0.0164	-0.0558	-0.0384
	(0.0050)	(0.0053)	(0.0073)	(0.0051)	(0.0053)	(0.0073)
Noon to 3 PM	0.0087	-0.0402	-0.0489	-0.0092	-0.0596	-0.0489
	(0.0051)	(0.0057)	(0.0076)	(0.0050)	(0.0056)	(0.0076)
3 PM to 6 PM	0.0075	-0.0572	-0.0647	0.0024	-0.0617	-0.0647
	(0.0065)	(0.0073)	(0.0097)	(0.0065)	(0.0073)	(0.0097)
6 PM to 9 PM	-0.0055	-0.0582	-0.0527	-0.0083	-0.0586	-0.0527
	(0.0066)	(0.0075)	(0.0100)	(0.0066)	(0.0075)	(0.0100)
9 PM to Midn.	-0.0111	-0.0547	-0.0437	-0.0126	-0.0554	-0.0437
	(0.0071)	(0.0077)	(0.0105)	(0.0071)	(0.0076)	(0.0105)
Observations	3,434,982	7,341,034	10,776,016	6,916,585	12,331,889	19,248,474

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours. Due to the presence of large prediction errors in some observations, the reported averages trim the p1 and p99 outliers.

Table 7: Prediction Results by Hour-Block

	(1)	(2)	(3)	(4)	(5)
Treatment (aggregate)	-0.0400	-0.0403	-0.0417	-0.0266	-0.0237
	(0.0068)	(0.0068)	(0.0071)	(0.0076)	(0.0076)
Midn. to 3 AM x Treat	-0.0375	-0.0411	-0.0426	-0.0273	-0.0245
	(0.0087)	(0.0082)	(0.0085)	(0.0087)	(0.0088)
3 AM to 6 AM x Treat	-0.0315	-0.0368	-0.0386	-0.0230	-0.0205
	(0.0088)	(0.0083)	(0.0086)	(0.0088)	(0.0089)
6 AM to 9 AM x Treat	-0.0205	-0.0201	-0.0213	-0.0064	-0.0033
	(0.0068)	(0.0064)	(0.0066)	(0.0074)	(0.0074)
9 AM to Noon x Treat	-0.0422	-0.0457	-0.0469	-0.0321	-0.0289
	(0.0064)	(0.0061)	(0.0065)	(0.0074)	(0.0074)
Noon to 3 PM x Treat	-0.0457	-0.0457	-0.0457	-0.0321	-0.0277
	(0.0067)	(0.0064)	(0.0068)	(0.0075)	(0.0076)
3 PM to 6 PM x Treat	-0.0528	-0.0475	-0.0500	-0.0339	-0.0320
	(0.0081)	(0.0078)	(0.0081)	(0.0084)	(0.0085)
6 PM to 9 PM x Treat	-0.0480	-0.0430	-0.0444	-0.0291	-0.0264
	(0.0083)	(0.0081)	(0.0084)	(0.0086)	(0.0087)
9 PM to Midn. x Treat	-0.0421	-0.0430	-0.0445	-0.0292	-0.0264
	(0.0084)	(0.0080)	(0.0083)	(0.0085)	(0.0086)
Observations	19,253,016	19,252,988	19,251,882	19,252,988	19,251,882
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours.

Table 8: Prediction Results by Type of Intervention

	(1)	(2)	(3)	(4)	(5)
HVAC Interventions: (aggregate)	-0.0444	-0.0447	-0.0464	-0.0319	-0.0292
	(0.0081)	(0.0081)	(0.0083)	(0.0092)	(0.0091)
Midn. to 3 AM x Treat	-0.0451	-0.0457	-0.0471	-0.0332	-0.0315
	(0.0102)	(0.0094)	(0.0096)	(0.0099)	(0.0099)
3 AM to 6 AM x Treat	-0.0387	-0.0407	-0.0421	-0.0282	-0.0265
	(0.0102)	(0.0094)	(0.0097)	(0.0099)	(0.0100)
6 AM to 9 AM x Treat	-0.0175	-0.0139	-0.0166	-0.0016	-0.0010
	(0.0075)	(0.0070)	(0.0072)	(0.0080)	(0.0079)
9 AM to Noon x Treat	-0.0355	-0.0401	-0.0430	-0.0278	-0.0273
	(0.0070)	(0.0066)	(0.0068)	(0.0077)	(0.0076)
Noon to 3 PM x Treat	-0.0393	-0.0412	-0.0426	-0.0289	-0.0271
	(0.0073)	(0.0070)	(0.0073)	(0.0080)	(0.0079)
3 PM to 6 PM x Treat	-0.0438	-0.0441	-0.0471	-0.0318	-0.0315
	(0.0094)	(0.0091)	(0.0094)	(0.0096)	(0.0096)
6 PM to 9 PM x Treat	-0.0514	-0.0472	-0.0494	-0.0346	-0.0338
	(0.0098)	(0.0095)	(0.0098)	(0.0100)	(0.0100)
9 PM to Midn. x Treat	-0.0487	-0.0488	-0.0502	-0.0364	-0.0345
	(0.0098)	(0.0090)	(0.0093)	(0.0095)	(0.0095)
Observations	14,939,789	14,939,763	14,938,822	14,939,763	14,938,822
Light Interventions: (aggregate)	-0.0620	-0.0626	-0.0638	-0.0510	-0.0478
	(0.0093)	(0.0092)	(0.0096)	(0.0105)	(0.0105)
Midn. to 3 AM x Treat	-0.0277	-0.0352	-0.0358	-0.0237	-0.0192
	(0.0108)	(0.0106)	(0.0108)	(0.0110)	(0.0112)
3 AM to 6 AM x Treat	-0.0240	-0.0355	-0.0370	-0.0241	-0.0205
	(0.0112)	(0.0107)	(0.0110)	(0.0111)	(0.0113)
6 AM to 9 AM x Treat	-0.0269	-0.0311	-0.0329	-0.0198	-0.0163
	(0.0085)	(0.0077)	(0.0078)	(0.0088)	(0.0089)
9 AM to Noon x Treat	-0.0596	-0.0610	-0.0617	-0.0498	-0.0452
	(0.0078)	(0.0072)	(0.0076)	(0.0087)	(0.0089)
Noon to 3 PM x Treat	-0.0641	-0.0587	-0.0572	-0.0476	-0.0408
	(0.0081)	(0.0076)	(0.0080)	(0.0090)	(0.0092)
3 PM to 6 PM x Treat	-0.0675	-0.0561	-0.0567	-0.0450	-0.0402
	(0.0102)	(0.0097)	(0.0101)	(0.0104)	(0.0107)
6 PM to 9 PM x Treat	-0.0565	-0.0475	-0.0481	-0.0361	-0.0316
	(0.0103)	(0.0100)	(0.0103)	(0.0106)	(0.0107)
9 PM to Midn. x Treat	-0.0425	-0.0436	-0.0439	-0.0321	-0.0273
	(0.0106)	(0.0105)	(0.0107)	(0.0109)	(0.0111)
Observations	12,940,115	12,940,089	12,939,223	12,940,089	12,939,223
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

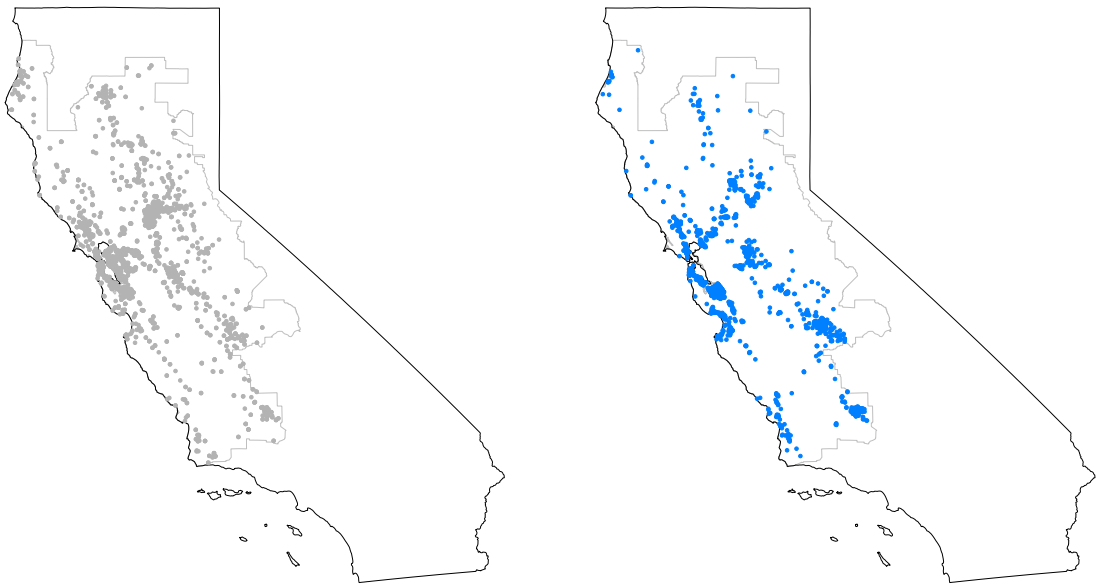
Notes: Standard errors clustered at the school level. Dependent variable is the prediction error in log electricity consumption consumption by blocks of three hours. Only schools that experienced either an HVAC or light upgrade are included in the treated sample, respectively.

Table 9: Ex-Post vs. Ex-Ante Savings

	(1)	(2)	(3)	(4)	(5)
Any Intervention:					
Energy consumption (kWh)	0.3009	0.3031	0.3167	0.2412	0.2471
	(0.0531)	(0.0536)	(0.0561)	(0.0531)	(0.0534)
Prediction error (kWh)	0.2517	0.2522	0.2534	0.2043	0.2098
	(0.0657)	(0.0663)	(0.0683)	(0.0650)	(0.0656)
Observations	19,253,017	19,252,989	19,251,978	19,252,989	19,251,978
HVAC Interventions:					
Energy consumption (kWh)	0.5841	0.5884	0.6238	0.5182	0.4797
	(0.1030)	(0.1039)	(0.1112)	(0.0998)	(0.1092)
Prediction error (kWh)	0.4669	0.4689	0.4698	0.4096	0.3838
	(0.1018)	(0.1028)	(0.1080)	(0.1026)	(0.1041)
Observations	14,941,851	14,941,824	14,940,960	14,941,824	14,940,960
Light Interventions:					
Energy consumption (kWh)	0.3861	0.3877	0.4061	0.2612	0.2592
	(0.0613)	(0.0608)	(0.0635)	(0.0640)	(0.0667)
Prediction error (kWh)	0.4196	0.4193	0.4211	0.3367	0.3283
	(0.0776)	(0.0783)	(0.0791)	(0.0868)	(0.0864)
Observations	12,948,240	12,948,213	12,947,422	12,948,213	12,947,422
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	No	Yes
Month of Sample FE	No	No	No	Yes	No
Month of Sample Ctrl.	No	No	No	No	Yes

Notes: Standard errors clustered at the school level. Estimates report the coefficient on expected savings, scaled over time as projects get implemented. A coefficient of one implies that ex-post savings were realized as expected on average.

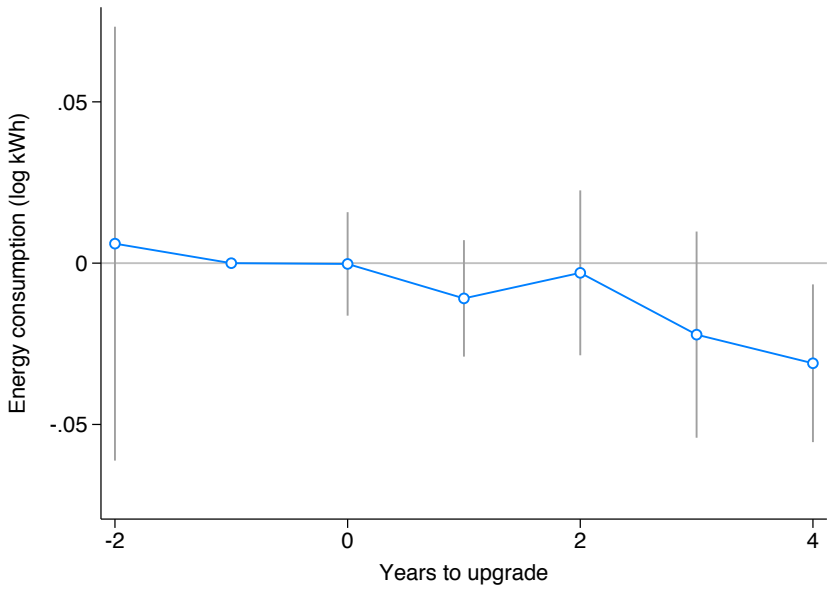
Figure 1: Locations of Treated and Untreated Schools



(a) Untreated (b) Treated

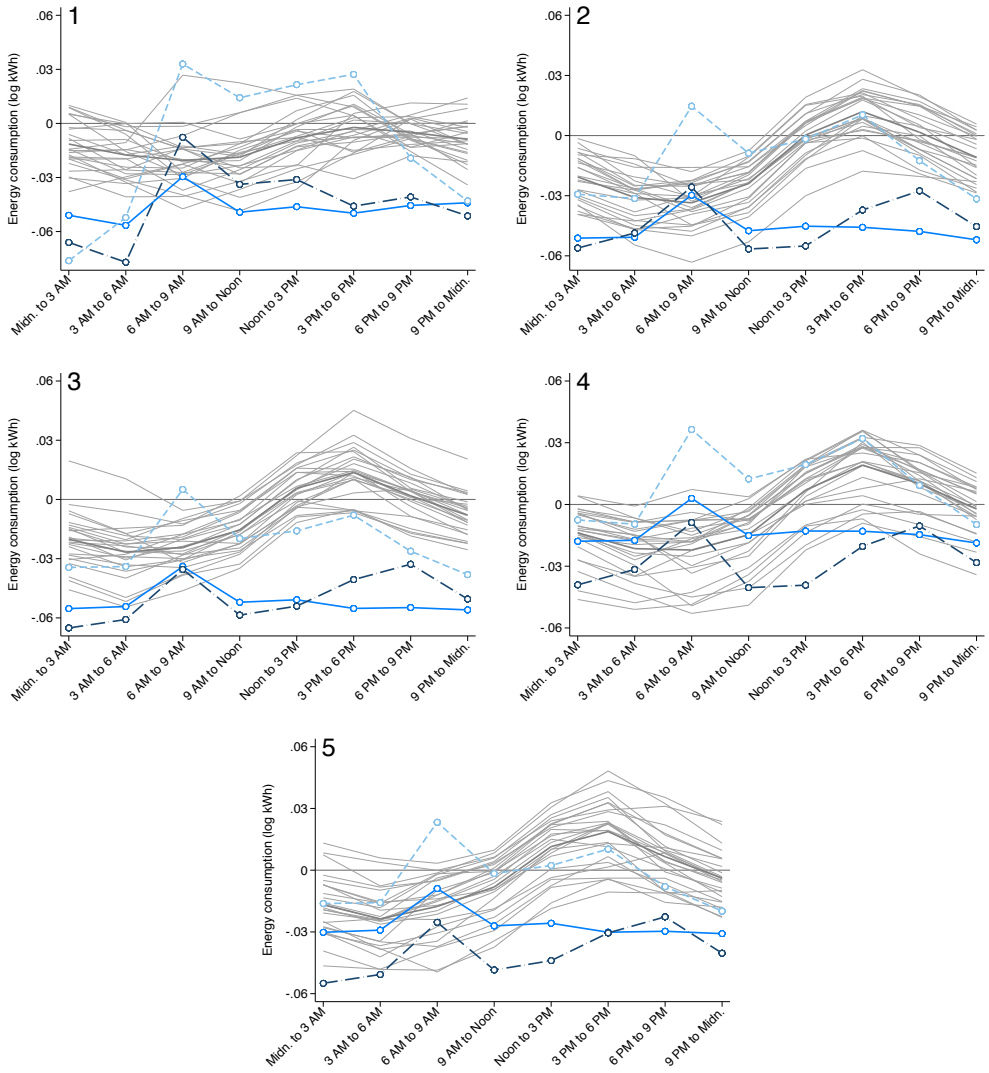
Notes: This figure displays the locations of schools in our sample. The left panel shows "untreated" schools that did not undertake any energy efficiency upgrades during our sample period, and the right panel shows "treated" schools that had at least one upgrade during our sample. There is substantial overlap in the locations of treated and untreated schools. The light gray outline shows the PG&E service territory.

Figure 2: Energy efficiency upgrades: Event study



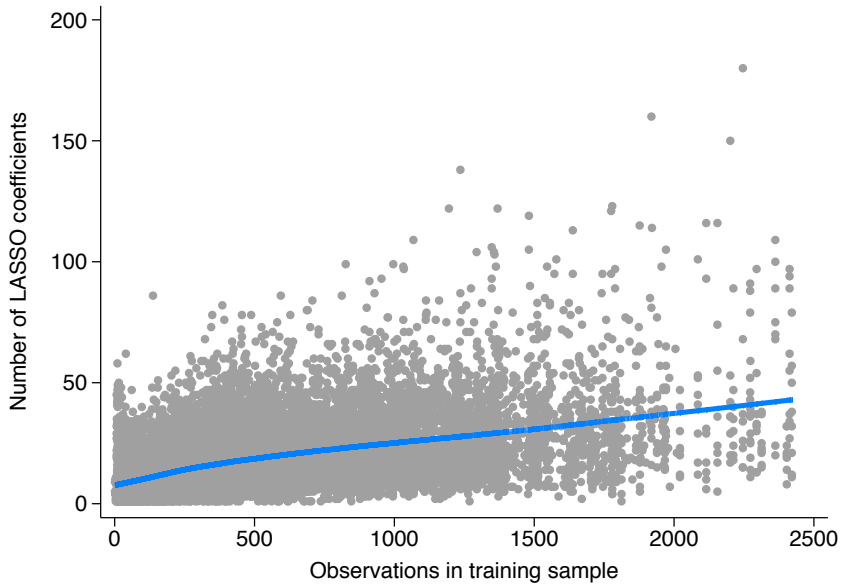
Notes: This event study figure displays the effects of installing any energy efficiency upgrade over time, having residualized school and block fixed effects. In order to leverage untreated observations, we randomly assign the untreated schools in our sample a treatment date. The point estimates presented here are the incremental changes in electricity consumption for treated units, relative to untreated units, in each time period. We find no statistically significant impact of energy efficiency upgrades prior to treatment, but a sharp decline in electricity consumption in the years following treatment. Because our panel is strongly unbalanced, we pool two or more years before treatment into a single coefficient, and four or more years after treatment into a single coefficient. We do not find a strong contemporaneous effect of energy efficiency upgrades, but given that many upgrades take place during the summer, we would not expect to see large effects during year zero.

Figure 3: Placebo Treatment Effects – Difference-in-Difference



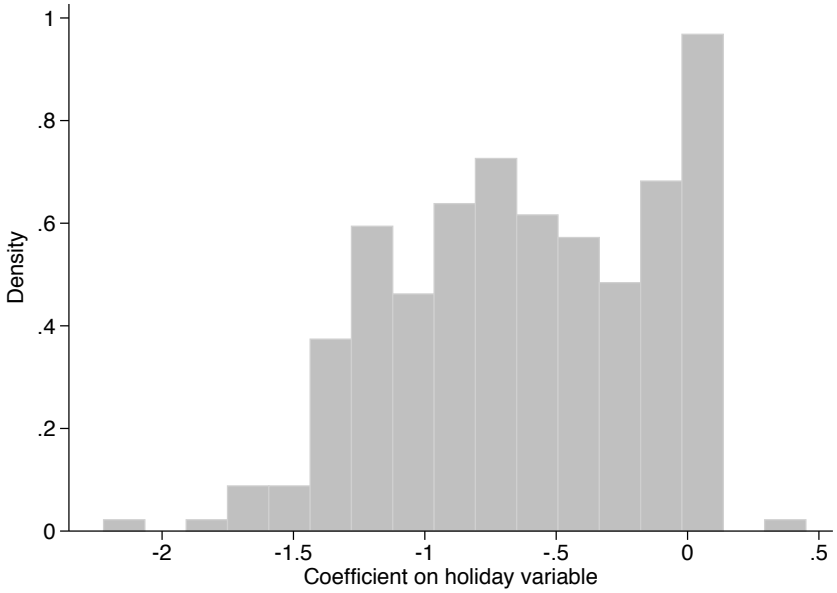
Notes: This figure presents difference-in-difference treatment effects using real and placebo data. Each panel corresponds to the column of the same number from Tables 2 and 3. The light gray lines present the placebo effects, with the solid gray line showing the average placebo effect. Placebo effects are generated using pre-treatment data only, and randomly assigning treatment status and timing according to the distribution present in real data. The solid blue line shows treatment effects from any upgrade; the dashed aqua line shows treatment effects for HVAC upgrades; and the dash-dotted navy line shows treatment effects from lighting upgrades, all using real data. Even with the most flexible specifications, which include school-by-hour-block fixed effects, we see marked hourly patterns in the placebos.

Figure 4: Number of LASSO coefficients by school



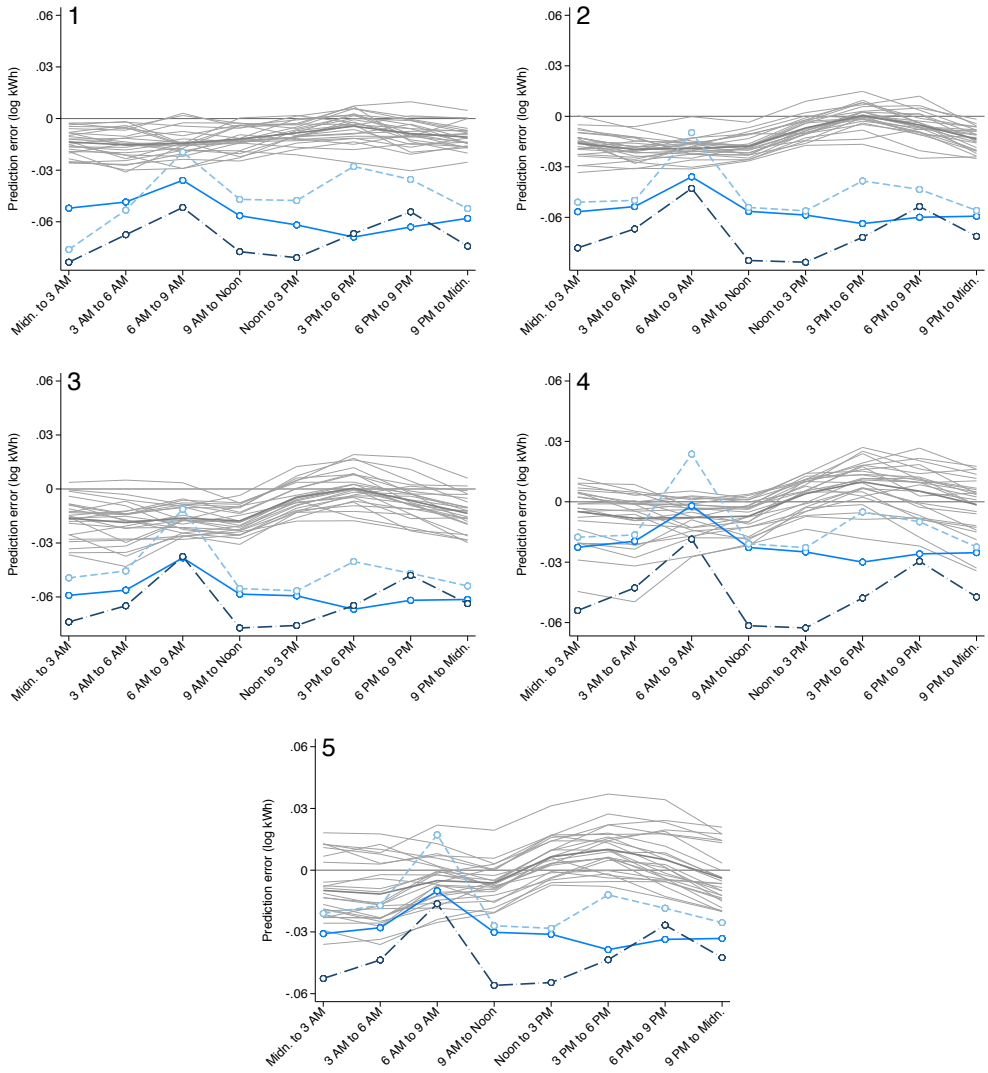
Notes: This figure shows the relationship between the number of observations for a school in the pre-treatment ("training") dataset and the number of variables the LASSO selects to include in the prediction model for that school, across every school in the sample. As expected, the more data available to the LASSO, the more variables it chooses to include. This provides evidence that the LASSO is not overfitting.

Figure 5: Holiday Effects in LASSO models



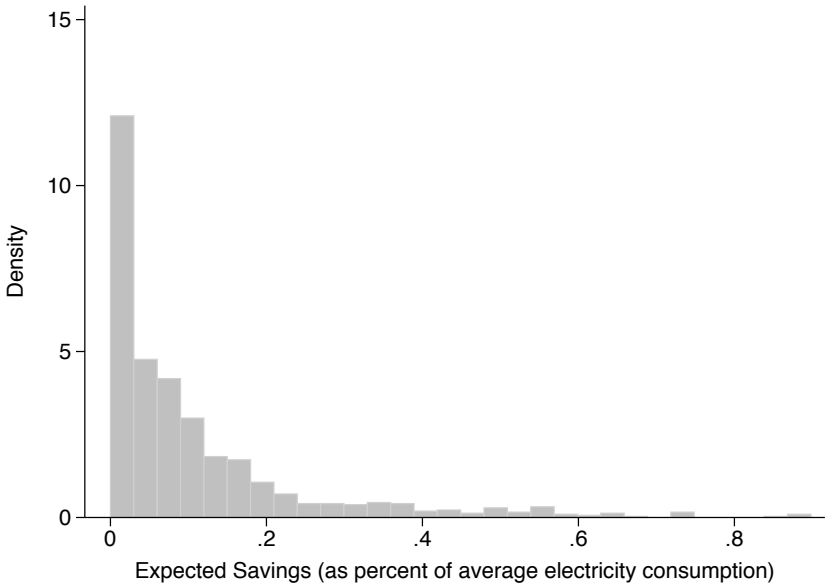
Notes: This figure displays the marginal effect of holidays on predicted energy consumption from the school-specific LASSOs.

Figure 6: Placebo Treatment Effects – Machine Learning



Notes: This figure presents machine learning treatment effects using real and placebo data. Each panel corresponds to the column of the same number from Tables 7 and 8. The light gray lines present the placebo effects, with the solid gray line showing the average placebo effect. Placebo effects are generated using pre-treatment data only, and randomly assigning treatment status and timing according to the distribution present in real data. The solid blue line shows treatment effects from any upgrade; the dashed aqua line shows treatment effects for HVAC upgrades; and the dash-dotted navy line shows treatment effects from lighting upgrades, all using real data. Unlike when we use a difference-in-difference method (as in Figure 3), we see little to no hourly patterns in the placebo treatment effects with machine learning.

Figure 7: Ex-ante energy savings



Notes: This figure displays the distributions of ex-ante expected savings as a percent over average electricity consumption. Only schools with interventions (i.e., non-zero expected savings are reported). The figure excludes schools for which expected savings are above 100%.