E2e Working Paper 032

Machine Learning from Schools about Energy Efficiency

Fiona Burlig, Christopher Knittel, David Rapson,

Mar Reguant, and Catherine Wolfram

September 2017

This paper is part of the E2e Project Working Paper Series.

E2e is a joint initiative of the Energy Institute at Haas at the University of California, Berkeley, the Center for Energy and Environmental Policy Research (CEEPR) at the Massachusetts Institute of Technology, and the Energy Policy Institute at Chicago, University of Chicago. E2e is supported by a generous grant from The Alfred P. Sloan Foundation.

Machine Learning from Schools about Energy Efficiency

Fiona Burlig
University of Chicago

Christopher Knittel
MIT

David Rapson
UC Davis

Mar Reguant
Northwestern University

Catherine Wolfram*
UC Berkeley

September 27, 2017

Abstract

In the United States, consumers invest billions of dollars annually in energy efficiency, often on the assumption that these investments will pay for themselves via future energy cost reductions. We study energy efficiency upgrades in K-12 schools in California. We develop and implement a novel machine learning approach for estimating treatment effects using high-frequency panel data, and demonstrate that this method outperforms standard panel fixed effects approaches. We find that energy efficiency upgrades reduce electricity consumption by 3 percent, but that these reductions total only 24 percent of *ex ante* expected savings. HVAC and lighting upgrades perform better, but still deliver less than half of what was expected. Finally, beyond location, school characteristics that are readily available to policymakers do not appear to predict realization rates across schools, suggesting that improving realization rates via targeting may prove challenging.

*Burlig: Department of Economics and Energy Policy Institute, University of Chicago, burlig@uchicago.edu. Knittel: Sloan School of Management and Center for Energy and Environmental Policy Research, MIT and NBER, knittel@mit.edu. Rapson: Department of Economics, UC Davis, dsrapson@ucdavis.edu. Reguant: Department of Economics, Northwestern University, CEPR and NBER, mar.reguant@northwestern.edu. Haas School of Business and Energy Institute at Haas, 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 at Haas 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, Yale University, and the 2016 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. Burlig was generously supported by the National Science Foundation's Graduate Research Fellowship Program under Grant DGE-1106400. All remaining errors are our own.

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 (International Energy Agency (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)). Indeed, it is not uncommon for analyses to project that the lifetime costs of these investments are negative. 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), and Gerarden, Newell, and Stavins (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.

A second strand of literature seeks to better understand the real-world savings and costs of energy efficiency investments. Analyses such as McKinsey & Company (2009) are based on engineering estimates of both the investment costs and the potential energy savings over time rather than field evidence. There are a variety of reasons why these engineering estimates might understate the costs consumers face or overstate savings. Economists have also pointed out that accurately measuring the savings from 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, Greenstone, and Wolfram (forthcoming)) and quasi-experimental (e.g., Allcott and Greenstone (2017), Levinson (2016), Myers (2015), and Davis, Fuchs, and Gertler (2014)) approaches to developing this counterfactual. In this paper, we leverage tools from machine learning to develop and implement a new approach for accurately estimating treatment effects using observational data. We apply our approach to energy efficiency upgrades in K-12 schools in California—an important extension of the previous literature which has focused on residential energy efficiency (Kushler (2015)). Our method can also be applied in a broad class of high-frequency panel data settings.

We take advantage of two recent advances, one technological and one methodological, to construct counterfactual energy consumption paths after energy efficiency investments. The first advance is the proliferation of high-frequency data in electricity markets, which provides a promising opportunity to estimate treatment effects associated with energy efficiency investments wherever advanced metering infrastructure (AMI, or "smart metering") is installed.[1] From a methodological perspective, high frequency data provide large benefits, but also presents new challenges. Using hourly 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

1. Over 50 percent of US households had smart meters as of 2016, and deployments are predicted to increase by over a third by 2020 (Cooper (2016)).

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)).

To overcome these challenges, we lean on the second advance: a set of new techniques in machine learning. Machine learning methods are increasingly popular in economics and other social sciences. They have been used to predict poverty and wealth (Blumenstock, Cadamuro, and On (2015), Engstrom, Hersh, and Newhouse (2016), Jean et al. (2016)), improve municipal efficiency (Glaeser et al. (2016)), understand perceptions about urban safety (Naik, Raskar, and Hidalgo (2015)), improve judicial decisions to reduce crime (Kleinberg et al. (2017)), and more. We extend this literature to selection-on-unobservables designs using high-frequency panel data. In particular, we use LASSO, a form of regularized regression, to generate *school-specific* prediction models of electricity consumption while avoiding overfitting. We train these models on pre-treatment data only, and use them to forecast counterfactual energy consumption paths in the absence of any energy efficiency investments. Using machine learning enables us to create flexible, data-driven models of energy use without fear of overfitting. The central insight of our approach is that machine learning methods, designed to produce predictions, can be used to generate counterfactuals in panel data settings, which we can then embed in a selection-on-unobservables context to estimate causal effects. Because we perform the prediction school-by-school, our approach becomes empirically tractable. 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.[2]

In particular, 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. We use pre-treatment data only to develop a rich model of electricity consumption at each school, and use these models to forecast electricity consumption into the post-treatment period. Because these models are trained using only data prior to the energy efficiency upgrades themselves, we can compare the resulting predictions of energy use to actual consumption data to estimate treatment effects.

Comparing our machine learning approach to standard panel fixed-effect approaches yields two primary findings. First, we show that estimates from standard panel fixed effects approaches are quite sensitive to the set of observations included as controls as well as to the fixed effects included in the specification. Our machine learning method yields estimates that are substantially more stable across specifications. Second, and perhaps more importantly, we find that the panel fixed effects method performs poorly in an event study check and a series of placebo tests. Even with rich school-by-time-of-day and month-of-sample fixed effects, this approach appears to be prone to bias. In sharp contrast, we see no evidence of systematic bias when we subject our machine learning approach to the same event study and placebo tests. This suggests that our machine learning method has the potential to outperform standard approaches in settings with high-frequency data.

Beyond our methodological innovation, we make a policy-relevant contribution to the literature

2. 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. In contemporaneous work, Varian (2016) provides an overview of causal inference targeted at scholars familiar with machine learning. He proposes using machine learning techniques to predict counterfactuals in a conceptually similar manner, although he does not implement this approach in an empirical setting.

on energy efficiency. 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. While 37 percent of electricity use in the United States in 2014 was residential, over half is attributable to commercial and industrial uses such as schools (Energy Information Administration (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.

Using our new method, we find that energy efficiency investments can lead to substantial energy savings in schools, but the accuracy of *ex ante* engineering estimates of these savings vary significantly by the type of the investment. Across all types of investments, energy efficiency appears to deliver between 2.9 and 4.5 percent reductions in electricity use. We also look at the two most prevalent upgrade categories in our sample: heating, ventilation, and air conditioning (HVAC), which makes up 51 percent of upgrades in our sample, and lighting, which makes up 22 percent of upgrades. Investments in HVAC upgrades deliver between 3.2 and 4.6 percent reductions in energy use, while lighting upgrades deliver savings of between 3.6 and 5.2 percent of energy consumption on average. The majority of these savings come during daytime hours, with reductions up to 6 percent during the school day. These estimates translate into substantial savings, with the average upgrade reducing energy use by around 60 kWh per school per day.

When we compare our savings estimates to ex ante expectations, however, we find that upgrades are substantially underperforming relative to engineering projections. Across all upgrade categories, the average upgrade delivers only 24 percent of expected savings. HVAC and lighting upgrades perform somewhat better, achieving 49 and 42 percent of expected savings, respectively. We reject realization rates of 100 percent—realized savings in line with engineering estimates—across all upgrade categories, even when we trim the sample to exclude outliers and estimate realization rates with an alternative treatment indicator. In addition to estimating realization rates on a school-by-school basis, we also ask whether the average measured savings correspond to average ex ante engineering predictions, regardless of where these upgrades occurred. This exercise yields higher realization rates of 55, 103, and 68 percent of expected savings among all, HVAC, and lighting upgrades, respectively. This suggests that while individual engineering projections correlate poorly with actual savings at particular schools, the predictions do a better job of approximating reality on average across the sample, in part because there is substantial heterogeneity across schools. In light of this, we investigate whether information that is readily available to policymakers can be used to predict which schools will have higher realization rates. Beyond basic information on location, however, we are unable to identify covariates that correlate strongly with higher realization rates, suggesting that improving realization rates via targeting may prove challenging in this setting.

The remainder of this paper 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 3.1). Section 3.2 introduces our machine learning methodology, and presents the results. In Section 4, we compare our savings estimates to the *ex ante* engineering projections to calculate realization rates. Section 5 concludes.

2 Context and data

Existing engineering 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, meaning that targeted programs to encourage investment in energy efficiency may yield particularly high returns among these establishments. 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 (California Energy Commission (2017)). To put this in perspective, it represents about one-third of what California spent 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 over the same time period (Barbose et al. (2013)). Though our sample 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 in which to isolate the impacts of energy efficiency. School buildings are all engaged in relatively similar activities, are subject to the same wide-ranging 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 school 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, Kok, and Quigley (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) and Eichholtz, Kok, and Quigley (2013)).

There is also a large grey literature evaluating energy efficiency programs, mostly through regulatory proceedings. Recent evaluations of energy efficiency programs for commercial customers, such as schools, in California find that actual savings are around 50 percent of projected savings for many efiiciency investments (Itron (2017a)) and closer to 100 percent for lighting projects (Itron (2017b)). The methodologies in these studies combine process evaluation (e.g., verifying the number

of light bulbs that were actually replaced) with impact evaluation, although the latter do not use meter-level data and instead rely on site visits by engineers to improve the inputs to engineering simulations. In this paper, we implement one of the first quasi-experimental evaluations of energy efficiency in schools.

2.1 Data sources

We use data from several sources. In particular, 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."[3]

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 by the district, rather than individual school. In order to estimate the effect of energy efficiency investments on electricity consumption, we required a concordance between meters and schools. 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 by NAICS code.[4] 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.[5] 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[6], a technology family (e.g., "HVAC", "Lighting", "Food service technology"), the number of units installed, the installation date, the expected lifetime of the project, the engineering-estimate of expected annual kWh savings, the incremental measure cost, and the PG&E upgrade incentive received by the school.[7] Many schools undertake multiple

3. Robustness checks suggest that our results are similar if we aggregate to hourly blocks.

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

5. 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.

6. 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"

7. We have opted not to use the cost data as we were unable to obtain a consistent definition of the variables related to costs.

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 also 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 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.[8]

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.[9] 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 that we use throughout the paper for computational efficiency. 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.[10]

[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 also look different across an array of observable characteristics from schools that did not receive these upgrades (see the last four columns of Table 1. 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

8. Bond data are from EdSource (edsource.org).

9. 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.

10. Table 1 does not summarize projected savings, since all untreated schools have expected savings of zero.

with electricity usage, it is important that we consider selection into treatment as a possible threat to econometric identification in this setting.

3 Empirical strategy and results

In this section, we describe our empirical approach and present results. We begin with a standard panel fixed effects strategy. Despite including a rich set of fixed effects in all specifications, we demonstrate that this approach is highly sensitive to the set of controls that we include in our analysis. Furthermore, routine checks - an event study analysis and placebo tests - demonstrate that this approach is prone to bias. We proceed by developing a novel machine learning methodology, wherein we generate school-specific models of electricity consumption to construct counterfactual electricity use in the absence of energy efficiency upgrades. We demonstrate that this method is substantially less sensitive to specification than our regression analysis, and show evidence that this method outperforms the panel fixed effects approach in standard validity checks.

3.1 Panel fixed effects approach

The first step of our empirical analysis is to estimate the causal impact of energy efficiency upgrades on electricity consumption. In an ideal experiment, we would randomly assign upgrades to some schools and not to others. In the absence of such an experiment, we begin by turning to standard quasi-experimental methods. We are interested in estimating the following equation:

$$Y_{itb} = \beta D_{it} + \alpha_{i,t,b} + \varepsilon_{itb} \tag{3.1}$$

where Y_{itb} is the log of energy consumption in kWh at school i on date t during 3-hour-block b. Our treatment indicator, D_{it}, is the fraction of expected energy savings that school i has installed by date t.[11] The coefficient of interest, β, is interpreted as the impact of going from no upgrades to 100% of the average portfolio of energy efficiency investments. $\alpha_{i,t,b}$ represents a variety of possible fixed effects approaches. Because of the richness of our data, we are able to include large numbers of multi-dimensional fixed effects, which non-parametrically control for observable and unobservable characteristics. In our most parsimonious specification, we control for school and hour-block fixed effects, accounting for time-invariant characteristics at each school, and for aggregate patterns over hours of the day. We present results from several specifications with increasingly stringent controls. Our preferred specification includes school-by-hour-block fixed effects, to control for differential patterns of electricity consumption across schools, and month-of-sample fixed effects, to control for macroeconomic shocks or time trends in energy consumption. As a result, our econometric identification comes from within-school-by-hour-block and within-month-of-sample differences between treated and untreated schools. Finally, ε_{itb} is an error term. Across all specifications, we cluster our standard errors at the school level to account for arbitrary within-school correlation. We trim the sample to exclude observations below the 1st and above the 99th percentile of the dependent variable.

11. More concretely, suppose a treated school undergoes two upgrades with expected savings of 0.5 kWh each. The variable will begin at 0, turning to 0.5 after the first upgrade, and to 1 after the second upgrade. This indicator is always zero for untreated schools.

3.1.1 Results

Table 2 reports results from estimating Equation (3.1) using five different sets of fixed effects. We find that energy efficiency upgrades resulted in between 1.4 and 4.7 percent reductions in energy use on average. These results are highly sensitive to the set of fixed effects included in the regression. Using our preferred specification - Column (5) in Table 2, which includes school-by-hour-block and month-of-sample fixed effects, we find that energy efficiency upgrades caused a 1.6 percent decline in energy consumption at treated schools, although estimates with a slightly more parsimonious, though reasonable, set of fixed effects indicate savings were over twice as large.

We also separately estimate results for three classes of upgrade: any upgrade, which includes all upgrades in the sample; HVAC upgrades, compared against an untreated group of schools that underwent no upgrades during our sample; and lighting upgrades, compared against the same untreated group as the HVAC upgrades. The results for HVAC are very similar to the results for any upgrade. HVAC upgrades resulted in between 1.7 and 4.8 percent reductions in energy consumption. Our preferred estimate for HVAC is a 1.8 percent reduction. Lighting interventions perform somewhat better, with effect sizes ranging from 2.4 to 5.5 percent savings, with our preferred estimate suggesting reductions of 2.5 percent savings. These results are all precisely estimated; the estimates in Column (5) are statistically significant at the 10 percent level for any upgrade and HVAC upgrades, and at the 5 percent level for lighting.

[Table 2 about here; Figure 2 about here]

Figure 2 presents results from this same panel fixed effects approach, but now we allow the treatment effects to vary by hour-block. We present results from two specifications, analogous to Columns (1) and (5) in Table 2. The light blue lines include only school and hour-block fixed effects; the darker blue lines have school-by-hour-block and month-of-sample fixed effects. Panel A presents results for any upgrade; Panel B shows HVAC results; and Panel C shows results for lighting. As in Table 2, adding more fixed effects attenuates the estimates toward zero. Across any and HVAC upgrades, the estimated savings are relatively stable throughout the day, with the exception of 6 AM to 9 AM. The increase in these morning hours may reflect schools ramping up equipment that is powered down over night to a greater extent after the retrofits. Lighting estimates display a somewhat different pattern, with the largest reductions coming during the main hours of school operation in our preferred specification. Note that energy consumption in levels is highest during normal operating hours (often twice as high as during the evening), so these figures mask substantial differences in total energy savings across hours.[12] Across the three upgrade types, the treatment effect patterns in timing of savings are highly sensitive to specification. The specification with less stringent controls estimates larger savings in the evening for all upgrade categories. This may reflect differences in load shapes among treated and untreated schools. The estimates are more stable across hour-blocks of the day in the regression with additional controls.

3.1.2 Robustness

The identifying assumption for the panel fixed effects model is that conditional on the set of controls in the model, treatment is as-good-as-randomly assigned, or formally, that $\mathbb{E}[\varepsilon_{itb}|\mathbf{X}] = 0$. In our

12. Boomhower and Davis (2017) 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.

preferred specification, this means that after removing school-by-hour-block-specific and month-of-sample-specific effects, treated and untreated schools need to be trending similarly. While we can never prove that this assumption holds, we can perform some standard checks to assess the validity of this assumption in this context.

Event study We begin by providing graphical results from an event study regression. These results shed light on the ability of our panel fixed effects approach to adequately control for underlying differences between treated and untreated schools over time. Figure 3 displays the impacts of energy efficiency upgrades in the quarters before and after an upgrade takes place. The x-axis plots quarters before and after the upgrade, with the month of replacement normalized to zero. We plot point estimates and 95 percent confidence intervals from a regression with our preferred set of fixed effects: school-by-hour-block and month-of-sample:

$$Y_{itb} = \sum_{q=-6}^{10} \beta_q \mathbf{1}[\text{Quarter to upgrade} = q]_{it} + \alpha_{i,t,b} + \upsilon_{itb} \tag{3.2}$$

where $\mathbf{1}[\text{Quarter to upgrade} = q]_{it}$ is an indicator for relative time in the sample, such that $q = 0$ is the quarter of upgrade, $q - 6$ is 6 quarters prior to the upgrade, and $q + 10$ is 10 quarters after the upgrade, etcetera. We measure treatment effects relative to $q = 0$.

[Figure 3 about here]

The results from this event study analysis present problems: we do not see strong evidence that energy consumption is substantially reduced after the upgrades. Furthermore, we see strong evidence of seasonal patterns in the estimates, even after including month-of-sample fixed effects. This may be reflective of seasonality in upgrade timing, as many schools install upgrades during holiday periods only. This suggests that, even using our preferred specification, treatment and control schools' energy consumption is likely not directly comparable.

Placebo test We next perform a placebo test to evaluate the extent to which the panel fixed effects approach is appropriately controlling for time-varying unobservable characteristics which may threaten our identification strategy. To do this, we begin with the set of schools that never underwent an energy efficiency upgrade during our sample. We randomly assign 40 percent of them to a "treatment" group, and keep the rest as our "untreated" group, and randomly assign the "treatment" group a new treatment date, and generate an indicator equal to 0 for all "untreated" schools, and equal to 1 for "treated" schools after their "treatment date". We re-estimate Equation (3.1) using this placebo treatment indicator as our regressor of interest. We allow the treatment effect to vary across each of our eight hour-blocks. We estimate these "treatment effects" using five different sets of fixed effects corresponding to the columns in Table 2 above. We repeat this process 50 times, storing the hour-block-specific coefficients.

[Figure 4 about here]

Figure 4 presents the results of this exercise. Each panel corresponds to a column from Table 2. Each light gray lines show the hour-block-specific treatment effects from one placebo run. The

darker gray line shows the mean treatment effect for each hour-block. If our panel fixed effects model were appropriately accounting for within-day differences between treated and untreated schools, we would expect these lines to be flat—without intra-day patterns—and centered around zero. Instead, across all specifications, we see marked patterns, with the largest "reductions" in energy consumption occurring between 3 and 6 AM, and "increases" in energy consumption occurring between 3 and 6 PM. These patterns persist even when we turn to our most stringent specifications: Panel 4 displays results from a regression with school-by-hour-block-by-month-of-year fixed effects and month of sample controls; Panel 5 displays results using school-by-hour-block and month-of-year fixed effects. Even these specifications, which explicitly control for school-specific hourly patterns, estimate non-zero placebo treatment effects. This suggests that our panel fixed effects model is not adequately controlling for within-day consumption patterns that vary across schools, and may return spurious or biased treatment effects.

Matching Our panel fixed effects approach is highly sensitive to the set of fixed effects, and the results of our event study and placebo tests above suggest that even our preferred specification does not fully control for time-varying differences between treated and untreated schools. In an attempt to remedy these issues, we implement a nearest neighbor matching approach, with the goal of reducing selection bias that results from differences between treated and untreated schools. We match each treated school to exactly one untreated school, using the mean, maximum, and standard deviation of energy consumption in each hour-block prior to treatment, demographic variables measured at the census block level (including the poverty rate and income), school-level variables (including enrollment, age of the school, grades taught, an academic performance index), and climate as observable characteristics to match on.

Matching presents a challenge in the school setting because energy efficiency upgrade decisions are typically made at the district, rather than the school, level. This means that any matching approach where treated and untreated schools come from the same district will likely induce selection bias: there was a reason, unobservable to the econometrician, that the treated school underwent an energy efficiency upgrade and the untreated school did not. On the other hand, forcing across-district matches undermines the ability of the matching approach to reduce observable differences between schools.

[Table 3 about here]

With this caveat in mind, we present results from three types of matches: unrestricted, restricted to schools in the same district, and restricted to schools in other districts for any type of intervention. Table 3 displays the results of this approach. As above, the results are quite sensitive to specification. When we allow unrestricted matches, we estimate treatment effects ranging from 2.9 percent reductions (in our preferred specification) to 5.4 percent reductions. These results are even more sensitive to the match candidates. When we restrict matches to be within the same school district, the estimates range from 1.5 percent reductions to 1 percent *increases* in energy consumption (0.1 percent increases in our preferred specification). Forcing matches to come from opposite districts results in similar results to the unrestricted match, though these estimates are somewhat attenuated, ranging from 2.1 percent reductions (in our preferred specification) to 5.1 percent reductions. 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.

3.2 Machine learning approach

As demonstrated above, even with a large set of high-dimensional fixed effects, the standard panel approach performs poorly on basic validity tests. We now take advantage of the richness of our data to develop a novel machine learning method for causal inference in panel data settings.

3.2.1 Methodology

Machine learning is fundamentally about making predictions that perform well out-of-sample. The central insight of our approach is that we can use machine learning methods to generate a counterfactual: we predict what would have occurred in the absence of treatment. To do this, we use machine learning methods on pre-treatment data to build unit-specific models of an outcome of interest, use these models to generate predicted outcomes in the post-treatment period, and then compare this predicted outcome to the realized outcome to compute treatment effects. This approach is data-driven, highly flexible, and computationally feasible. While we apply our method in the context of energy efficiency upgrades, it could in principle be used in a wide variety of settings where researchers have access to rich panel data.

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 high 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 algorithmically trade off bias and variance to achieve out-of-sample performance. This enables the econometrician to select models from a much wider space than would be possible with trial-and-error.

This paper contributes to a small but rapidly growing economics literature on machine learning for causal inference.[13] The existing work in this area has focused on two areas. First, researchers have been using machine learning tools to estimate heterogeneous effects in randomized trials while minimizing concerns about "cherry-picking" (Athey and Imbens (2015)).[14] Second, and more closely related to our work, economists are leveraging machine learning to improve selection-on-observables designs. McCaffrey, Ridgeway, and Morral (2004) propose a method analogous to propensity score matching but using machine learning for model selection. Wyss et al. (2014) force covariate "balance" by directly including balancing constraints in the machine learning algorithm used to predict selection into treatment. Finally, Belloni, Chernozhukov, and Hansen (2014) propose 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.

13. Athey (2017) and Mullainathan and Spiess (2017) provide useful overviews.

14. These methods are incredibly useful when units are randomly assigned to treatment. In our context, however, non-random selection makes this method undesirable, as the partitioning itself may introduce greater selection bias.

We extend the literature on machine learning for causal inference to selection-on-unobservables designs. Rather than using machine learning to predict selection into treatment, we leverage untreated time periods in high-frequency panel data to create unit-specific predictions of the outcome of interest in the absence of treatment. We can then estimate treatment effects by comparing real outcomes to predicted outcomes between treated and untreated periods. This enables us to combine machine learning methods with a standard panel fixed approach, using within-unit within-time-period variation for identification.

Prediction We use machine learning to generate school-specific prediction models of electricity consumption at the hour-block level. In particular, we begin with pre-treatment data only. 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 of in-sample calendar dates.[15] Data after the treatment or pseudo-treatment date is set aside and not used for the purposes of generating the model.[16]

We use the Least Absolute Shrinkage and Selection Operator (LASSO), a form of regularized regression, to generate a model of energy consumption at each school.[17] We begin with a set of 3,000 possible covariates for each school-block separately, including day of the week, a holiday dummy, a seasonal spline, a temperature spline, and interactions between all of these variables.[18] The LASSO separates the pre-treatment data from one school at a time into "training" and "testing" sets. The algorithm finds the model with the best fit in the training data, and then tests the out-of-sample fit of this model in the testing set, thereby guarding against overfitting.[19] Ultimately, the algorithm returns a prediction model for each school. Generating school-specific models is important in our context, where there is a large degree of heterogeneity in energy consumption patterns across schools. Creating predictions on an individual school-by-hour-block basis allows our method to be extremely flexible, which helps to remove the types of bias displayed in the panel fixed effects approach above.

We examine the resulting prediction models to ensure that the LASSO is yielding useful outputs. We begin with the number of covariates in each model. The LASSO algorithm attempts to balance the number of explanatory variables and the prediction errors of the model, in a way that the optimal model for any given school will not include all of the candidate regressors. In fact, we find that the optimal models usually include between 50 and 100 variables. Importantly, however, the

15. 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.

16. 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.

17. We use LASSO in this context because of its existing popularity in economics, but adapting this method to instead use ridge, as recommended by Abadie and Kasy (2017) in certain settings, would be straightforward as it is nested in the `glmnet` library in *R* that we use in our implementation. Other prediction algorithms, such as random forests, could also be easily incorporated by modifying the prediction algorithm in the prediction step.

18. Because we are estimating school-block-specific models, each covariate is also essentially interacted with a school fixed effect and a block fixed effect—meaning that the full covariate space includes over 6,000,000 candidate variables. To make the approach computationally tractable, we estimate a LASSO model one school-block at a time.

19. This process is repeated several times, with a new randomly-selected "training" and "testing" dataset each time. We use the `glmnet` library default method, which uses a cross-validation procedure to select the tuning parameter of the LASSO.

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 for each school-block, showcasing the flexibility of selecting a different model for each school. Panel A of Figure 5 displays the relationship between the amount of training data and the number of non-zero coefficients in the prediction model at every school in the sample. Intuitively, the LASSO selects fewer covariates for schools with smaller training samples - this is indicative of the algorithm guarding against overfitting. As the training set gets larger, so too does the number of covariates, up to a point - there does appear to be diminishing returns from adding regressors in the prediction models.[20]

[Figure 5 about here]

We can also inspect the selected covariates themselves. Because we expect holidays to dramatically impact electricity consumption in K-12 schools, our holiday indicator provides a useful illustration of the results of the LASSO.[21] Panel B of Figure 5 shows the coefficient on the holiday dummy (and its interactions - we allow up to 50 per school) in each school-specific prediction model. The LASSO selected nearly 22,000 holiday variables across the more than 2,000 schools in our sample. In addition, the coefficients on these variables are overwhelmingly negative, which is in keeping with our ex ante predictions about the effects of holidays on energy consumption. This suggests that the LASSO-selected models reflect real-world electricity use.

The final output of the prediction step in our method 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.

Estimation Armed with predicted energy consumption at each school, we turn to the estimation of treatment effects. Our ultimate goal is to compare our models' predictions of energy consumption with real energy use. In the absence of other confounding factors, the difference between our predicted counterfactual energy consumption and our data on electricity use would be the causal impact of energy efficiency upgrades. Here, we present a series of estimators based on this idea, but designed to estimate treatment effects in the presence of time-varying changes in energy consumption.

We begin with a test of our method: we compute prediction errors—the average difference between the realized (log) energy consumption and its prediction—at *untreated* schools in the post-"treatment" period:[22]

$$\hat{\beta}^U = \frac{1}{(1-P)I}\frac{1}{r}\sum_{i=PI+1}^{I}\sum_{t=1}^{r}(y_{it} - \hat{y}_{it})$$

20. This is a good feature, as the LASSO algorithm works best in environments in which the true model is sparse, i.e., even if we had lots of data, the algorithm would not pick all candidate variables in the model.

21. We define "holidays" 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 any long spells of inactivity at the schools.

22. Recall that we assigned every untreated school a random "treatment" date. We use only pre-"treatment" data to train untreated schools' models and validate our predictions out of sample.

where there are I total units in the sample, and P is the proportion of treated units, such that the first PI units are treated and the remaining $(1-P)I$ are untreated; there are $m+r$ total time periods, split into $[-m+1, 0]$ pre-treatment periods and $(0, r]$ post-treatment periods; y_{it} is realized energy consumption in school i at time t, and $\hat{y}_{it}$ is predicted energy consumption.[23] If the model has good performance out of sample, $\hat{\beta}^U$ should be zero in expectation. Figure 6 displays the results of this estimator: our "treatment effect," an 0.2 percent increase in energy consumption, is statistically indistinguishable from zero and if anything, positive—the opposite sign from what we would expect an energy efficiency upgrade to deliver.[24]

We may be concerned about trends in energy use over time confounding this simple comparison. To test this, we can extend our estimator to compare prediction errors in the post-treatment period with prediction errors in the pre-treatment period:

$$\hat{\beta}^{UD} = \hat{\beta}^U - \frac{1}{(1-P)I}\frac{1}{m}\sum_{i=PI+1}^{I}\sum_{t=-m+1}^{0}(y_{it} - \hat{y}_{it})$$

Figure 6 shows that after controlling for changes over time, $\hat{\beta}^{UD}$ yields an 0.4 percent decline in energy consumption. We again fail to reject the null that the treatment effect is zero among untreated schools. These two tests provide suggestive evidence that our machine learning models are performing as expected.

[Figure 6 about here]

We can now leverage predicted energy consumption to estimate treatment effects at treated schools. We begin with the simplest estimator:

$$\hat{\beta}^T = \frac{1}{PI}\frac{1}{r}\sum_{i=1}^{PI}\sum_{t=1}^{r}(y_{it} - \hat{y}_{it}),$$

which is analogous to $\hat{\beta}^U$, but with treated rather than untreated schools. If energy efficiency upgrades deliver savings, $\hat{\beta}^T$ should be negative, as predicted energy use, generated without any knowledge of the upgrade, will overestimate actual energy consumption. This is exactly what we see in Figure 6: a treatment effect of -4.4 percent, statistically significant at the 1 percent level.

As with the untreated schools, we can also compare treated schools to themselves over time:

$$\hat{\beta}^{TD} = \hat{\beta}^T - \frac{1}{PI}\frac{1}{m}\sum_{i=1}^{PI}\sum_{t=-m+1}^{0}(y_{it} - \hat{y}_{it})$$

We again expect this to be negative, and it is: Figure 6 shows the treatment effect estimate of -5.0 percent, statistically significant at the 1 percent level.

To the extent that there are systematic differences between the prediction and the observed outcomes for untreated schools during the post period, and to the extent that these differences

23. To avoid clutter, we do not include the 'i' subscripts on m and r (and generally abstract from issues associated with unbalanced panels) although those parameters differ by school, as described above.

24. For this and all estimators below, we cluster our standard errors at the school level.

reflect trends and biases in the predictive model that are common across schools, we can use these differences as a bias correction for the treated schools by estimating:

$$\hat{\beta}^{PD} = \hat{\beta}^{T} - \hat{\beta}^{U},$$

which yields a post-differenced corrected treatment effect estimate of -4.7 percent (statistically significant at the 1 percent level and shown in Figure 6) under the assumption of common trends and shocks between treated and untreated schools.

We can also estimate a "triple difference" that exploits the differences in predictions between treated and untreated schools during the pre- and post-period, by taking the differences of the before and after estimators at treated and untreated schools:

$$\hat{\beta}^{DD} = \hat{\beta}^{TD} - \hat{\beta}^{UD}.$$

This difference will tend to provide very similar results to those only using post data, as the corrections using pre-treatment data are relatively small. Using this estimator, we find that energy efficiency upgrades caused a 4.9 percent reduction in energy consumption, significant at the 1 percent level, and shown in Figure 6. Note that this triple difference relies on the same identifying assumptions as the difference-in-difference estimator described in the regression methodology section above, namely, that conditional on covariates, treated and untreated schools are trending similarly. The key difference is that for this estimator to be identified, we need treated and untreated schools to be trending similarly in *prediction errors*, rather than in energy consumption.

Taken together, these results suggest that our machine learning method is delivering causal estimates of the impact of energy efficiency on electricity use. Estimates for untreated schools are very close to zero, as expected, and all of the estimators using treated schools find treatment effects between -4.4 and -5.0 percent.

3.2.2 Results

We now combine the machine learning approach with the same rich sets of fixed effects from our panel fixed effects approach. Table 4 displays the results from estimating Equation (3.1) with prediction errors for logged consumption in kWh as the dependent variable.[25] As in Table 2 which presents the panel fixed effects approach above, we show treatment effect estimates for five different specifications and three different intervention types. Using the machine learning approach, we find that upgrades reduce energy consumption by between 2.9 and 4.5 percent on average. In our preferred specification, using school-by-hour-block and month-of-sample fixed effects, we estimate a reduction of 3.1 percent. The estimates appear to be somewhat less sensitive to different controls than the panel fixed effects estimates above. HVAC upgrades reduce energy consumption by between 3.2 and 4.6 percent; using our preferred specification, we estimate savings of 3.2 percent. Finally, lighting interventions save between 3.6 and 5.2 percent. Our preferred specification yields a treatment effect of 4.1 percent for lighting interventions. These estimates are tightly estimated;

25. Because the regression is in logs, the prediction model is for log consumption, as opposed to consumption in levels.

every point estimate in Table 4 is statistically significant at the 1 percent level.[26]

[Table 4 and Figure 7 about here]

Figure 7 displays hour-block-specific results from the machine learning approach. Panel A shows results for any upgrade; panel B shows results for HVAC upgrades only; and panel C shows results for lighting upgrades only. The lighter blue lines present results with school and hour-block fixed effects only, as in Column (1) of Table 4. The darker blue lines present results from our preferred specification, as in Column (5) of Table 4, and uses school-by-hour-block and month-of-sample fixed effects. In contrast with the panel fixed effects results from Figure 2, the machine learning estimates are much less sensitive to specification. Not only are the overall treatment effect estimates from the two specifications remarkably similar, but the treatment effects also appear to follow the same within-day pattern regardless of specification. Furthermore, whereas the panel fixed effects approach showed the largest treatment effects during the evening, the machine learning estimates show the largest savings during peak hours of school operation: 9 AM to 3 PM - also among the most valuable hours for reductions as discussed above. This pattern holds across all three types of upgrades. For any upgrade and HVAC upgrades, we again see some evidence of morning ramping, as the 6 AM to 9 AM block shows the smallest savings. Overall, these estimates appear to be consistent with patterns of energy savings that we would have expected from these interventions.

3.2.3 Robustness

As with the standard panel fixed effects approach, the identifying assumption is that, conditional on controls treatment is as-good-as-randomly assigned, or formally, that $\mathbb{E}[\varepsilon_{itb}|\mathbf{X}] = 0$. In our preferred specification, this means that after removing school-by-hour-block-specific and month-of-sample-specific effects, treated and untreated schools need to be trending similarly *in prediction errors*, as our dependent variable is now the difference between predicted and actual log energy consumption. This is analogous to having included a much richer set of control variables on the right-hand side of our regression. In a sense, the machine learning methodology enables us to run a much more flexible model in a parsimonious and computationally tractable way. Below, we subject the machine learning approach to the same checks as the panel fixed effects method and find that it performs substantially better.

Event study We again begin with graphical evidence from an event study regression of log prediction errors on indicator variables for quarters relative to treatment. Figure 8 displays the point estimates and 95 percent confidence intervals from estimating Equation (3.2) with log prediction errors as the dependent variable, and including school-by-hour-block and month-of-sample fixed effects, as in Column (5) of Table 4. We normalize the quarter of treatment to be zero for all schools.

[Figure 8 about here]

26. We cluster our standard errors at the school level to account for arbitrary within-school correlation. Because we care about the expectation of the prediction, rather than the prediction itself, our standard errors are unlikely to be substantially underestimated by failing to explicitly account for our forecasted dependent variable.

Figure 8 shows no statistically significant treatment effects in the 6 quarters prior to an energy efficiency upgrade. Unlike in Figure 3, the point estimates do not exhibit strong seasonal patterns. Furthermore, after the energy efficiency upgrades occur in quarter 0, we see a marked shift downwards in energy consumption. This treatment effect, of an approximately 3 percent reduction in energy use, is stable and persists up to 10 quarters after the upgrade occurs. This event study figure provides strong evidence to suggest that the machine learning approach - unlike the panel fixed effects approach above - is properly specified, and effectively controlling for time-varying observable and unobservable differences between treated and untreated schools.

Placebo test We perform the same placebo exercise as in Section 3.1.2 to evaluate the effectiveness of the machine learning approach in capturing within-day patterns in energy consumption. We again begin with untreated schools only; assign 40 percent to a fake treatment status, with a treatment effect size of 0. We also generate a fake treatment indicator, which turns on only in the machine learning post-training period for "treated" schools. We re-estimate (3.1) with this fake treatment indicator as the independent variable of interest, allowing the effect to vary across each of our eight hour-blocks. We estimate our "treatment effects" using the five fixed effects specifications in Table 4, and repeat the process 50 times.

[Figure 9 about here]

Figure 9 shows the results of this exercise, with each numbered panel corresponding to a column from Table 4. Each gray line depicts the results from one placebo draw; the thick gray line shows the mean of the point estimates across each hour-block. If the model is properly specified, we expect to see flat lines centered at zero. Unlike in the panel fixed effects approach, we find that the estimated treatment effects are close to zero on average, amd display only minimal within-day patterns. These results suggest that our machine learning method is less prone than the panel fixed effects approach to generating spurious treatment effect estimates.

4 Policy implications

Using our machine learning approach, we find that energy efficiency upgrades reduce school energy consumption by between 2.9 and 4.5 percent on average.[27] These savings represent real reductions in energy consumption which in turn translate into financial savings among schools. At the same time, it is important to compare these savings to the *ex ante* engineering projections, which schools used to make their investment decisions. This is particularly important because we observe a wide set of energy efficiency upgrades in our sample, ranging from replacing a few fixtures, to upgrading a whole HVAC system. In this section, we estimate a treatment effect that is proportional to projected savings information that was available *ex ante*. This allows us to compare across heterogeneous interventions, and to properly contextualize and interpret our savings estimates.

27. The panel fixed effects approach yields somewhat smaller estimates, ranging from 1.4 percent to 4.3 percent.

4.1 Comparing realized and expected savings

To compare the realized savings from energy efficiency upgrades to *ex ante* expected savings, we estimate eequations of the following form:

$$Y_{itb} = -\gamma S_{it} + \alpha_{i,t,b} + \epsilon_{itb}, \tag{4.1}$$

where Y is electricity consumption (in kWh) for the panel fixed effects approach, and prediction error (in kWh) for the machine learning approach.[28] S_{it} is the cumulative expected savings installed at school i by time t, normalized to be in units of kWh, which we compute for each upgrade from annualized engineering estimates.[29] As above, $\alpha_{i,t,b}$ represents a flexible set of fixed effects, ranging from school and hour-block fixed effects only to school-by-hour-block and month-of-sample fixed effects.

The coefficient of interest is γ. Note that these regressions are in levels, as opposed to logs, so that a coefficient of $\gamma = 1$ can be interpreted as *ex post* realized savings matching, on average, *ex ante* estimated energy savings at the school level. A coefficient larger than one would suggest that the observed energy savings are larger than those predicted *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 5 about here, Figure 10 about here]

Table 5 presents the results for all interventions, HVAC interventions, and lighting interventions for the same five fixed effects specifications shown above. 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. Figure 10 displays the results from Column (5) of Table 5, which includes school-by-hour-block and month-of-sample fixed effects, graphically. We find that, across all specifications and upgrade types, realized savings are substantially lower than expected savings. The panel fixed approach yields realization rate estimates between 22 and 29 percent across all upgrades; using machine learning, we estimate realization rates between 24 percent (our preferred specification) and 30 percent. Our realization rate estimates are somewhat higher for HVAC and lighting upgrades, with machine learning estimates ranging from 46 percent to 56 percent and 42 to 51 percent, respectively. All of the realization rates in Table 5 are statistically different from 100 percent at conventional levels.

Energy efficiency policy discussions sometimes distinguish between "net" and "gross" savings, where the former excludes inframarginal energy efficiency investments that customers would have made even in the absence of an energy efficiency subsidy program. Because the machine learning approach provides school-specific counterfactuals, it allows us to disentangle the extent to which untreated schools in our sample (i.e., schools who are not receiving rebates from the utility) are also reducing their consumption over time. As shown in Figure 6 above, we estimate zero savings on average among our untreated group. This is reassuring, as it suggests that the low realized

28. Because the prediction here is for consumption in levels, as opposed to logs, we re-run the LASSO model to improve fit and preserve the linearity of the error term in the prediction.

29. For a concrete example, suppose that a school undergoes two upgrades, one that is expected to save 25 kWh, and the other expected to save 75 kWh when normalized to the hourly level. S_{it} will be equal to zero before the first upgrade, to 25 after the first upgrade, and to 100 after the second upgrade.

savings are not driven by unmeasured efficiency upgrades at untreated schools, and are instead likely driven by overly optimistic *ex ante* predictions or rebound.

4.2 Sensitivity to measurement error

Another possible explanation for our low realization rates is measurement error in our expected savings data. There are many ways in which energy efficiency upgrade data can fail to reflect on-the-ground realities. First, a plausible dimension of mismeasurement is in upgrade dates. Our dataset ostensibly reports the date an energy efficiency upgrade was installed, but if some schools instead reported the date an upgrade was initiated or paid for, if an upgrade spanned multiple days, or if there is simply enumeration error, by treating all "installation dates" in our sample as treatment dates, we may estimate realization rates that are biased towards zero. A second important dimension of measurement error is in expected savings. Our own estimates, as well as other recent work (e.g. Fowlie, Greenstone, and Wolfram (forthcoming)), suggest that *ex ante* engineering estimates do not accurately reflect real-world savings. Even if these estimates accurately described the impacts of a given intervention *on average*, the engineering projections could easily fail to account for school-specific characteristics impacting the performance of the energy efficiency upgrades in question. Even if such errors were to cancel on average across schools, this type of mismatch could lead us to estimate a relatively low realization rate when comparing realized savings to estimated savings on a school-by-school basis.

The vast heterogeneity across measures also can complicate estimating the effectiveness of these energy efficiency interventions. Figure 11 shows the distribution of expected savings relative to average energy consumption among the treated schools in our sample. The majority of interventions are expected to save less than 10 percent of average energy consumption, but there remains a substantial right tail, with some schools expected to reduce energy consumption by an unrealistically large amount. These data points could result from measurement error in expected savings, or from a mismatch between schools and interventions.[30]

[Figure 11 about here]

Including the schools at the right tail of the distribution in Figure 11 will lead us to estimate small realization rates. For example, if a school installs an upgrade that is projected to reduce its energy consumption by 100 percent, but in reality, this upgrade only reduces consumption by 5 percent, we will estimate a realization rate of 5 percent. If this same upgrade were instead only expected to reduce consumption by 10 percent, we would instead estimate a realization rate of 50 percent. This highlights the potential consequences of measurement error in expected savings on our realization rate estimates.

In light of the potential for measurement error in this context, we test the sensitivity of our realization rate estimates to a variety of sample trimming approaches. Table 6 displays the results of this exercise. Column (1) replicates the results from column (5) in Table 5, in which we regress either electricity consumption or prediction errors on cumulative expected savings in kWh. In this baseline specification, we trim the sample such that the observations below the 1st and above

30. The true right tail of this distribution is even longer. For graphical purposes, we remove the 15 schools for which expected savings exceed 100 percent of average energy consumption from Figure 11.

the 99th percentile of the dependent variable are excluded. Column (2) explores the relevance of measurement error in the timing of interventions by replacing our continuous treatment indicator with a variable that is equal to zero for all pre-treatment periods, and turns to the average expected savings over the post-treatment period as soon as a school installs its first energy efficiency upgrade. Using this alternative treatment indicator, we find substantially large realization rates: 44, 58, and 58 percent for any, HVAC, and lighting upgrades with the machine learning approach. This suggests that measurement error in the timing of treatment, or in the relative size of each intervention could be attenuating our earlier realization rate estimates. We next use this same "binary" treatment indicator, but now trim the sample to exclude observations outside the 1st and 99th percentile in *expected savings* (Column (3)), or trim on both expected savings and the dependent variable (Column (4)). We estimate somewhat larger realization rates in Column (3), though our estimates in Column (4) are similar to those in Column (2). In every case, the point estimates remain below 1, and we can statistically reject 100 percent realization rates in all cases.

[Table 6 about here]

Equation (4.1) generates estimates of the average relationship between *ex ante* expected savings at a particular school and that same school's realized savings. In the presence of substantial measurement error, or in cases where school-specific idiosyncrasies cause substantial mismatch between individual estimates of projected savings and realized savings, this estimator will result in relatively low realization rates. In Column (5) of Table 6, we present an alternative calculation. Here, we generate an estimate of the average expected savings in levels in the sample by estimating Equation (3.1) with energy consumption or prediction errors in levels as the dependent variable.[31] We divide the implied average treatment effect from this regression by the average expected savings across energy efficiency upgrades in the sample (weighted to be representative of the estimation sample from the regression). This estimator essentially asks: do the average savings that we can measure in the data correspond to the average *ex ante* engineering estimates, irrespective of which school undertook each upgrade?[32] Using this alternative estimator, we find average realization rates of 55 percent for any intervention, 103 percent for HVAC interventions, and 68 percent for lighting interventions using the machine learning approach.[33] These results underscore the importance of accounting for measurement error and heterogeneity in evaluations of energy efficiency upgrades. While Table 6 suggests that measurement issues are partially responsible for the low realization rates we estimate in our baseline specification, these challenges do not appear to explain the full gap between the *ex ante* estimates of expected savings and realized savings in our context.

31. We use a binary treatment indicator here, which generates an estimate of average realized savings in kWh.

32. We must exercise caution when trying to compare the results from this estimator to estimates from Equation (4.1). The ratio of average savings need not be equal to the estimate of the expected ratio, due to Jensen's inequality. If the savings realization rate were homogeneous across projects, then the two estimates would be equivalent. If realization rates were larger (smaller) for larger projects, then the realization rate using this method should be larger (smaller). In our case, larger projects tend to be correlated with lower realization rates, thus potentially making this ratio estimator smaller. The presence of potential measurement error complicates the comparative statics further, but goes in the opposite direction, contributing to making this estimator larger than the regression approach, which could be attenuated.

33. The panel fixed effects approach is quite sensitive to our chosen set of fixed effects. For the specification we report here, including school-by-hour-block and month-of-sample fixed effects, we find larger realization rates in Column (5) than Column (1) for HVAC and lighting interventions, but substantially lower realization rates for any upgrade.

4.3 Realization rate heterogeneity

The previous results suggest that heterogeneity in realization rates across schools can influence our estimates of realization rates. Given the richness of our electricity consumption data, we can take an additional approach and estimate treatment effects for each school separately.[34] To compute these school-specific estimates, we regress prediction errors in kWh on a school-specific dummy variable, which turns to one during the post-treatment period (or, for untreated schools, the post-training period from the machine learning model). The resulting estimates represent the difference between pre- and post-treatment energy consumption at each school individually. We can then use these school-specific estimates to understand the distribution of treatment effects, and try to unveil potential systematic patterns across schools.

Panel A of Figure 12 displays the relationship between these school-specific savings estimates and expected savings for treated schools. We find a positive correlation between estimated savings and expected savings, consistent with the findings in Table 5, although there is substantial noise in the school-specific estimates. Once we trim outliers in expected savings, we recover a slope of 43 percent, consistent with the findings in Table 6. Panel B presents a comparison of the school-specific effects between treated and untreated schools. The estimates at untreated schools are much more tightly centered around zero, which helps validate our methodology, and suggests that there may be meaningful information in the school-specific estimates. In contrast, the distribution of treated school estimates is shifted towards additional savings, consistent with schools having saved energy as a result of their energy efficiency upgrades. These results suggest that energy efficiency projects successfully deliver savings, although the relationship between the savings that we can measure and the *ex ante* predicted savings is noisy.

[Figure 12 about here]

In light of our relatively low realization rates, we next try to project these school-specific estimates onto information that is readily available to policymakers, in an attempt to find predictors of higher realization rates. We do this by regressing our school-specific treatment effects onto a variety of covariates via quantile regression, in order to remove the undue influence of outliers in these noisy estimates.[35] We include one observation per treated school in our sample, and weight the observations by the length of the time series of energy data for each school. All variables are centered around their mean and normalized by their standard deviation, such that we can interpret the constant of this regression as the median realization rate.

[Table 7 about here]

Table 7 presents the results of this exercise. Column (1) shows that the median realization rate for treated schools using this approach is around 50 percent, consistent with the aggregate

34. The identification assumptions to obtain school-specific treatment effects are much stronger than when obtaining average treatment effects, as changes in consumption at the school level will be confounded with the treatment effect. Therefore, these estimates should not be taken as a precise causal measure of savings at any given school, but rather as a first step that allows us to then project heterogeneous estimates onto school-specific covariates for descriptive purposes.

35. Ideally, we would also use a quantile regression approach in our high-frequency data, which would assuage potential concerns about outliers. Because we rely on a large set of high-dimensional fixed effects for identification, however, this is computationally intractable.

estimates in Tables 5-6. Column (2) shows that average realization rates are somewhat larger for HVAC and lighting interventions, although this difference is not statistically significant. Latitude and longitude appear to be associated with larger realization rate in column (3). This effect could be driven by many potential factors correlated with location, such as weather, demographics, etc. Columns (4)-(5) control for some additional covariates, such as total enrollment, the Academic Performance Index and poverty rate, although these do not appear to significantly explain the performance of the measures.[36]

The lack of correlation between individual school-specific realization rates and the covariates suggests that, in our data, it is difficult to identify which schools are most likely to delivering savings in line with *ex ante* expectations. This implies that uncovering "low-hanging fruit" may be challenging, and, furthermore, that improving the success of the types of energy efficiency upgrades in our sample via targeting will likely prove difficult. That said, uncovering patterns in the heterogeneity of realization rates is particularly challenging in our setting. First, our sample of treated schools is relatively small—there are just over 1,000 observations in these quantile regressions, and each of the schools is subject to its own idiosyncrasies, leading to concerns about collinearity and omitted variables bias. Second, savings rates are noisily measured in our application, both because the numerator (realized savings) is a noisy estimate, and the denominator (expected savings) is also likely measured with error, deteriorating the signal-to-noise ratio and introducing potential bias. It is possible that in samples with more homogeneous energy efficiency projects, and with a larger pool of treated units, one could identify covariates that predict higher realization rates. This in turn could be used to inform targeting procedures to improve average performance.

5 Conclusions

In this paper, we leverage high-frequency data on electricity consumption and develop a novel machine learning method to estimate the causal effect of energy efficiency upgrades at K-12 schools in California. In our new machine learning approach, we use untreated time periods in high-frequency panel data to generate school-specific prediction of energy consumption that would have occurred in the absence of treatment, and then compare these predictions to observed energy consumption to estimate treatment effects.

We document three main findings. First, we show that our machine learning approach outperforms standard panel fixed effects approaches on two important dimensions: first, the machine learning treatment effect estimates are less sensitive to specification choice than the panel fixed effects estimates, which vary dramatically with the set of included control variables and observations; and second, the panel fixed effects approach appears prone to bias in a graphical event study analysis and a placebo test, while the machine learning method does not exhibit these biases in these standard checks. Our approach is general, and can be applied to a broad class of econometric settings where researchers have access to relatively high-frequency panel data. Second, using this approach, we find that energy efficiency investments reduced energy consumption by between 2.9 and 4.5 percent on average, with somewhat higher savings for HVAC and lighting investments, which deliver reductions between 3.2 and 4.6 percent and 3.6 and 5.2 percent, respectively. These

36. We explored a variety of other potential demographic variables, but we did not find any clear correlation with realization rates.

savings appear be driven by hours when school is in session, though we do find evidence of smaller savings during evening and night hours. These savings represent real reductions in energy use, on the order of 60 kWh per school per day, which translate into financial savings for schools which may be facing tight budget constraints. Finally, when we compare our estimated savings to *ex ante* engineering projections, we find low realization rates. Overall, upgrades delivered approximately 24 percent of expected savings. For HVAC and lighting upgrades, realized savings are closer to *ex ante* expectations, with realization rates of 49 and 42 percent, respectively. In all cases, we can reject realization rates of 100 percent—with energy efficiency investments generally delivering less than half of what was expected.

We test the extent to which measurement error and heterogeneity drive our conclusions. While we find evidence of significant mismeasurement in expected savings—for example, schools whose expected savings are greater than 100 percent of their average energy consumption—even with a variety of sample trimming approaches and alternative treatment indicators designed to account for these issues, we still find realization rates substantially below 100 percent. In our most optimistic approach, we find realization rates of 65 percent for all upgrades, 71 percent for HVAC upgrades, and 40 percent for lighting upgrades. Yet, we still reject realization rates of 100 percent, suggesting that measurement error is not solely responsible for our low realization rate estimates. We also find evidence of heterogeneous realization rates across schools. When we compare average realized savings to average expected savings in the sample, regardless of where these upgrades took place, we find substantially higher realization rate estimates: 55 percent across all upgrades, 103 percent for HVAC upgrades, and 68 percent for lighting upgrades. This suggests that heterogeneity in realization rates across schools is important. Using school-specific treatment effect estimates, we find a noisy relationship between expected savings and estimated savings, with evidence of heterogeneity across schools. Given this heterogeneity, we attempt to use information that is readily available to policymakers to predict which schools will have higher realization rates. Beyond basic information on location, we are unable to identify school characteristics that strongly predict higher realization rates. This suggests that without collecting additional data, improving realization rates via targeting may prove challenging.

This paper represents an important extension of the energy efficiency literature to a non-residential sector. We demonstrate that, in keeping with evidence from residential applications, energy efficiency upgrades deliver substantially lower savings than expected *ex ante*. These results have implications for policymakers and building managers deciding over a range of capital investments, and demonstrates the importance of real-world, *ex post* program evaluation to determine the effectiveness of energy efficiency. Beyond energy efficiency applications, our machine learning method provides a way for researchers to estimate causal treatment effects in high-frequency panel data settings, hopefully opening avenues for future research on a variety of topics that are of interest to applied microeconomists.

References

Abadie, Alberto, and Maximilian Kasy. 2017. "The Risk of Machine Learning." *Working paper.*

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

———. 2017. *Measuring the Welfare Effects of Residential Energy Efficiency Programs.* Technical report. National Bureau of Economic Research Working Paper No. 23386.

Athey, Susan. 2017. "Beyond prediction: Using big data for policy problems." *Science* 355 (6324): 483–485.

Athey, Susan, and Guido Imbens. 2015. "Recursive Partitioning for Heterogeneous Causal Effects." *Proceedings of the National Academy of Sciences* 113 (27): 7353–7360.

Barbose, Galen L., Charles A. Goldman, Ian M. Hoffman, and Megan A. Billingsley. 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.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–650.

Blumenstock, Joshua, Gabriel Cadamuro, and Robert On. 2015. "Predicting Poverty and Wealth from Mobile Phone Metadata." *Science* 350:1073–1076.

Boomhower, Judson, and Lucas Davis. 2017. "Do Energy Efficiency Investments Deliver at the Right Time?" National Bureau of Economic Research Working Paper No. 23097.

California Energy Commission. 2017. *Proposition 39: California Clean Energy Jobs Act, K-12 Program and Energy Conservation Assistance Act 2015-2016 Progress Report.* Technical report.

Cicala, Steve. 2017. "Imperfect Markets versus Imperfect Regulation in U.S. Electricity Generation." National Bureau of Economic Research Working Paper No. 23053.

Cooper, Adam. 2016. *Electric Company Smart Meter Deployments: Foundation for a Smart Grid.* Technical report. Institute for Electric Innovation.

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

Eichholtz, Piet, Nils Kok, and John M. Quigley. 2013. "The Economics of Green Building." *Review of Economics and Statistics* 95 (1): 50–63.

Energy Information Administration. 2015. *Electric Power Monthly.* Technical report.

Engstrom, Ryan, Jonathan Hersh, and David Newhouse. 2016. "Poverty in HD: What Does High Resolution Satellite Imagery Reveal about Economic Welfare?" *Working Paper.*

Fisher, Anthony, Michael Haneman, Michael Roberts, and Wolfram Schlenker. 2012. "The Economic Impacts of Climate Change: Evidence from Agricultural Output and Random Fluctuations in Weather: Comment." *American Economic Review* 102 (7): 1749–1760.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram. Forthcoming. "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program." *Quarterly Journal of Economics.*

Gerarden, Todd D, Richard G Newell, and Robert N Stavins. 2015. *Assessing the Energy-Efficiency Gap.* Technical report. Harvard Environmental Economics Program.

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

Glaeser, Edward, Andrew Hillis, Scott Duke Kominers, and Michael Luca. 2016. "Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy." *American Economic Review: Papers & Proceedings* 106 (5): 114–118.

International Energy Agency. 2015. *World Energy Outlook.* Technical report.

Itron. 2017a. *2015 Custom Impact Evaluation Industrial, Agricultural, and Large Commercial: Final Report.* Technical report.

———. 2017b. *2015 Nonresidential ESPI Deemed Lighting Impact Evaluation: Final Report.* Technical report.

Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353:790–794.

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

Kahn, Matthew, Nils Kok, and John Quigley. 2014. "Carbon emissions from the commercial building sector: The role of climate, quality, and incentives." *Journal of Public Economics* 113:1–12.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2017. "Human Decisions and Machine Predictions." *Working Paper.*

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

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

Levinson, Arik. 2016. "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." *American Economic Review* 106 (10): 2867–2894.

McCaffrey, Daniel, Greg Ridgeway, and Andrew Morral. 2004. "Propensity Score Estimation with Boosted Regression for Evaluating Causal Effects in Observational Studies." *RAND Journal of Economics* 9 (4): 403–425.

McKinsey & Company. 2009. *Unlocking energy efficiency in the U.S. economy.* Technical report. McKinsey Global Energy and Materials.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives* 31 (2): 87–106.

Myers, Erica. 2015. "Asymmetric information in residential rental markets: implications for the energy efficiency gap." *Working Paper.*

Naik, Nikhil, Ramesh Raskar, and Cesar Hidalgo. 2015. "Cities Are Physical Too: Using Computer Vision to Measure the Quality and Impact of Urban Appearance." *American Economic Review: Papers & Proceedings* 106 (5): 128–132.

Varian, Hal R. 2016. "Causal inference in economics and marketing." *Proceedings of the National Academy of Sciences* 113 (27): 7310–7315.

Wyss, Richard, Alan Ellis, Alan Brookhart, Cynthia Girman, Michele Funk, Robert LoCasale, and Til Sturmer. 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.

Table 1: Average characteristics of schools in the sample

Category	Untreated	Any intervention		HVAC interventions		Lighting interventions	
		Treated	T-U	Treated	T-U	Treated	T-U
Hourly energy consumption (kWh)	34.16	57.05	22.89	63.52	29.36	61.63	27.46
	(35.91)	(72.23)	[<0.01]	(81.94)	[<0.01]	(86.58)	[<0.01]
First year in sample	2012	2010	-2	2009	-2	2010	-2
	(1.71)	(1.87)	[<0.01]	(1.71)	[<0.01]	(1.86)	[<0.01]
Total enrollment	553.49	722.27	168.78	765.94	212.45	747.70	194.21
	(386.43)	(483.25)	[<0.01]	(507.95)	[<0.01]	(522.88)	[<0.01]
Academic perf. index (200-1000)	789.41	793.74	4.33	792.78	3.37	786.50	-2.91
	(98.52)	(89.65)	[0.31]	(89.30)	[0.49]	(80.17)	[0.57]
Bond passed within 2 yrs (0/1)	0.30	0.31	0.00	0.30	-0.05	0.28	-0.07
	(0.46)	(0.46)	[0.83]	(0.46)	[0.05]	(0.45)	[0.01]
Bond passed within 5 yrs (0/1)	0.39	0.40	0.01	0.41	-0.04	0.39	-0.06
	(0.49)	(0.49)	[0.66]	(0.49)	[0.10]	(0.49)	[0.03]
High school graduates (percent)	23.51	23.29	-0.22	23.65	0.15	24.08	0.57
	(12.35)	(11.70)	[0.69]	(11.67)	[0.81]	(10.92)	[0.39]
College graduates (percent)	20.08	20.29	0.21	19.45	-0.63	19.72	-0.36
	(12.34)	(11.97)	[0.70]	(11.84)	[0.32]	(11.90)	[0.59]
Single mothers (percent)	20.11	19.39	-0.73	19.99	-0.12	19.88	-0.23
	(19.06)	(18.19)	[0.39]	(18.47)	[0.90]	(18.35)	[0.83]
African American (percent)	5.73	6.24	0.51	5.47	-0.26	5.87	0.14
	(9.35)	(8.51)	[0.20]	(6.29)	[0.55]	(7.05)	[0.77]
Asian (percent)	9.46	11.34	1.89	12.27	2.81	9.60	0.14
	(13.70)	(15.96)	[<0.01]	(16.83)	[<0.01]	(12.48)	[0.85]
Hispanic (percent)	41.99	43.68	1.68	45.60	3.60	46.22	4.22
	(28.58)	(26.94)	[0.18]	(27.07)	[0.01]	(25.61)	[0.01]
White (percent)	34.31	30.71	-3.60	29.83	-4.48	29.73	-4.58
	(26.73)	(24.40)	[<0.01]	(23.79)	[<0.01]	(24.40)	[<0.01]
Average temperature (° F)	60.18	60.71	0.53	61.17	0.99	60.92	0.74
	(4.59)	(3.75)	[<0.01]	(3.63)	[<0.01]	(3.87)	[<0.01]
Latitude	37.68	37.47	-0.21	37.37	-0.31	37.43	-0.25
	(1.13)	(1.03)	[<0.01]	(1.01)	[<0.01]	(1.07)	[<0.01]
Longitude	-121.61	-121.25	0.36	-121.05	0.56	-121.19	0.42
	(1.04)	(1.11)	[<0.01]	(1.14)	[<0.01]	(1.12)	[<0.01]
Number of schools	1055	1039		628		493	

Notes: This table displays average characteristics of the treated and untreated schools in our sample, by type of intervention. Standard deviations are in parentheses, with p-values of the difference between treated and untreated schools in brackets. "Untreated" schools underwent no energy efficiency upgrades for the duration of our sample. Schools in the "Any," "HVAC," and "Lighting" categories had at least one intervention in the respective category installed during the sample period. In all cases, the "T-U" column compares treated schools to the schools that installed zero upgrades. Each row is a separate calculation, and is not conditional on the other variables reported here. There is substantial evidence of selection into treatment: treated schools tend to consume more electricity; have been in the sample longer; are larger; are in areas with more minorities; are in hotter locations; and are further to the north and east.

Table 2: Panel fixed effects results by type of intervention

	(1)	(2)	(3)	(4)	(5)
Any intervention:	-0.043	-0.043	-0.047	-0.014	-0.016
	(0.007)	(0.007)	(0.007)	(0.008)	(0.008)
Observations	19,113,434	19,113,434	19,112,574	19,112,574	19,113,434
HVAC interventions:	-0.043	-0.042	-0.048	-0.017	-0.018
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
Observations	14,840,990	14,840,990	14,840,266	14,840,266	14,840,990
Lighting interventions:	-0.052	-0.051	-0.055	-0.024	-0.025
	(0.009)	(0.009)	(0.009)	(0.011)	(0.011)
Observations	12,849,542	12,849,541	12,848,929	12,848,929	12,849,541
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	Yes	No
Month of Sample Ctrl.	No	No	No	Yes	No
Month of Sample FE	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with the log of hourly energy consumption (averaged across "blocks" of three hours) as the dependent variable. The independent variable is defined as the fraction of total expected savings that have been installed by time t, and takes on values from 0 to 1 in the treated schools, and 0 always in the control schools. A coefficient of -0.045, for example, means that going from 0 to 100% of a school's expected savings delivers energy savings of approximately 4.5% on average. Standard errors, clustered at the school level, are in parentheses. All samples are trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. Regressions for HVAC and light interventions include only schools with an intervention of this type and pure untreated schools that never underwent energy efficiency interventions of any kind during the sample period.

Table 3: Matching results – any intervention

	(1)	(2)	(3)	(4)	(5)
Any district:	-0.050	-0.051	-0.054	-0.030	-0.029
	(0.012)	(0.012)	(0.013)	(0.013)	(0.012)
Same district:	-0.015	-0.015	-0.010	0.010	0.001
	(0.014)	(0.014)	(0.015)	(0.015)	(0.014)
Opposite district:	-0.047	-0.047	-0.051	-0.026	-0.021
	(0.012)	(0.012)	(0.013)	(0.013)	(0.012)
Observations	4,828,122	4,828,108	4,826,977	4,826,977	4,828,108
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	Yes	No
Month of Sample FE	No	No	No	No	Yes
Month of Sample Ctrl.	No	No	No	Yes	No

Notes: This table reports results from estimating Equation (3.1), with the log of hourly energy consumption (averaged across "blocks" of three hours) as the dependent variable. As above, the independent variable is defined as the fraction of total expected savings that have been installed by time t. The untreated group in these regressions is chosen via nearest-neighbor matching. In particular, we match one untreated school to each treated 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 untreated 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 variables measured at the census block level, including the poverty rate, log of per capita income, school-level variables (enrollment; age of the school; grades taught; an academic performance index; and climate). These estimates are relatively sensitive to which schools are included. Standard errors, clustered at the school level, are in parentheses. All samples are trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable.

Table 4: Machine learning results by intervention type

	(1)	(2)	(3)	(4)	(5)
Any intervention:	-0.044	-0.045	-0.045	-0.029	-0.031
	(0.007)	(0.007)	(0.008)	(0.008)	(0.008)
Observations	19,113,602	19,113,602	19,112,664	19,112,664	19,113,602
HVAC interventions:	-0.044	-0.045	-0.046	-0.032	-0.032
	(0.008)	(0.008)	(0.008)	(0.009)	(0.009)
Observations	14,841,225	14,841,225	14,840,437	14,840,437	14,841,225
Lighting interventions:	-0.052	-0.052	-0.052	-0.036	-0.041
	(0.009)	(0.009)	(0.009)	(0.010)	(0.010)
Observations	12,849,701	12,849,701	12,849,024	12,849,024	12,849,701
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	Yes	No
Month of Sample Ctrl.	No	No	No	Yes	No
Month of Sample FE	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with the prediction errors the log of hourly energy consumption (averaged across "blocks" of three hours) as the dependent variable. The independent variable is defined as the fraction of total expected savings that have been installed by time t, and takes on values from 0 to 1 in the treated schools, and 0 always in the control schools. Standard errors, clustered at the school level, are in parentheses. All samples are trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. Regressions for HVAC and light interventions include only schools with an intervention of this type and pure untreated schools that never underwent energy efficiency interventions of any kind during the sample period. All regressions include a control for being in the post-training period for the machine learning.

Table 5: Realization rates by intervention type

	(1)	(2)	(3)	(4)	(5)
Any intervention:					
Energy consumption (kWh)	0.259	0.272	0.289	0.228	0.220
	(0.052)	(0.054)	(0.056)	(0.054)	(0.054)
Prediction error (kWh)	0.295	0.291	0.286	0.244	0.244
	(0.094)	(0.094)	(0.097)	(0.095)	(0.093)
Observations	19,110,690	19,110,690	19,109,832	19,109,832	19,110,690
HVAC interventions:					
Energy consumption (kWh)	0.493	0.505	0.565	0.435	0.445
	(0.099)	(0.101)	(0.099)	(0.093)	(0.098)
Prediction error (kWh)	0.561	0.552	0.543	0.456	0.489
	(0.156)	(0.156)	(0.164)	(0.161)	(0.157)
Observations	14,841,361	14,841,361	14,840,641	14,840,641	14,841,361
Light interventions:					
Energy consumption (kWh)	0.375	0.394	0.416	0.301	0.302
	(0.087)	(0.089)	(0.091)	(0.100)	(0.096)
Prediction error (kWh)	0.507	0.503	0.503	0.415	0.422
	(0.112)	(0.112)	(0.115)	(0.125)	(0.123)
Observations	12,849,868	12,849,867	12,849,255	12,849,255	12,849,867
School FE, Block FE	Yes	Yes	Yes	Yes	Yes
School-Block FE	No	Yes	Yes	Yes	Yes
School-Block-Month FE	No	No	Yes	Yes	No
Month of Sample Ctrl.	No	No	No	Yes	No
Month of Sample FE	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (4.1), with energy consumption in kWh or prediction errors in kWh as the dependent variable. The independent variable is the cumulative *ex-ante* expected savings from energy efficiency implemented by time t, scaled to the hourly level. A coefficient of one, therefore, implies that *ex-post* realized savings matched expected savings on average. Standard errors, clustered at the school level, are in parentheses. All samples are trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. In all cases, we can reject realization rates of zero and one at the 95 percent level.

Table 6: Realization rates: alternative estimation procedures

	(1)	(2)	(3)	(4)	(5)
Any intervention:					
Energy consumption (kWh)	0.220	0.238	0.500	0.452	0.062
	(0.054)	(0.067)	(0.084)	(0.070)	(0.092)
Prediction error (kWh)	0.244	0.437	0.652	0.465	0.552
	(0.093)	(0.109)	(0.100)	(0.065)	(0.088)
Observations	19,110,690	19,113,764	19,276,456	18,914,829	18,911,601
HVAC interventions:					
Energy consumption (kWh)	0.445	0.429	0.637	0.579	0.492
	(0.098)	(0.112)	(0.123)	(0.121)	(0.235)
Prediction error (kWh)	0.489	0.577	0.708	0.540	1.026
	(0.157)	(0.127)	(0.140)	(0.104)	(0.222)
Observations	14,841,361	19,113,764	19,276,456	18,914,829	14,702,384
Light interventions:					
Energy consumption (kWh)	0.302	0.309	0.377	0.334	0.429
	(0.096)	(0.077)	(0.093)	(0.072)	(0.156)
Prediction error (kWh)	0.422	0.479	0.396	0.351	0.676
	(0.123)	(0.122)	(0.101)	(0.062)	(0.132)
Observations	12,849,867	19,113,764	19,276,456	18,914,829	12,732,420
Estimation method					
Time-varying treatment	X				
Binary treatment		X	X	X	
Realized / expected savings					X
Trimming					
Dependent variable	X	X		X	X
Expected savings			X	X	X

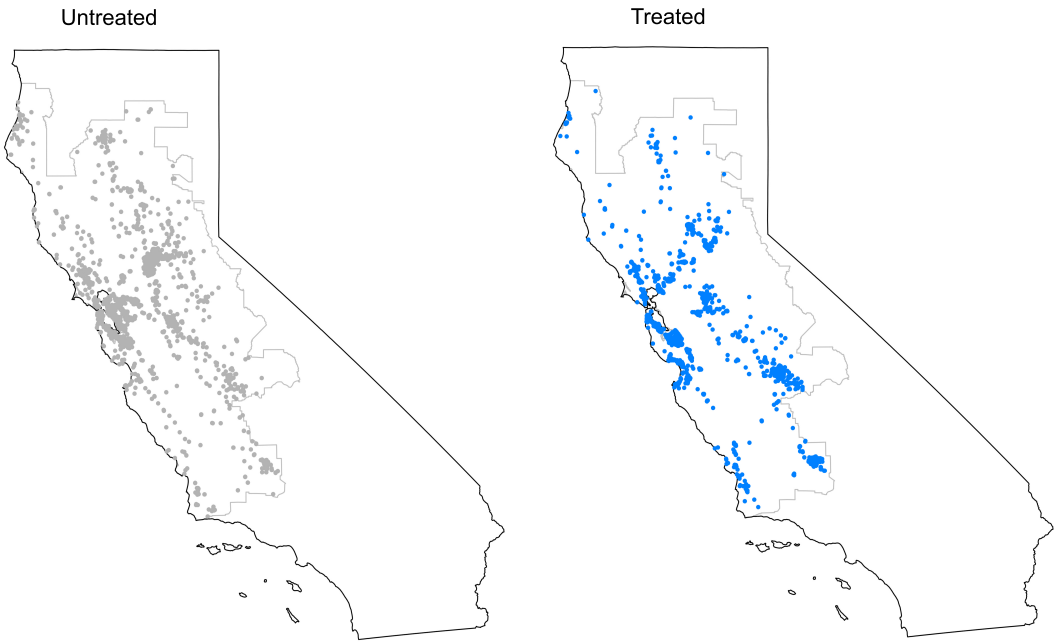
Notes: This table presents estimated realization rates and 95 percent confidence intervals for any intervention, HVAC interventions, and lighting interventions. Column (1) repeats the estimates from Column (5) of Table 5, that is, uses the cumulative *ex ante* expected savings installed by time t as the treatment variable, and the sample is trimmed to exclude observations below the 1st and above the 99th percentile of the dependent variable. Column (2) uses the same sample as Column (1), but instead uses a treatment indicator equal to 0 prior to treatment, and to the average expected savings installed in the post-treatment period after treatment begins. Column (3) uses the same dependent variable as Column (2), but trims schools to exclude the 1st percentile and 99th percentile of expected savings. Column (4) repeats the analysis from Column (3), and trims the 1st and 99th percentile of both expected savings and the dependent variable. Column (5) shows average realized savings divided by average expected savings during the treatment period, trimming the 1st and 99th percentile of expected savings and the dependent variable. All columns include school-by-block and month-of-sample fixed effects. Standard errors, clustered by school, are in parenthesis. In all but one case (HVAC machine learning in Column (5)), we reject realization rates of 1 at 95 percent confidence.

Table 7: Realization rate heterogeneity

Variable	(1)	(2)	(3)	(4)	(5)
Constant	0.495	0.440	0.537	0.610	0.676
	(0.064)	(0.125)	(0.161)	(0.144)	(0.156)
HVAC (0/1)		0.018	0.044	-0.001	0.109
		(0.131)	(0.170)	(0.152)	(0.168)
Lighting (0/1)		0.048	-0.091	-0.118	-0.252
		(0.127)	(0.160)	(0.142)	(0.157)
Longitude			0.233	0.204	0.319
			(0.187)	(0.169)	(0.188)
Latitude			0.319	0.342	0.430
			(0.124)	(0.113)	(0.125)
Average temperature (° F)			-0.028	-0.017	-0.100
			(0.150)	(0.135)	(0.145)
Total enrollment				0.082	0.069
				(0.071)	(0.077)
Academic perf. index (200-1000)					-0.037
					(0.086)
Poverty rate					0.054
					(0.092)
Number of schools	1,004	1,004	978	930	858

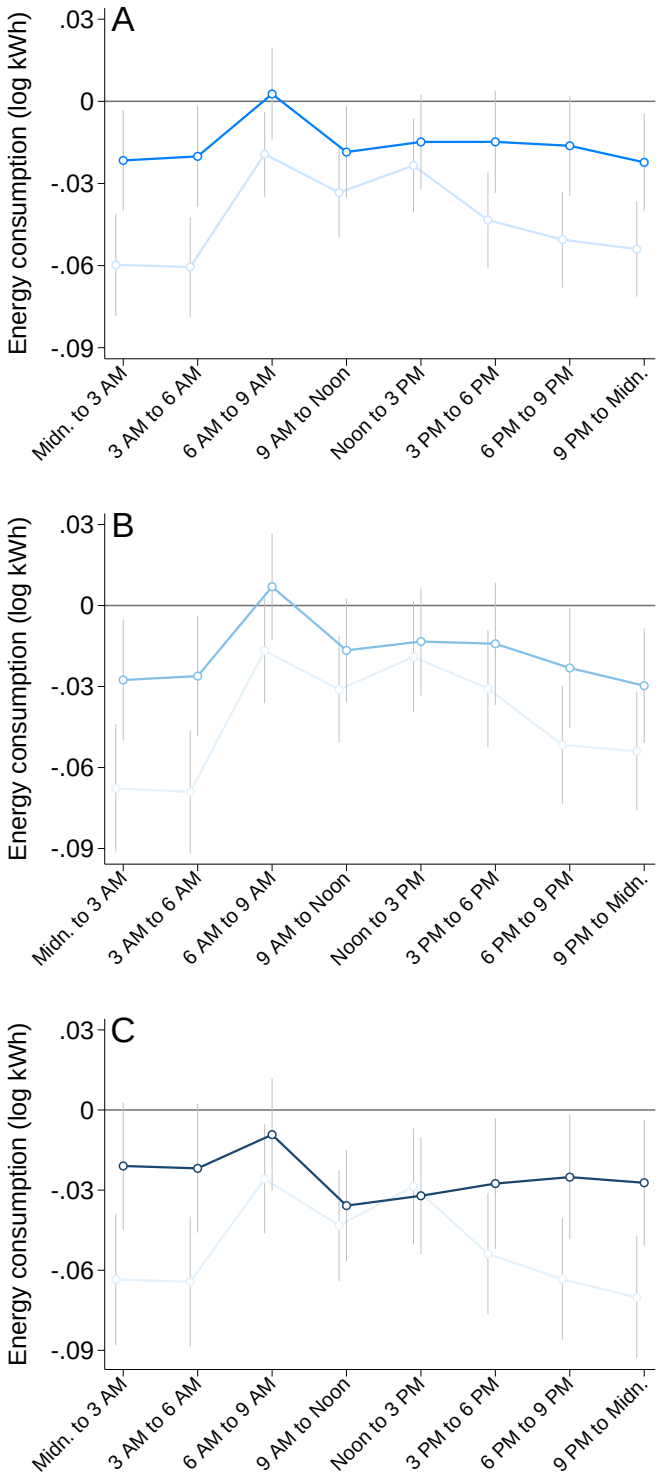
Notes: This table presents results from median regressions of school-specific realization rates on a variety of covariates. The school-specific realization rates are estimated from a regression of prediction errors (in kWh) on school-specific treatment indicators and school-by-block-by-month fixed effects, where we trim the dependent variable to exclude observations outside the 1st and 99th percentile. This table presents results for treated schools only. All estimates are weighted by the number of observations at each school. Standard errors, robust to heteroskedasticity, are in parentheses.

Figure 1: Locations of untreated and treated schools



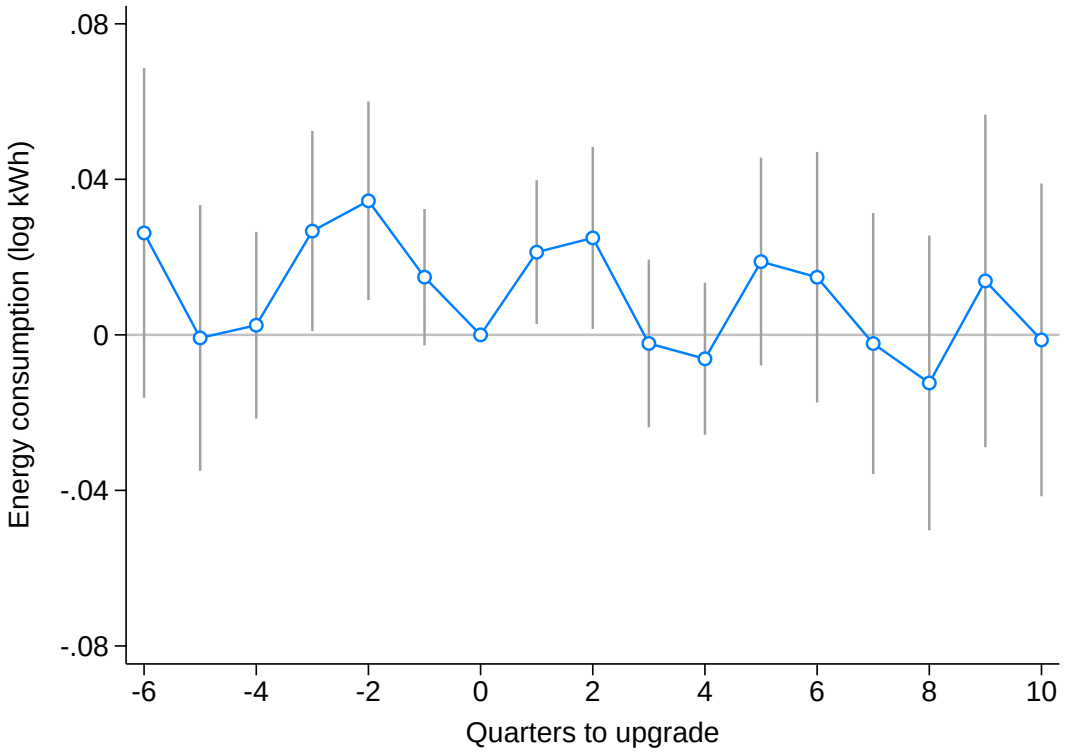
Notes: This figure displays the locations of schools in our sample. "Untreated" schools, in gray on the left, did not undertake any energy efficiency upgrades during our sample period. "Treated" schools, in blue on the right, installed at least one upgrade during our sample. There is substantial overlap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

Figure 2: Panel fixed effects results by hour-block



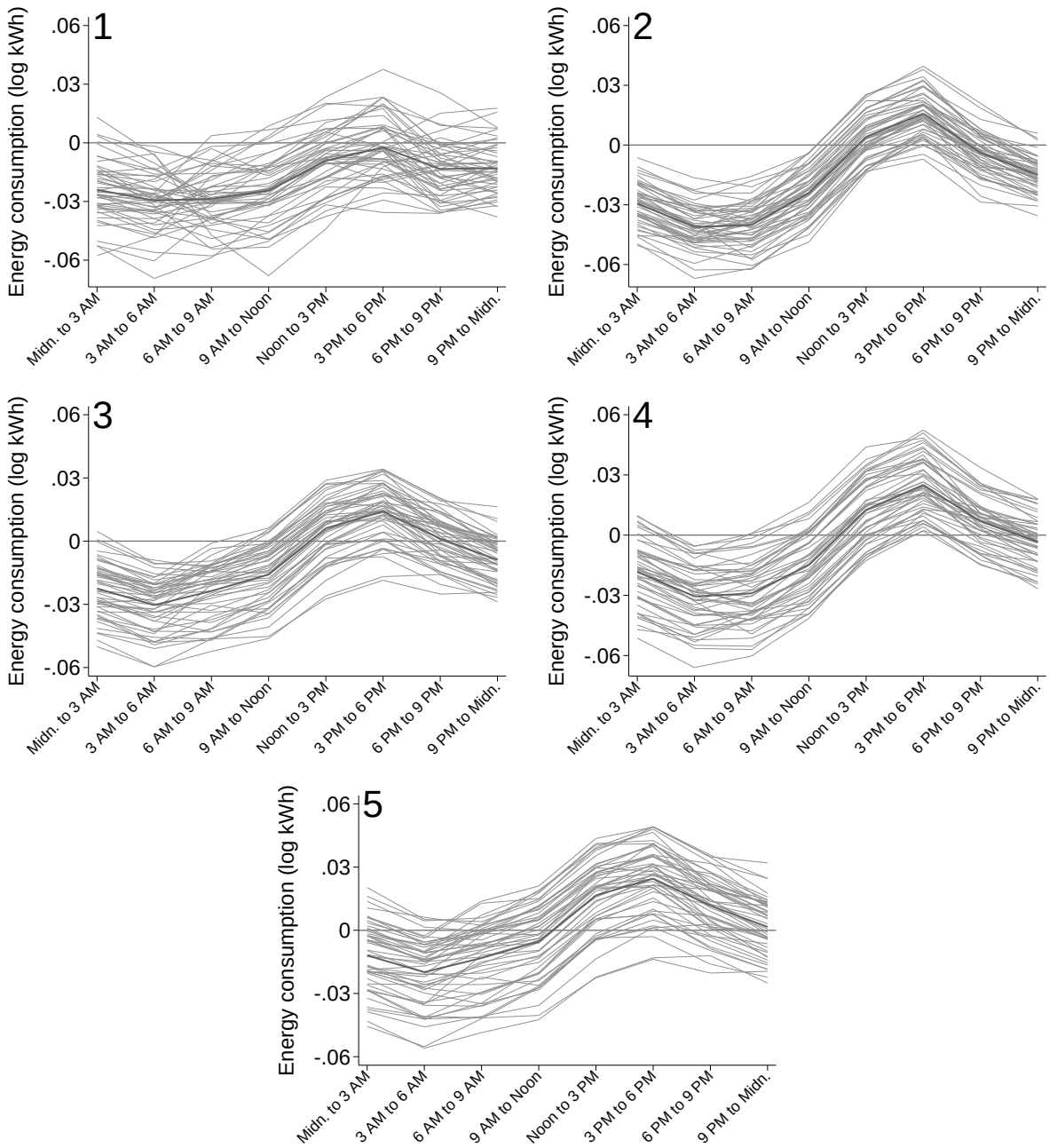
Notes: This figure presents treatment effects and 95 percent confidence intervals for each three-hour block of the day estimated using the log of energy consumption in kWh as the dependent variable. We present two specifications - corresponding to Columns (1) and (5) in Table 2. The first (light blue) has only school and block fixed effects; whereas the second (dark blue) has school-by-block and month-of-sample fixed effects. Panels A, B, and C present results for any intervention, HVAC interventions (compared against untreated schools only), and lighting interventions (same control group as B). Standard errors are clustered by school, and the sample has been trimmed to exclude observations outside the 1st and 99th percentile of the dependent variable.

Figure 3: Panel fixed effects event study



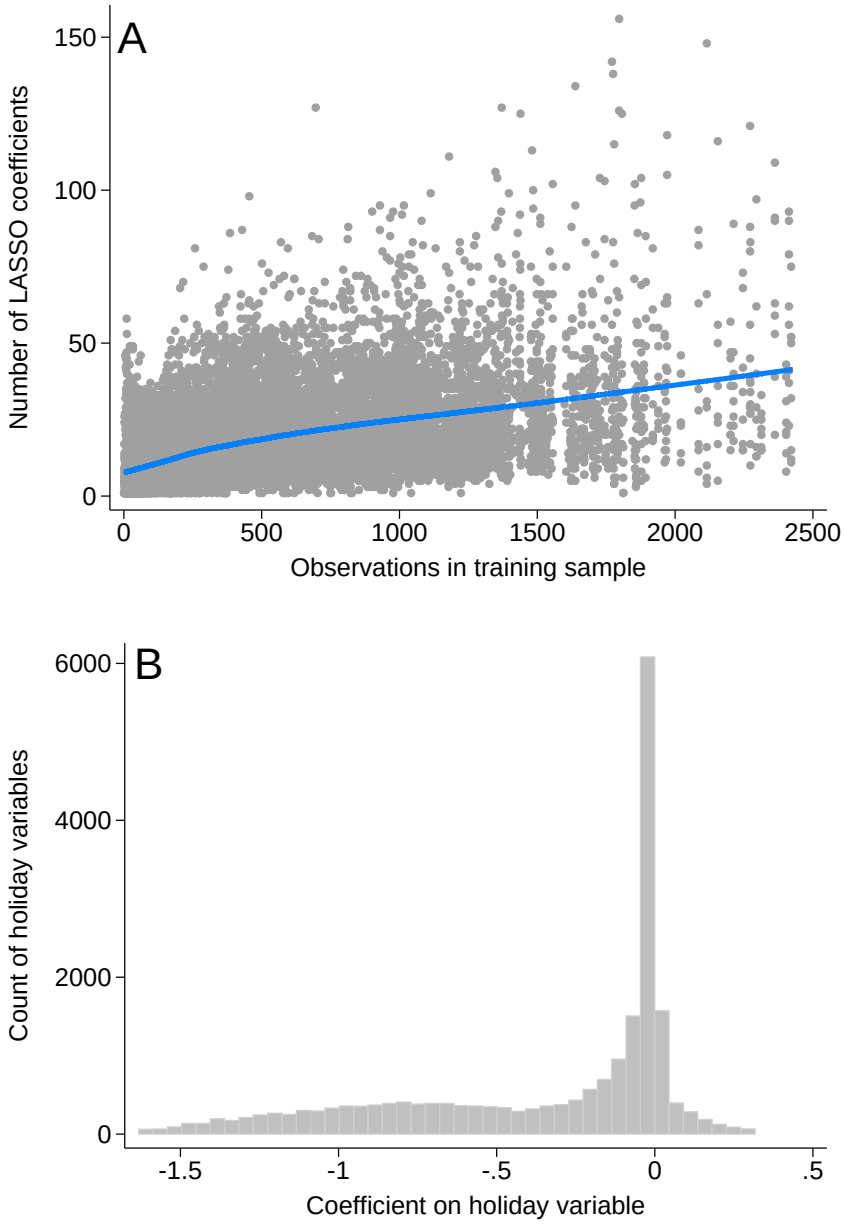
Notes: This figure displays point estimates and 95 percent confidence intervals from event study regressions of energy consumption before and after an energy efficiency upgrade. We estimate Equation (3.2) with the log of hourly electricity consumption (in kWh, averaged by three hour block) as the dependent variable. We normalize time relative to the quarter each school undertook its first upgrade. The underlying regression corresponds to Column (5) of Table 2, with school-by-block and month-of-sample fixed effects, and includes both treated and untreated schools. Standard errors are clustered by school, and the sample has been trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. Even with flexible controls, these estimates display strong patterns - perhaps reflecting seasonality in upgrade timing. We also do not see strong evidence of a shift in energy consumption as a result of energy efficiency upgrades.

Figure 4: Panel fixed effects placebo treatment effects



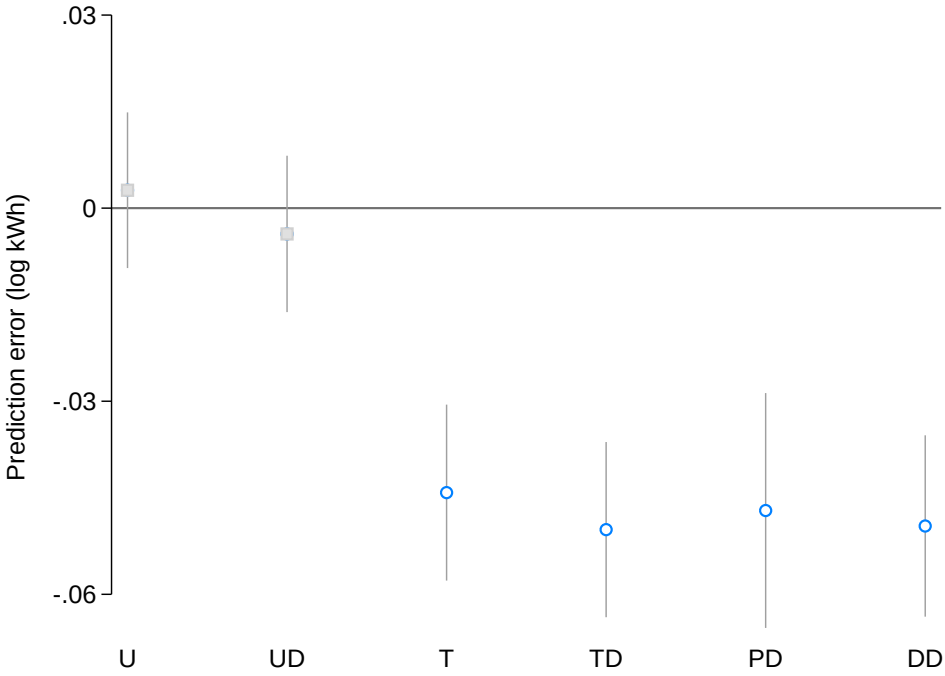
Notes: This figure displays the results of a placebo exercise with 50 runs. For each run, we begin with the untreated schools only, and randomly assign half to "treatment," which begins after a randomly-selected date. We then estimate Equation (3.1) with the log of hourly energy consumption as the dependent variable, and the placebo treatment indicator ("treat" × "post") as the independent variable. Each gray line displays the hour-block specific point estimates from one run. The dark gray line shows the average point estimate for each hour-block. Panel numbers indicate different controls, and correspond to the columns of Table 2. If our regression model were properly specified, we would expect to see flat treatment effects centered on zero. Even with the most flexible specifications, which include school-by-hour-block fixed effects, however, we see marked hourly patterns in the placebos.

Figure 5: Machine learning diagnostics



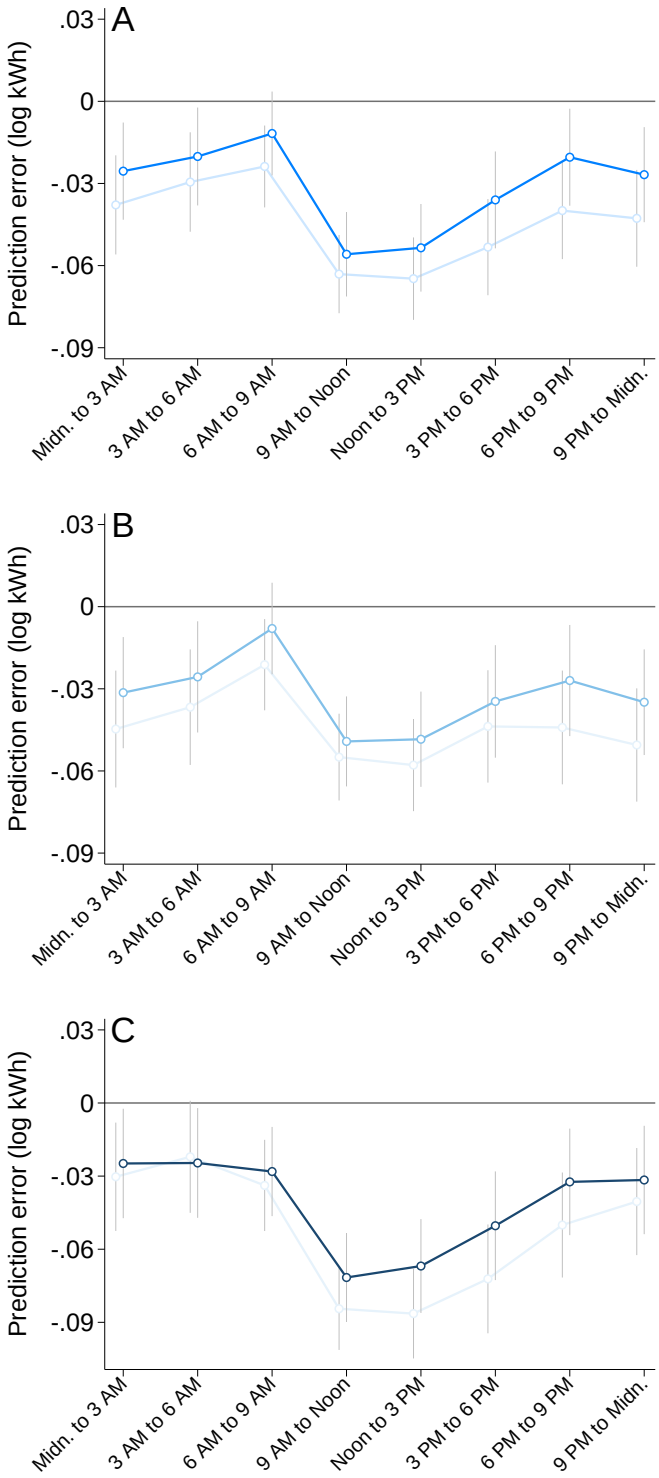
Notes: This figure presents two checks of our machine learning methodology. Panel A displays the relationship between the number of observations in the pre-treatment ("training") dataset and the number of variables LASSO selects to include in the prediction model for each school in the sample. Schools with very few training observations yield sparse models. As expected, the larger the training sample, the more flexible the prediction model becomes. This suggests that the LASSO is not overfitting. Panel B displays the marginal effect of holiday indicators in each school-specific prediction model. 2,137 schools' prediction models include at least one interaction with a holiday indicator (we allowed for up to 50 in the LASSO model), for a total of 21,953 holiday variables across all schools. Furthermore, the majority of the coefficients on these models are negative, which suggests that the LASSO model is picking up patterns that we would expect to be present in the data.

Figure 6: Comparing machine learning estimators



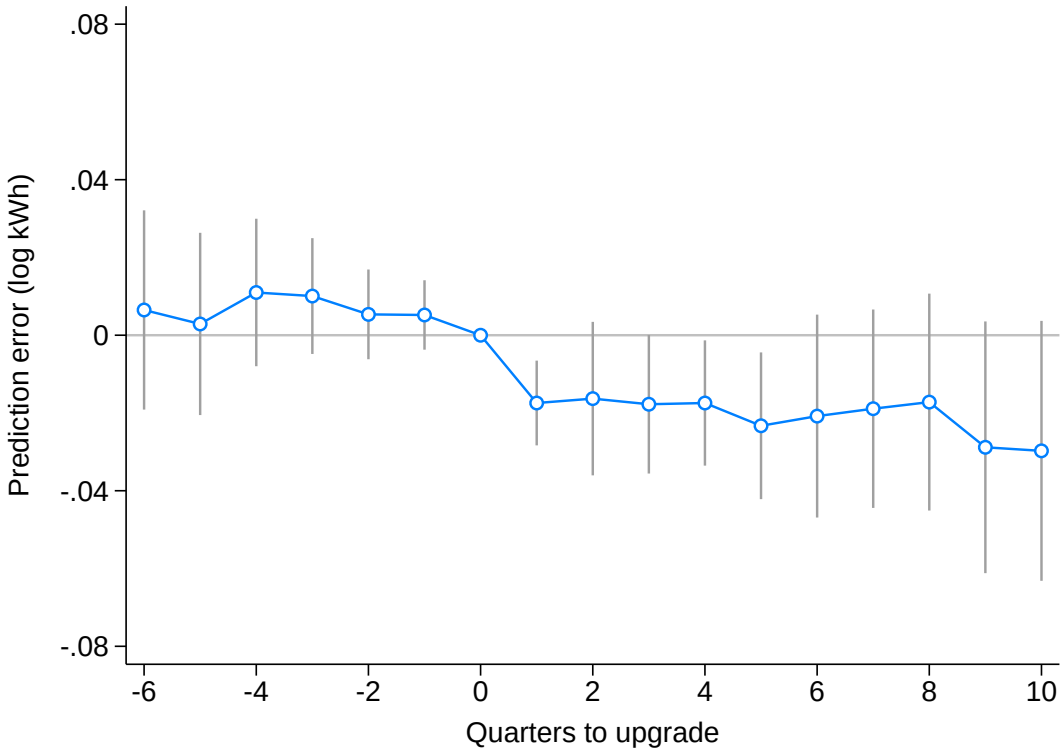
Notes: This figure shows average treatment effects, in the form of prediction errors based on log electricity consumption in kWh (averaged across three-hour "blocks"), from a variety of different machine learning estimators. The effect marked U shows prediction errors (real energy consumption minus predicted energy consumption) in untreated schools in the post period only; UD presents prediction errors in the untreated group in the post period minus pre-period prediction errors for the untreated group. We expect these effects to be close to zero, as they use untreated schools only. The effect marked T presents prediction errors in treated schools in the post-period only. TD presents prediction errors in the treated group in the post period minus pre-period prediction errors for the treated group. PD presents post-period prediction errors in the treated group minus post-period prediction errors in the untreated group. Finally, DD presents the prediction errors in the post- minus the pre-period for the treated group minus prediction errors in the post- minus pre-period for the treated group. For all estimators, we cluster our standard errors at the school level.

Figure 7: Machine learning results by hour-block



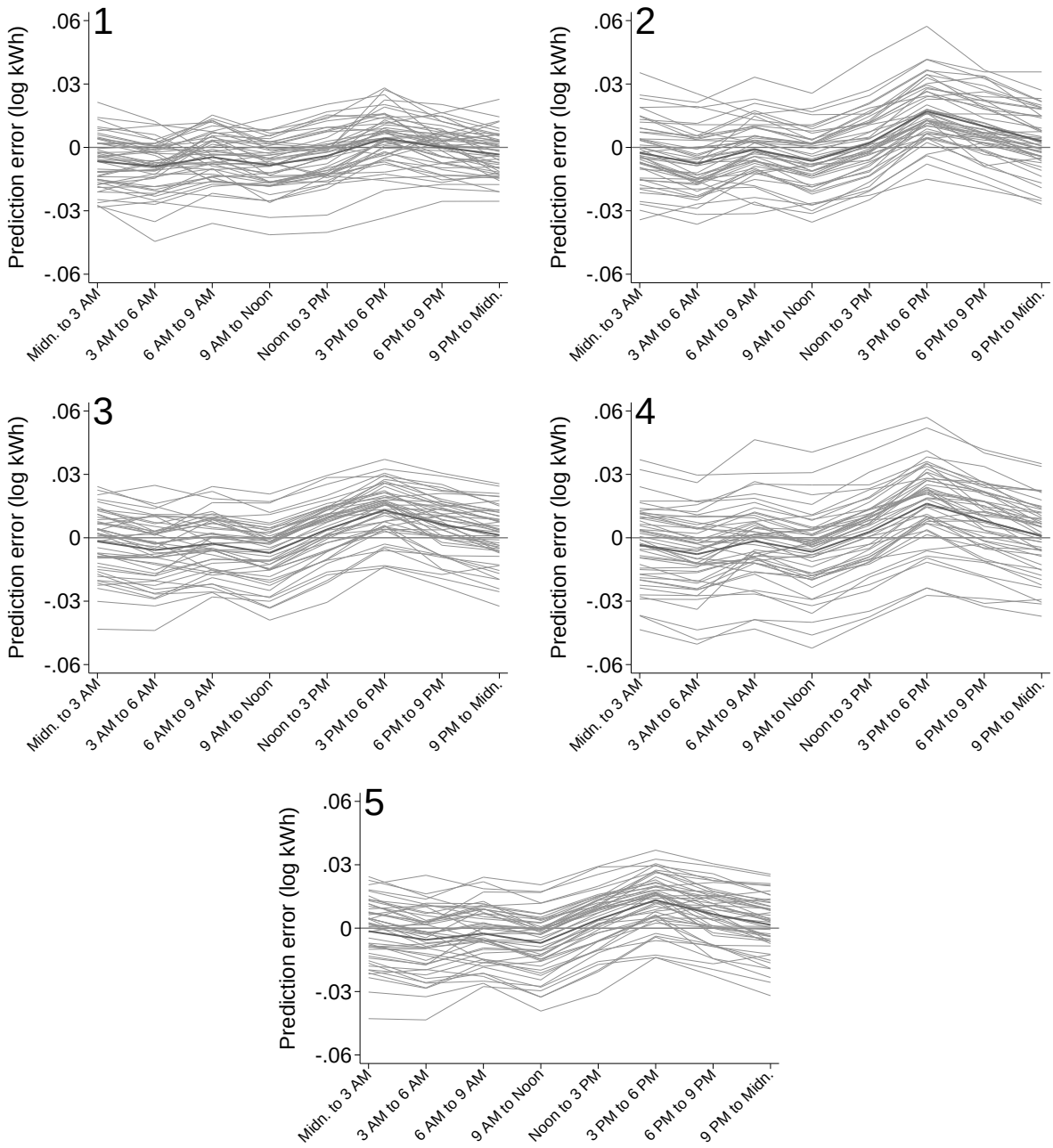
Notes: This figure presents treatment effects for each three-hour block of the day estimated using prediction errors based on log electricity consumption in kWh (averaged across three-hour "blocks") as the dependent variable. We present two specifications - corresponding to Columns (1) and (5) in Table 4. The first (light blue) has only school and block fixed effects; whereas the second (dark blue) has school-by-block-by-month-of-year fixed effects, as well as a month of sample control. Panels A, B, and C present results for any intervention, HVAC interventions (compared against untreated schools only), and lighting interventions (same control group as B). Standard errors are clustered by school, and the sample has been trimmed to exclude observations outside the 1st and 99th percentile of the dependent variable.

Figure 8: Machine learning event study



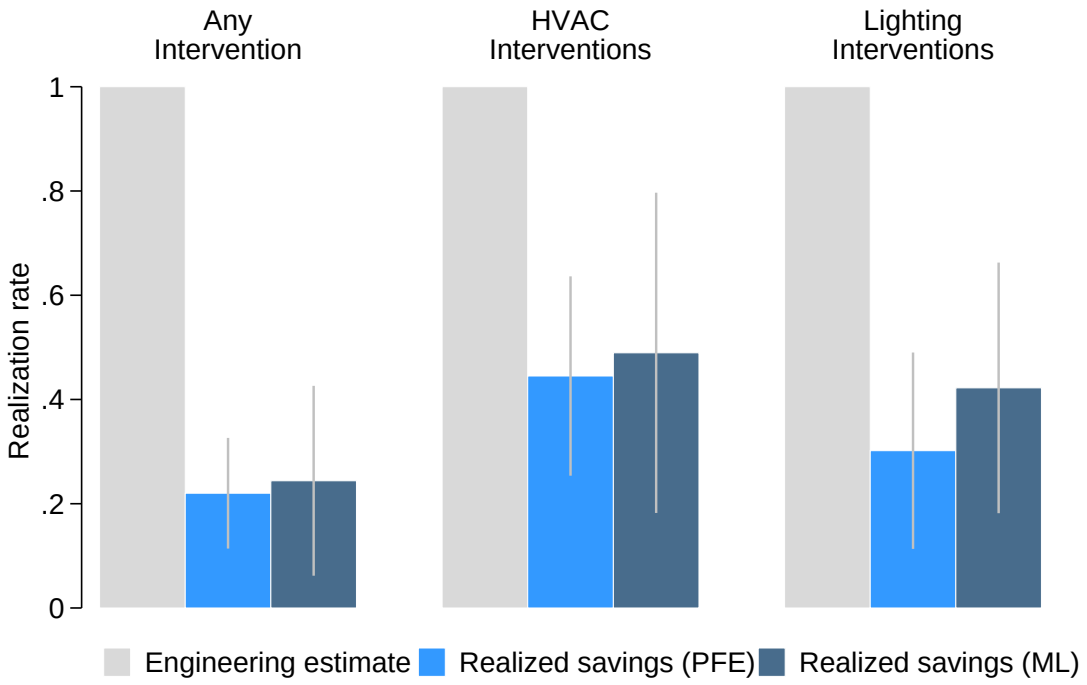
Notes: TThis figure displays point estimates and 95 percent confidence intervals from event study regressions of energy consumption before and after an energy efficiency upgrade. We estimate Equation (3.2) with prediction errors based on log electricity consumption in kWh (averaged across three-hour "blocks") as the dependent variable. We normalize time relative to the quarter each school undertook its first upgrade. The underlying regression corresponds to Column (5) of Table 4, with school-by-block and month-of-sample fixed effects, and includes both treated and untreated schools. Standard errors are clustered by school, and the sample has been trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. Unlike the regression estimates displayed in Figure 3, there is a clear change in energy consumption after the installation of energy efficiency upgrades, which persists more than a year after the upgrade. Furthermore, we fail to reject differential energy consumption between treated and untreated schools prior to the upgrades.

Figure 9: Machine learning placebo treatment effects



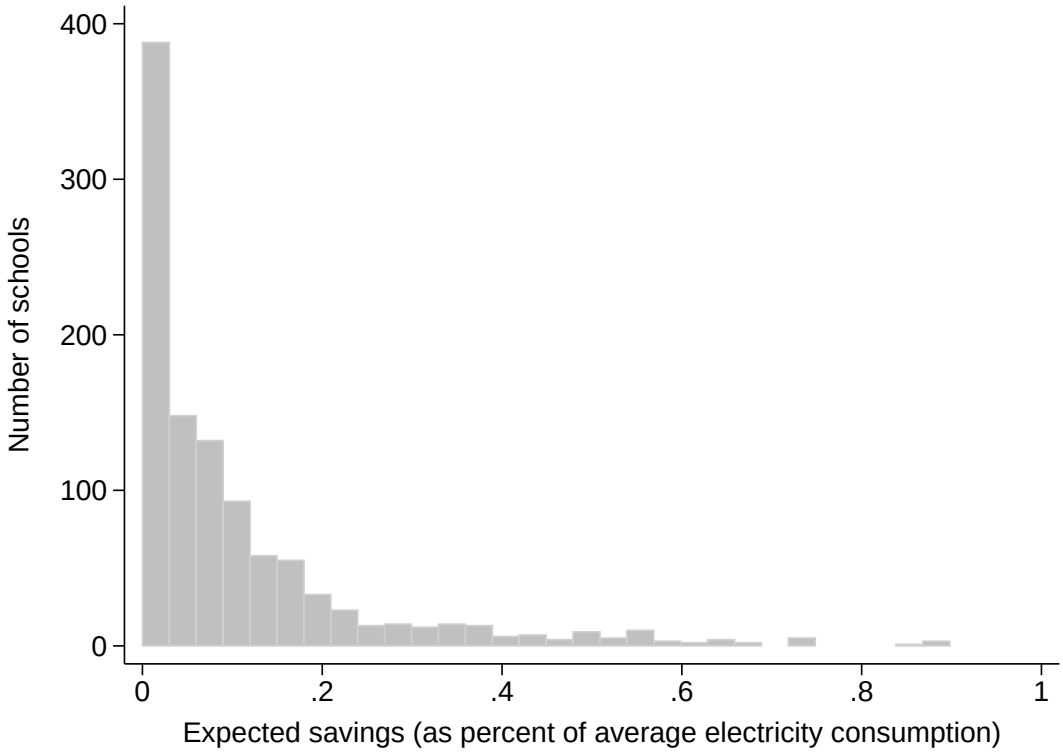
Notes: This figure displays the results of a placebo exercise with 50 runs. For each run, we begin with the untreated schools only, and randomly assign half to "treatment," which begins after a randomly-selected date. We then estimate Equation (3.1) with prediction errors based on log electricity consumption in kWh (averaged across three-hour "blocks") as the dependent variable, and the placebo treatment indicator ("treat" × "post") as the independent variable. Panel numbers indicate different controls, and correspond to the columns of Table 4. Each gray line displays the hour-block specific point estimates from one run. The dark gray line shows the average point estimate for each hour-block. If our regression model were properly specified, we would expect to see flat treatment effects centered on zero. Unlike in Figure 4, we see that the placebos are close to zero on average and exhibit only minor fluctuations over the course of the day.

Figure 10: Realization rates by intervention type



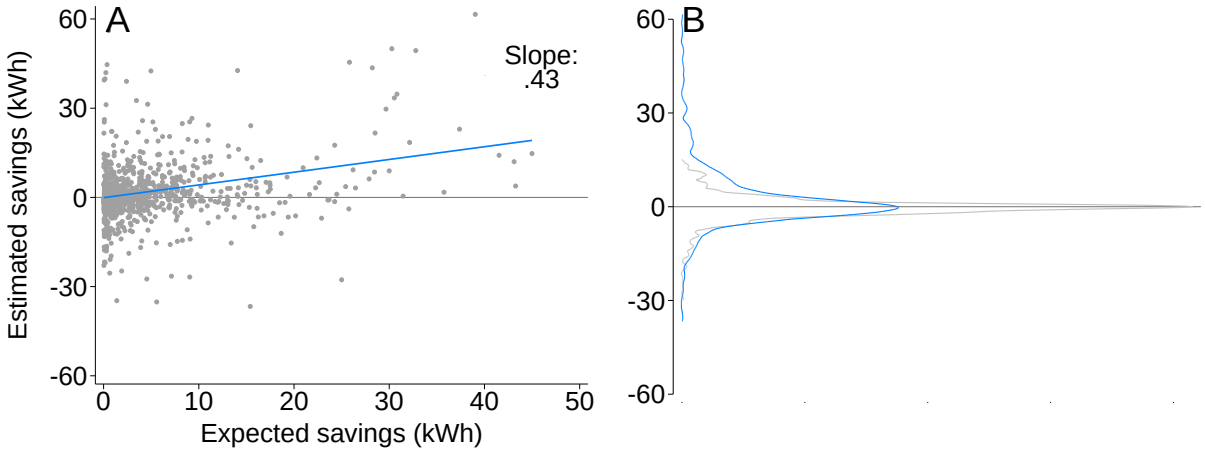
Notes: This figure displays realization rate estimates and 95 percent confidence intervals from estimating Equation (4.1). We present results for the Panel fixed effects method (log energy consumption in kWh as the dependent variable) and the machine learning method (prediction errors based on log electricity consumption in kWh as the dependent variable) against the engineering estimate of 1. The regressions in this figure include school-by-block and month-of-sample fixed effects, and correspond to Column (5) in Table 5. Standard errors are clustered at the school level, and samples are trimmed to exclude observations below the 1st or above the 99th percentile of the dependent variable. Realization rates are below 0.6 for all types of intervention, with the HVAC interventions appearing to outperform interventions in any category. In all cases, we can reject the engineering estimate of 1 at 95 percent confidence.

Figure 11: Expected savings relative to consumption



Notes: This figure shows the distribution of expected savings relative to average electricity consumption among the treated schools in our sample. While the median school is expected to reduce its energy consumption by 6 percent of its average consumption, the mean school has expected savings equal to 15 percent of its average consumption, reflecting the substantial right tail. This figure excludes the 15 schools where expected savings exceed 100 percent of average energy consumption. This figure suggests that there are irregularities in reported expected savings, as the expected savings-to-consumption ratios in the right tail appear unrealistically high.

Figure 12: School-specific effects



Notes: This figure displays school-specific savings estimates. We generate these estimates by regressing prediction errors in kWh onto an intercept and school-by-post-training dummies. We trim the sample in these regressions to exclude observations below the 1st and above the 99th percentile of the dependent variable. The coefficients on these dummies are the savings estimates. Panel A compares estimated savings with expected savings among treated schools only. This method produces a realization rate of 0.43 (weighted by the number of observations per school after removing outliers in expected savings), though there is substantial heterogeneity. Panel B displays kernel densities of estimated savings in the untreated group (gray line) and estimated savings in the treated group (blue line). While the untreated group distribution is narrow and centered around zero, the treated group appears shifted towards more savings.