MIT CEEPR

MIT Center for Energy and Environmental Policy Research

Working Paper Series

Machine Learning from Schools about Energy Efficiency

Fiona Burlig, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram

FEBRUARY 2019

CEEPR WP 2019-003

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

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

June 20, 2018

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. Measuring the returns to energy efficiency investments requires estimates of counterfactual energy consumption, and recent research suggests that industry standard approaches to measuring savings may be overstating the gains from energy efficiency considerably. We develop and implement a machine learning approach for estimating treatment effects using high-frequency panel data, which are now widely available from smart meters. We study the effectiveness of energy efficiency upgrades in K-12 schools in California, and demonstrate that the machine learning method outperforms standard panel fixed effects approaches. We find that energy efficiency upgrades deliver only 53 percent of *ex ante* expected savings on average, and find a similarly low correlation between school-specific predictions of energy savings and realized savings. We see suggestive evidence that HVAC and lighting upgrades perform closer to *ex ante* expectations, as do smaller upgrades. However, we are unable to predict high realization rates using readily available demographic information, making targeting-based improvements challenging.

JEL Codes: Q4, Q5, C4
Keywords: energy efficiency; machine learning; schools; panel data

[*] 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. Wolfram: 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, Harvard, the Colorado School of Mines, the University of Arizona, Arizona State University, Texas A & M, Iowa State, Boston College, the University of Maryland, Kansas State, Yale University, Columbia University, University of Warwick, the University of Virginia, New York University, the University of Pennsylvania, Carnegie Mellon, 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 (2016a), 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 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 combine machine learning techniques with a differences-in-differences estimator to derive causal impacts.

In particular, we use LASSO, a form of regularized regression with cross-validation, 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. 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 panel fixed effects model to estimate causal effects. Because we perform the prediction school-by-school, our approach becomes empirically tractable.[2]

We match hourly electricity consumption data from public K-12 schools in California to energy efficiency upgrade 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.

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

[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)).

[2] To our knowledge, this is one of the first papers in economics to incorporate machine learning methods into a differences-in-differences design in an applied setting. 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. 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 his approach in an empirical setting.

included in the specification. Our machine learning method yields estimates that are more stable across specifications. Second, we find that the panel fixed effects method performs poorly in an event study check and a series of other robustness tests. Even with rich school-by-time-of-day and month-of-sample fixed effects, this approach appears to be prone to bias. In contrast, we see no evidence of systematic bias when we subject our machine learning approach to the same event study test.

We also make a policy-relevant contribution to the literature 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 machine-learning method, we find that energy efficiency investments installed in California's K-12 schools are substantially underperforming relative to average *ex ante* engineering projections of expected savings. The average energy upgrade delivers only about 53 percent of expected savings. We also look specifically at the two most prevalent upgrade categories in our sample: heating, ventilation, and air conditioning (HVAC) and lighting, which together make up 74 percent of upgrades. We find that HVAC and lighting upgrades deliver 75 and 65 percent of expected savings on average, while all other interventions only achieve a realization rate of 42 percent, though these estimates are noisy. In addition to estimating realization rates on an average basis, we also estimate the school-specific correlation between engineering estimates and realized savings, which is 45 percent overall, 40 percent for HVAC upgrades, and 43 percent for lighting upgrades, somewhat lower than the overall estimates. We also find suggestive evidence that smaller upgrades perform closer to *ex ante* expected savings than larger upgrades. We demonstrate that while measurement error is likely present in our setting, our low realization rates are driven largely by real energy efficiency underperformance rather than attenuation bias. Finally, we explore the extent to which we are able to predict realization rates using easily-observable school characteristics. We are unable to identify covariates that correlate strongly with higher realization rates, suggesting that a targeting-based approach to reducing the gap between *ex ante* predictions and *ex post* outcomes 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 panel fixed approach methodology and present realization rate estimates using these standard tools (Section 3.1). Section 3.2 introduces our machine learning methodology and presents the results. In Section 4, we discuss measurement error, heterogeneity, school-specific realization rates, and targeting. 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)). This 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 efficiency 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.

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.[3] Next, they used GPS coordinates attached to each meter to match meters from this universe to school sites, using school location data from the California Department of Education. This results in a good but imperfect match between meters and schools. In some cases, multiple school sites match to one or more meters. This can often be resolved by hand, and was wherever possible, but several "clusters" remain. We use only school-meter matches that did not need to be aggregated. Robustness tests suggest that the results presented here do not change substantively when we include these "clusters." Our final sample includes 1,905 schools.

The PG&E data also describe energy efficiency upgrades at the schools as long as the school

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

applied for rebates from the utility.[4] 2,484 upgrades occurred at 920 schools between January 2008 and December 2014. For each energy efficiency measure installed, our data include the measure code, the measure description[5], a technology family (e.g., "HVAC", "Lighting", "Food service technology"), the number of units installed, the installation date, the expected lifetime of the project, the engineering-estimate of expected annual kWh savings, the incremental measure cost, and the PG&E upgrade incentive received by the school.[6] Many schools undertake multiple upgrades, either within or across categories. We include all upgrades in our analysis, and break out results for the two most common upgrade categories: HVAC and lighting. Together, these two categories make up over 74 percent of the total upgrades, and nearly 70 percent of the total projected savings in our sample.

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.[7]

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

2.2 Summary statistics

Table 1 displays summary statistics for the data described above, across schools with and without energy efficiency projects. Of the 1,905 schools in the sample, 920 undertook at least one energy efficiency upgrade. 566 schools installed only HVAC upgrades, and 439 received only lighting upgrades. There are 985 "untreated" schools that did not install any energy efficiency upgrades during our sample period. Our main variable of interest is electricity consumption, which we observe

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

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

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

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

[8] We performed our own sample cleaning procedure on the data from these stations, dropping observations with unreasonably large fluctuations in temperature, and dropping stations with more than 10% missing or bad observations. The raw data are available with a free login from http://mesowest.utah.edu/.

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, or approximately 5 percent of average annual electricity consumption. Savings are a slightly larger share of consumption for schools with lighting interventions.[9]

[Table 1 and Figure 1 about here]

The first three columns of Table 1 highlight measurable differences between treated and untreated schools. Treated schools consume substantially more electricity, appear in our sample earlier, are larger, and tend to be located to the southeast of untreated schools.

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). Because these schools are different on a range of observable characteristics, and because these indicators may be correlated with electricity usage, it is important that we consider selection into treatment as a possible threat to econometric identification in this setting.

3 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 control schools that we include in our analysis. Furthermore, a routine event study check demonstrates that this approach is prone to bias. We proceed by implementing a 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

Energy savings 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:

$$Y_{itb} = \beta D_{it} + \alpha_{itb} + \varepsilon_{itb} \tag{3.1}$$

[9] We do not summarize expected savings in Table 1, as all untreated schools have expected savings of zero.

where Y_{itb} is energy consumption in kWh at school i on date t during 3-hour-block b.[10] Our treatment indicator, D_{it}, is a dummy indicating that school i has undertaken at least one energy efficiency upgrade by date t. The coefficient of interest, β, is interpreted as the average savings in kWh/hour at a treated school. α_{itb} represents a variety of possible fixed effects approaches. Because of the richness of our data, we are able to include many multi-dimensional fixed effects, which non-parametrically control for observable and unobservable characteristics that vary across schools and time periods. Finally, ε_{itb} is an error term. To speed computation time, we estimate these regressions on data that we first collapse to the monthly level, and weight such that we recover results that are equivalent up to first order to our estimates on the disaggregated data.

We present results from several specifications with increasingly stringent controls. 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. 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 common 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. Across all specifications, we cluster our standard errors at the school level to account for arbitrary within-school correlations.

Realization rates In addition to estimating impacts of energy efficiency upgrades on energy consumption, we compare these estimates to average *ex ante* estimates of expected savings. We follow the existing energy efficiency literature in calculating realization rates.[11] Specifically, we calculate the realization rate as $\hat{\beta}$ divided by the average expected savings for upgrades in our sample.[12] If our *ex post* estimate of average realized savings matches the *ex ante* engineering estimate, we will estimate a realization rate of one. Realization rates below (above) one imply that realized savings are lower (higher) than expected savings.

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 energy consumption reductions of between 1.3 and 3.9 kWh per hour. 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 1.3 kWh reduction in energy consumption at treated schools. Estimates with a more parsimonious set

[10] We re-normalize energy consumption that our results will be in units of kWh consumed per hour.

[11] Davis, Fuchs, and Gertler (2014), Fowlie, Greenstone, and Wolfram (forthcoming), Levinson (2016b), Kotchen (2017), Novan and Smith (2018), and Allcott and Greenstone (2017) all use this method.

[12] We weight this average to account for the length of time each upgrade is installed.

of fixed effects, however, indicate savings nearly three times as large. These results are all precisely estimated; all estimates are statistically significant at the 1 percent level.[13]

[Table 2 about here]

Using this panel fixed effects approach, we find evidence that energy efficiency upgrades reduced school electricity consumption. However, these upgrades appear to under-deliver relative to *ex ante* expectations. In all specifications, we find realization rates below one, and we can reject them to be equal to one for all but specification (3). Our estimated realization rates range from 0.87 to 0.29 in our preferred specification. This suggests that energy savings in schools are not as large as expected.

3.1.2 Panel fixed effects robustness

Graphical analysis 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 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 perform a standard event study analysis to assess the validity of this assumption in this context. The event study sheds light on the ability of our panel fixed effects approach to adequately control for underlying differences between treated and untreated schools that vary over time. Figure 2 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[14]:

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

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 2 about here]

[13] In Appendix Table A.1, we present results using two-way clustering on school and month of sample, allowing for arbitrary dependence within schools and across schools within a time period. The results remain highly statistically significant using these alternative approaches.

[14] Appendix Figure A.1 shows the results for each of the five specifications in Table 4. All specifications display marked cyclical patterns, and in no case do we see strong evidence that energy consumption declined sharply after the upgrades were installed.

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, treated and untreated schools' energy consumption is likely not directly comparable.

Matching In order to address selection concerns, we conduct a nearest neighbor matching exercise, in which we use observable characteristics of treated schools to find similar untreated schools. Because the decision to invest in energy efficiency upgrades is often made at the district, rather than school, level, matching is conceptually challenging in this context. Allowing treated schools to match to any similar untreated school will likely induce selection bias by comparing schools that were chosen to be treated in a manner unobservable to the econometrician to those chosen not to be treated; on the other hand, forcing schools to match outside of their district can create problems with poor overlap. Appendix Table A.2 displays the results, using three different candidate control groups: all untreated schools; schools in the same district as the treated school only; and schools in other districts only. These results are highly sensitive to specification and the selected control group, providing further evidence that the standard panel fixed effects approach is unstable.[15]

Trimming As an additional robustness test, we examine the importance of outliers in driving our realized savings measures. Given that our energy consumption regressions are in levels, the results could be particularly sensitive to the presence of outliers. Appendix Table A.3 presents results from our preferred specification in which we trim the data to exclude the 1st and 99th or 2nd and 98th percentile of the dependent variable. Trimming reduces realization rates substantially, to 0.09 and 0.04, respectively. This is another cause for concern about the (lack of) stability of our realization rate estimates.

Taken together, the results from our main effects, event study check, matching approach, and trimming test demonstrate that the standard panel fixed effects approach is highly sensitive to specification and likely to be prone to bias, despite the rich set of fixed effects we are able to include in our preferred specification.

3.2 Machine learning approach

Even with a large set of high-dimensional fixed effects, the standard panel approach performs poorly on basic validity tests. In order to address some of these issues, we take advantage of the richness

[15] The synthetic control estimator, described by Abadie, Diamond, and Hainmueller (2010) is a natural alternative to the matching approach we use here. In our machine learning approach described below, we allow information from other untreated schools to inform our prediction of school i's energy consumption, in the spirit of this method.

of our data and implement a novel machine learning method for causal inference in panel data settings.

3.2.1 Methodology

Machine learning is optimized to make 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 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 estimate 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 isolate the effect of any particular variable, but rather to generate a good overall prediction. With the goal of achieving high predictive power, machine learning techniques enable the researcher to build an extremely flexible statistical prediction model, which can consider 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.[16] 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); Chernozhukov et al. (2018))).[17] 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. Belloni, Chernozhukov, and Hansen (2014) propose a "double selection" approach, using machine learning 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. Our approach is the most similar in spirit to Athey et al. (2017), in which the authors implement a matrix completion approach for estimating

[16] Athey (2017) and Mullainathan and Spiess (2017) provide useful overviews.

[17] These methods are useful when units are randomly assigned to treatment. In our context, however, non-random selection makes these approaches undesirable, as the partitioning itself may introduce selection bias.

counterfactuals in panel data.[18]

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-effect 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.[19] We do not use data after the treatment or pseudo-treatment date for the purposes of generating the model.[20]

In our baseline approach, 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. We begin with a set of many potential 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.[21] As a complement to these school-specific variables, we also allow for consumption at control schools to be a potential predictor, in the spirit of the synthetic control literature (Abadie, Diamond, and Hainmueller (2010)). Note that we generate a purely static prediction. That is, we do not include any lags or time trends in the prediction model, because we are generating predictions that we use substantially out of sample and these dynamics could dramatically impact predictions far into the future. The underlying assumption necessary for the predictions to be accurate is that units are in a relatively static environment, at least on average, which seems reasonable in this particular application.

As is standard with machine learning methods, the LASSO algorithm 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

[18] Because our setting necessitates allowing for a large amount of heterogeneity, this method is computationally infeasible in our context.

[19] 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.

[20] As an example, suppose that we observe 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.

[21] 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.

testing set, thereby guarding against overfitting. We repeat this process several times using tenfold cross-validation, with a new randomly-selected "training" and "testing" dataset each time. We tune the `glmnet` method to perform such cross-validation using a block-bootstrap approach, in which each week is considered to be a potential draw. This allows us to take into account potential autocorrelation in the data.[22]

Ultimately, we generate a model of energy consumption for each school-block. Training models for each school separately is critical in our context, where we observe a large amount of heterogeneity in energy consumption patterns across schools. Furthermore, generating block-specific models for each school makes our prediction approach extremely flexible while remaining relatively parsimonious.

Before proceeding, we examine the resulting prediction models. 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, such that the optimal model for any given school will not include all of the candidate regressors. In fact, we find that the prediction models constructed via cross-validation typically select fewer than 100 variables. While we find that the joint set of variables selected across all schools and blocks covers the majority of the candidate space, the selected models differ greatly across schools, highlighting the importance of allowing for flexible predictions across schools and hour-blocks. Panel A of Figure 3 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. Given that the LASSO algorithm works best in environments in which the true underlying model is sparse, the fact that the number of selected covariates asymptotes suggests that the LASSO is well-suited to this context.

[Figure 3 about here]

We can also inspect the selected covariates individually. 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.[23] Panel B of Figure 3 shows the coefficient on the holiday dummy (and its interactions - we allow up to 50 per school) in each school-block-specific prediction model. The LASSO selected nearly 5,390 holiday variables across the more than 16,000 school-blocks in our sample. We also find that, across models, holidays are negatively associated

[22] In the time series literature, other cross-validation approaches have been suggested where the training sample is gradually expanded over time. This is appropriate for methods where the prediction exercise includes a dynamic component (e.g., using recent lags to predict the future). Because our predictions are static, we use the standard cross-validation approach here.

[23] 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 exact dates of summer vacations, although the seasonal splines should help account for any long spells of inactivity at the schools.

with energy consumption. This suggests that the LASSO-selected models reflect real-world electricity use. We also find that the LASSO procedure often choose to include consumption at other untreated schools; the median school-block model includes over ten such covariates.

Panel C of Figure 3 shows the variables selected by each of the school-block models for treated and untreated schools separately. Nearly all of the models include an intercept, and over 90 percent of the models include consumption at at least one untreated school. Season and temperature variables are each included in nearly half of the models. Many models also include interactions between temperature and weekday dummies. This again demonstrates the substantial heterogeneity in prediction models across schools, and suggests that our machine learning method yields counterfactual predictions that are substantially more flexible than their traditional panel fixed effects analogue.

Finally, we use these models to construct predicted energy consumption for each observation in the sample. For each school, we predict block-specific energy consumption by plugging observed values of the selected covariates into the model. We generate a full model-predicted counterfactual time series for each school, which will allow us to compare our model predictions with actual energy consumption data throughout the sample.

While we use LASSO to generate these predictions, in principle, many different machine learning methods could be used for the prediction step. We experimented with six variants on the LASSO: we allow the LASSO to select from either the "basic" variables only, consumption at (other) untreated schools only, or both sets of variables. For each set of variables, we estimate predictions using two different tuning parameters – the cross-validated optimal λ, or λ plus one standard error, the default in `glmnet`. Finally, we also generate two sets of predictions using random forest. In the first, we produce school-by-block-specific predictions, as in the LASSO. In the second, we generate school-specific predictions only – a less flexible model. We test the performance of each of these methods by estimating the correlation between the prediction and actual energy consumption in the post-training period for untreated schools only, which allows us to examine overall fit completely out of sample. Table 3 displays the results of this exercise, showing the distribution of correlations between data and predictions across these six methods. The LASSO methods which use untreated schools perform better than those with the "basic" variables only, including the random forest models. We proceed with the method displayed in Column (4) of Table 3, which includes both basic variables and untreated schools, and uses `glmnet`'s default tuning parameter, as this model performs slightly better than the other options.[24]

[Table 3 about here]

[24] Appendix Figure A.2 displays hour-specific treatment effects estimated with each of these prediction methods, and the results are insensitive to the choice of method. We could also extend this set of methods to include ridge regression (as recommended in some settings by Abadie and Kasy (2017), or other prediction approaches. Because our chosen LASSO method performs relatively well, we proceed with this method in the interest of parsimony.

Estimation Armed with predicted energy consumption at each school, we can now estimate 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 realized energy consumption and our machine-learning-based predictions— at *untreated* schools only.[25] We construct two "treatment effect estimates": first, the average prediction error in the post-treatment period only, and next, the average difference between prediction errors in the post period and the pre period. We implement these estimators as a regression, and cluster the standard errors at the school level. We expect these estimates to be close to zero; this is suggestive evidence that our machine learning predictions are serving as good counterfactuals. The markers labeled "U" and "UD" in Figure 4 shows the results of this exercise for post-period-only and pre-minus-post estimation, respectively. As expected, we find that both of these estimates are statistically indistinguishable from zero. They are also both slightly negative, which suggests that there may be a time trend in energy consumption.

[Figure 4 about here]

Next, we compute treatment effects using four similar estimators. First, we compute the average prediction error in the post-treatment period for treated schools only, labeled "T" in Figure 4. Second, we compute the average difference between this prediction error and the pre-treatment prediction error for treated schools only ("TD"). Third, we use the untreated schools to bias-correct the estimator, and take the difference between "T" and "U" to form "PD": the average difference in prediction errors between treated and untreated schools in the post-treatment period only. Finally, we compare "TD" and "UD": the average difference in prediction errors between treated and untreated schools between the pre- and post-treatment periods, labeled "DD" in Figure 4. We find savings of between 3.8 and 4.2 kWh per hour for each estimator, with slightly smaller savings for the estimators that compare treated and untreated groups, in keeping with the bias correction.

3.2.2 Results

There are still potential confounding factors that could affect these average effects, such as time trends that the simple difference does not capture, or systematic school-specific errors or outliers that drive the means. 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

[25] 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.

Equation (3.1) with prediction errors in consumption in kWh as the dependent variable. We also present realization rates estimated using our machine learning method.

Table 4 reports results from estimating Equation (3.1) using five different fixed effects specifications. We find that energy efficiency upgrades resulted in energy consumption reductions of between 2.4 and 4.0 kwh per hour - substantially less sensitive than the panel fixed effects estimates presented in Table 2. In our preferred specification (Column (5)), which includes school-by-hour-block and month-of-sample fixed effects, we find that energy efficiency upgrades reduced electricity use by 2.4 kWh per hour in treated schools relative to untreated schools. These results are both larger and more stable across specifications than the panel fixed effects results above. Our results are highly statistically significant.[26]

[Table 4 about here]

We again compare these results to the *ex ante* engineering estimates to form realization rates. Our estimated realization rates range from 0.89 to 0.53 in our preferred specification. These realization rates are statistically different than zero and larger than the estimates from our panel fixed effects approach, but still imply that realized savings were substantially lower than *ex ante* expectations, although we cannot reject a realization rate of one in the first three specifications.

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 4 above, we estimate close-to-zero savings on average among our untreated group. This is reassuring, as it suggests that the low realized savings are not driven by unmeasured efficiency upgrades at untreated schools, and are instead likely driven by overly optimistic *ex ante* predictions or rebound.

3.2.3 Machine learning robustness

As with the standard panel fixed effects approach, the identifying assumption underlying the results in Table 4 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

[26] In Appendix Table A.4, we present results two-way clustering on school and month of sample. The results remain highly statistically significant using these alternative approaches. 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.

predicted and actual 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. Here, we subject the machine learning approach to the same event study check as the panel fixed effects method and find that it performs substantially better. We also show results for a variety of different machine learning algorithms, and demonstrate that the results are robust across these methods. Finally, we test the sensitivity of our machine learning estimates to trimming, and find that the results are much more stable than in the panel fixed effects approach.

Graphical analysis We present graphical evidence from an event study regression of prediction errors on indicator variables for quarters relative to treatment. Figure 5 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.[27]

[Figure 5 about here]

Figure 5 shows flat treatment effects in the 6 quarters prior to an energy efficiency upgrade. Unlike in Figure 2, 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 kWh reduction in energy use, is stable and persists up to 10 quarters after the upgrade occurs, though the later quarters are more noisily estimated. This event study figure provides strong evidence to suggest that the machine learning approach —unlike the panel fixed effects approach above— is more effectively controlling for time-varying differences between treated and untreated schools.

Alternative prediction approaches How sensitive are our results to our use and implementation of the LASSO algorithm? Depending on the underlying data, different algorithms may be more effective than others (Mullainathan and Spiess (2017)). As described in Section 3.2.1, the LASSO appears to generate well-behaved models. Furthermore, we find similar out-of-sample prediction effectiveness in untreated schools across our choice of tuning parameters and potential covariates, as well as when we train our models using a random forest algorithm rather than a LASSO algorithm. Similarly, we find that our main results look remarkably similar across these techniques. Appendix Table A.5 shows the results where we estimate (3.1) with different prediction algorithm approaches. We find energy savings between 2.86 kWh per hour and 1.95 kWh per hour. Using our preferred LASSO approach (Column (4)), we estimate savings of 2.36 kWh per hour. These

[27] Appendix Figure A.3 shows the results for each of the five specifications in Table 4. All specifications display a marked downward shift in energy consumption after the upgrade.

estimates translate into realization rates of 0.64, 0.44, and 0.53, respectively.[28] These estimates are generally not statistically distinguishable, and our preferred approach lies in the middle of the range of estimated realization rates, suggesting that the machine learning approach is not highly sensitive to our chosen prediction method.

As an additional robustness check, rather than training our models on pre-treatment data only and forecasting these models into the post-treatment period, we instead train our models on *post*-treatment data only and forecast these models into the *pre*-treatment period. If the models are performing as expected, we should recover similar results.[29] Appendix Table A.6 presents the results of this exercise. When we train the model on post-treatment data, we find energy savings estimates of between 3.73 kWh per hour and 1.96 kWh per hour (in our preferred specification), or realization rates of 0.84 and 0.44, respectively. We also estimate a version where we pool the data from both the "forward" and "backward" predictions, and recover realization rates ranging from 0.87 to 0.48. These results are qualitatively consistent with and statistically indistinguishable from our main results.

Finally, we implement a variant on the Belloni, Chernozhukov, and Hansen (2014) "double selection" method. This approach allows for proper inference with model selection. We adapt it to the panel setting as follows. We first estimate a LASSO to predict the timing of treatment. We estimate a second LASSO to predict electricity consumption, and finally estimate a third LASSO with time as the dependent variable, which allows for trends. We then regress energy consumption on treatment timing and the union of the non-zero-coefficient variables from all three LASSOs. To make this computationally tractable, we apply the selection of variables on a school-by-school basis. Finally, we residualize each dependent variable by the full set of controls and regress residualized prediction errors on the residualized treatment date error and the residualized time error with all schools pooled.[30] The final panel of Appendix Table A.6 presents the results. We find realization rates ranging from 1.26 to 0.56 with our preferred specification. The double LASSO approach is substantially more sensitive to the inclusion of time fixed effects than our main machine learning approach, but with our preferred specification, we find a very similar realization rate of 0.56 to our main estimate of 0.53.

Trimming As a final robustness check, we test the extent to which our machine learning results vary as we exclude outliers. As with the panel fixed effects approach, we trim observations below the 1st (2nd) and above the 99th (98th) percentile of the dependent variable – now defined as prediction errors in energy consumption. Appendix Table A.7 shows the results for our preferred

[28] Appendix Figure A.2 shows hour-block-specific treatment effects for all of the machine learning methods shown here. The hourly patterns are very similar across methods.

[29] We set up these regressions so as to recover estimates that are the same sign as the main effect.

[30] By Frisch-Waugh-Lovell, these procedures are equivalent, but computationally less taxing, than estimating the full unresidualized regression.

specification. We find trimmed energy savings estimates of 2.33 and 2.25 kWh per hour using 1 percent and 2 percent trimming, yielding realization rates of 0.55 and 0.54, respectively. These results are substantially more stable than their panel fixed effects counterparts.

In contrast with the standard panel fixed effects approach, our machine learning method delivers results that are substantially less sensitive to specification and outliers. Furthermore, our machine learning method outperforms the panel fixed effects approach on our event study check, suggesting that it is less prone to bias. Finally, our predictions appear to be well-behaved, and our results are insensitive to our choice of prediction method. Taken together, this suggests that our machine learning approach is more effective than the standard panel fixed effects method.

4 Discussion

Using our machine learning approach, we estimate an overall realization rate of 53 percent: realized energy savings are just over half of *ex ante* expected savings. In this section we discuss the impact of measurement error on our results and the role of heterogeneity, including school-specific realization rates and the potential for targeting to improve program effectiveness.

4.1 Measurement error

One 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 months, or if there is simply recording 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.

The vast heterogeneity across measures also can complicate estimating the effectiveness of these energy efficiency interventions. Figure 6 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.[31]

[31] 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 6.

[Figure 6 about here]

Including the schools at the right tail of the distribution in Figure 6 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 sample trimming approach. Table 5 displays the results of this exercise. Column (1) replicates the realization rates from Column (5) in Table 4. In Column (2), we remove schools with expected savings below the 1st or above the 99th percentile. We find a realization rate of 0.55, quantitatively similar to our main result, which suggests that measurement error in expected savings is not driving our results.[32]

[Table 5 about here]

In estimating realized savings, we use a simple post-treatment indicator as the independent variable of interest. This may attenuate our treatment effect towards zero as there may be measurement error in the timing of treatment.[33] We address this with an alternative specification in which we take into account the timing of treatment for schools that install more than one upgrade:

$$Y_{itb} = \delta N_{it} + \alpha_{i,t,b} + \varepsilon_{itb} \tag{4.1}$$

where N_{it} is the number of upgrades installed in school i by time t. Column (3) of Table 5 displays the results of this exercise. Using this alternative treatment variable, we find a substantially smaller realization rate of 0.29. If anything, this suggests that our main realization rate estimate of 0.53 is conservative relative to a null hypothesis of 1, which would imply that the *ex ante* engineering estimate is correct. However, this alternative specification is also likely subject to attenuation bias if there is measurement error in the timing of treatment, as the treatment variable now adjusts every time a new upgrade occurs. The fact that using this alternative definition of treatment leads us to estimate *smaller* realization rates is consistent with there being a substantial amount of measurement error in the treatment dates, leading us to prefer our main specification with a binary treatment indicator.

[32] Appendix Table A.8 presents analogous results using the panel fixed effects approach.

[33] Note that if we had perfect measures of treatment timing, the resulting estimates would not be attenuated by the fact that schools underwent multiple upgrades, as we scale our expected savings measure by the amount of time spent at each level of expected savings.

Overall, these two exercises suggest that measurement error is an important consideration for this literature. While we provide suggestive evidence that measurement error matters in our context, it does not appear to explain the entire gap between *ex ante* engineering estimates and realized outcomes.

4.2 Heterogeneity

In light of our low overall realization rate estimates, we now seek to understand whether these realization rates vary by upgrade type or size, which is informative for policymakers deciding which (if any) upgrades to subsidize.[34]

Upgrade type We begin by evaluating realization rate heterogeneity by upgrade type. In particular, we compute separate realization rates for HVAC and lighting upgrades. These upgrades together account for over 74 percent of the upgrades and nearly 70 percent of the expected savings in our sample. To estimate separate realization rates for these categories, we re-estimate Equation 3.1 with separate indicators for HVAC, lighting, and other upgrades:

$$Y_{itb} = \lambda D_{it}^{\text{HVAC}} + \gamma D_{it}^{\text{lighting}} + \tau D_{it}^{\text{other}} + \alpha_{i,t,b} + \varepsilon_{itb} \tag{4.2}$$

This allows us to isolate the impact of these upgrades from all other upgrade types. The second panel of Table 5 shows our realization rate estimates for these upgrades.

We find that HVAC and lighting upgrades achieve much higher realization rates than our estimate for all of the upgrades in our sample. Using our main specification, shown in Column (1), We find a realization rate of 0.75 for HVAC upgrades and 0.65 for lighting upgrades, and only 0.42 for all other upgrades.[35] This suggests that HVAC and lighting upgrades come closer to the *ex ante* engineering predictions than the average upgrade in the sample, though these estimates are quite noisy. We also subject these heterogeneous effects estimates to the same measurement error checks discussed above. We again find that HVAC and lighting upgrades tend to outperform other categories.

Upgrade size Next, we investigate the relationship between upgrade size and realization rates. To do this, we estimate energy savings separately for schools with above-median upgrade sizes and schools with below-median upgrade sizes. We split schools using two definitions of upgrade

[34] Because our focus in this paper is on realization rates, which are determined by overall savings, we do not focus here on heterogeneity of treatment effects by time. As Borenstein (2002) and Boomhower and Davis (2017) point out, however, the value of energy savings varies over time. We estimate hour-block-specific treatment effects, and present the results in Appendix Figure 7. We find evidence that the largest reductions occur during the school day – consistent with our results picking up real, rather than spurious, energy savings. This is suggestive that that the reductions in our sample are happening at high-value times, though peak power consumption hours in California occur between 4 and 8 PM, after the largest estimated reductions from the energy efficiency upgrades in our sample.

[35] Appendix Table A.8 presents results using the panel fixed effects approach.

size: first, by total expected savings in kWh, and second, by expected savings relative to average consumption. After dividing schools into expected savings groups, we estimate group-specific energy savings:

$$Y_{itb} = \sum_g \beta^g D_{it} \cdot \mathbf{1}[\text{quartile} == q] + \alpha_{i,t,b} + \varepsilon_{itb} \tag{4.3}$$

We then compute realization rates by dividing our savings estimates by the average expected savings in the relevant group. Table 6 presents the results. We find a realization rate of 1.27 (0.92) for below-median schools when we split on absolute (relative) savings, and a realization rate of 0.53 (0.47) for above-median schools. While the point estimates suggest that schools with below-median expected savings (in both absolute and relative terms) have higher realization rates than schools with above-median expected savings, the results are extremely noisy.

[Table 6 about here]

4.3 School-specific realization rates

Up to this point, we have estimated realization rates by comparing our estimates of the energy savings from energy efficiency upgrades on average ($\hat{\beta}$ from estimating Equation (3.1)) to the average expected savings in the sample. This estimator — standard in the literature — 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? For policy makers who wish to learn about the program-wide effectiveness of energy efficiency upgrades, this is the relevant estimator.

However, we may also be interested in the correlation between expected savings *at a given school* and realized savings *at that same school.* In particular, for a (risk averse) principal or district administrator deciding whether or not to invest in energy efficiency, this "school-specific" realization rate is likely to be relevant. In order to estimate these school-specific realization rates, we estimate equations of the following form:

$$Y_{itb} = -\gamma S_{it} + \alpha_{itb} + \epsilon_{itb}, \tag{4.4}$$

where Y is the prediction error (in kWh per hour), 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.[36] As above, α_{itb} 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

[36] For a concrete example, suppose that a school undergoes two upgrades, one that is expected to save 25 kWh per hour, and the other expected to save 75 kWh per hour. S_{it} will be equal to zero before the first upgrade, to 25 after the first upgrade, and to 100 after the second upgrade.

effects.

The coefficient of interest is γ. These regressions are scaled 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.[37] 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.

Columns (4), (5), and (6) of Table 5 presents the results. For all upgrades in the sample, we find realization rates of 0.45, 0.65, and 0.08, without trimming, trimming the 1st and 99th percentile of expected savings, and using a time-varying treatment indicator, respectively.[38] We find similar patterns as with the average regressions, though again, our realization rate estimates are quite noisy. We can reject realization rates of one in nearly all cases. Using this approach, the evidence is less strongly in favor of HVAC upgrades performing better than other upgrades. These results corroborate our main estimates, and suggest that energy efficiency upgrades are falling short of *ex ante* expectations both across the program and on a site-specific basis.

4.4 Targeting

Given the richness of our electricity consumption data, we can potentially go beyond average treatment effect estimation and instead estimate treatment effects for each school separately. Of course, the identification assumptions to obtain school-specific treatment effects are much stronger than when obtaining average treatment effects, as coincidental changes in consumption at a given school will be confounded with its estimated 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. 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 recover potential systematic patterns across schools.

Panel A of Figure 8 displays the relationship between these school-specific savings estimates and expected savings for treated schools. We find a positive correlation between estimated savings

[37] We must exercise caution when trying to compare the results from this estimator to estimates from Equation (3.1). 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). 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.

[38] See Appendix Table A.8 for analogous results using the panel fixed effects approach.

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 53 percent, the same as our main realization rate estimate. 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 8 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.[39] 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 close to 80 percent. Column (2) shows that median realization rates are somewhat larger for HVAC and lighting interventions, although this difference is not statistically significant. We add latitude and longitude in Column (3), but these are not significantly correlated with realization rates. Columns (4)-(5) control for yet more covariates, including total enrollment, the Academic Performance Index and the poverty rate. We do not find statistically significant correlations between these observable characteristics and realization rates.[40] In Column (6), also look at the relationship between expected savings and realization rates. As above, we find that schools with larger expected savings tend to have lower realization rates.

Ultimately, we do not detect clear relationships between school characteristics and realization rates. This suggests that uncovering "low-hanging fruit" may be challenging, and, furthermore,

[39] 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.

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

that improving the success of the types of energy efficiency upgrades in our sample via targeting will likely prove difficult. That said, several features of our setting make recovering this type of pattern challenging. First, our sample of treated schools is relatively small—there are fewer than 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. 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 Conclusion

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 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, 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 in conjunction with our preferred fixed effects specification, we find that energy efficiency investments reduced energy consumption by 2.4 kWh per hour on average. While these energy savings are real, they represent only 53 percent of *ex ante* expected savings, and we can reject a realization rate of 100 percent. 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.

Finally, we explore heterogeneity in realization rates. We find some evidence that HVAC upgrades outperform other upgrades, and that upgrades with smaller expected savings come closer to *ex ante* engineering estimates. We attempt to use information that is readily available to policymakers to predict which schools will have higher realization rates, but 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 in determining 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, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program."

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, Mohsen Bayati, Nikolay Doudchenko, Guido Imbens, and Khashayar Khosravi. 2017. *Matrix Completion Methods for Causal Panel Data Models.* Working Paper 1710.10251. arXiv.

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.

Borenstein, Severin. 2002. "The Trouble With Electricity Markets: Understanding California's Restructuring Disaster." *Journal of Economic Perspectives* 16 (1): 191–211.

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.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val. 2018. *Generic Machine Learning Inference on Heterogenous Treatment Effects in Randomized Experiments.* Working Paper, Working Paper Series 24678. National Bureau of Economic Research.

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.

Kotchen, Matthew J. 2017. "Longer-Run Evidence on Whether Building Energy Codes Reduce Residential Energy Consumption." *Journal of the Association of Environmental and Resource Economists* 4 (1): 135–153.

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

———. 2016b. "How Much Energy Do Building Energy Codes Save? Evidence from California Houses." *American Economic Review* 106 (10): 2867–94.

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.

Novan, Kevin, and Aaron Smith. 2018. "The Incentive to Overinvest in Energy Efficiency: Evidence from Hourly Smart-Meter Data." *Journal of the Association of Environmental and Resource Economists* 5 (3): 577–605.

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

	Untreated	Any intervention		HVAC interventions		Lighting interventions	
Characteristic		Treated	T-U	Treated	T-U	Treated	T-U
Hourly energy use (kWh)	33.3	57.4	24.0	62.9	29.6	61.0	27.7
	(34.4)	(72.8)	[<0.01]	(82.1)	[<0.01]	(87.2)	[<0.01]
First year in sample	2012	2010	-2	2009	-2	2010	-2
	(1.7)	(1.8)	[<0.01]	(1.6)	[<0.01]	(1.8)	[<0.01]
Total enrollment	541	729	188	766	225	746	205
	(365)	(488)	[<0.01]	(515)	[<0.01]	(522)	[<0.01]
Acad. perf. index (200-1000)	788	794	6	794	5	785	-3
	(99)	(89)	[0.21]	(89)	[0.32]	(79)	[0.56]
Bond passed, last 2 yrs (0/1)	0.3	0.2	-0.0	0.3	-0.0	0.2	-0.0
	(0.4)	(0.4)	[0.33]	(0.4)	[0.76]	(0.4)	[0.26]
Bond passed, last 5 yrs (0/1)	0.4	0.4	0.0	0.4	0.0	0.4	0.0
	(0.5)	(0.5)	[0.97]	(0.5)	[0.30]	(0.5)	[0.21]
High school graduates (%)	23.5	23.3	-0.2	23.7	0.2	24.2	0.8
	(12.2)	(11.6)	[0.76]	(11.7)	[0.74]	(10.6)	[0.26]
College graduates (%)	19.9	20.3	0.4	19.4	-0.5	19.4	-0.5
	(12.3)	(12.0)	[0.53]	(11.9)	[0.47]	(11.8)	[0.47]
Single mothers (%)	20.4	19.3	-1.1	19.8	-0.6	20.6	0.2
	(19.1)	(18.5)	[0.24]	(18.8)	[0.57]	(19.0)	[0.83]
African American (%)	5.6	6.1	0.5	5.4	-0.2	6.2	0.6
	(9.2)	(8.0)	[0.24]	(6.0)	[0.58]	(7.3)	[0.25]
Asian (%)	9.0	11.6	2.7	12.6	3.6	9.7	0.8
	(13.1)	(16.1)	[<0.01]	(17.1)	[<0.01]	(12.1)	[0.31]
Hispanic (%)	42.4	43.6	1.2	45.4	3.0	46.4	4.0
	(28.7)	(26.8)	[0.38]	(27.1)	[0.05]	(25.4)	[0.01]
White (%)	34.5	30.8	-3.7	29.9	-4.6	29.4	-5.1
	(26.9)	(24.4)	[<0.01]	(23.8)	[<0.01]	(24.2)	[<0.01]
Average temp. (° F)	60.1	60.8	0.8	61.3	1.3	61.0	1.0
	(4.1)	(3.5)	[<0.01]	(3.4)	[<0.01]	(3.6)	[<0.01]
Latitude	37.7	37.5	-0.2	37.4	-0.3	37.4	-0.2
	(1.2)	(1.0)	[<0.01]	(1.0)	[<0.01]	(1.1)	[<0.01]
Longitude	-121.6	-121.2	0.4	-121.0	0.6	-121.1	0.5
	(1.0)	(1.1)	[<0.01]	(1.1)	[<0.01]	(1.1)	[<0.01]
Number of schools	985	920		566		439	

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 hotter locations; and are located southeast of untreated schools.

Table 2: Panel fixed effects results

	(1)	(2)	(3)	(4)	(5)
Treat × post	-3.24	-3.24	-3.88	-2.24	-1.30
	(0.45)	(0.45)	(0.45)	(0.48)	(0.47)
Observations	19,193,084	19,193,084	19,192,744	19,192,744	19,193,084
Realization rate	0.73	0.73	0.87	0.50	0.29
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 hourly energy consumption in kWh (averaged across "blocks" of three hours) as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on energy consumption by the average expected savings in the sample.

Table 3: Correlation between data and predictions across machine learning methods

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
10th percentile	-0.23	0.08	-0.12	0.13	-0.08	0.12	0.09	-1.70
25th percentile	0.18	0.21	0.30	0.33	0.28	0.32	0.30	-0.15
50th percentile	0.44	0.43	0.62	0.64	0.61	0.63	0.52	0.42
75th percentile	0.61	0.58	0.86	0.86	0.85	0.85	0.67	0.63
90th percentile	0.72	0.69	0.93	0.93	0.93	0.93	0.76	0.71
Method	LASSO	LASSO	LASSO	LASSO	LASSO	LASSO	RF	RF
Block-specific model	X	X	X	X	X	X	X	
Basic variables	X	X	X	X			X	X
Untreated schools $-i$			X	X	X	X		
Tuning parameter	Min	1SE	Min	1SE	Min	1SE		

Notes: This table reports the R^2 of the prediction models for untreated schools during the post-treatment period. As that these predictions are completely out-of-sample, and therefore extreme outliers may be a concern, we present the distribution of the R^2. Columns 1 through 6 display predictions generated via LASSO, while Columns 7 and 8 show predictions generated using a random forest algorithm. In all but Column 8, we generate prediction models for each school-hour-block separately. The "basic variables" include day of the week, a holiday dummy, a seasonal spline, a temperature spline, and all of their their multi-way interactions. In Columns 3, 4, 5, and 6, we include energy consumption at all (other) untreated schools as candidate variables. For the LASSO estimates, we report results for two tuning parameters: "Min," which minimizes the root mean squared error, or "1SE," which chooses a slightly more parsimonious model than Min, but which has a root mean squared error that remains within one standard error of Min. Overall, we find that the LASSO model where we allow for both basic variables and untreated school consumption, with a 1SE tuning parameter, provides the best overall fit.

Table 4: Machine learning results

	(1)	(2)	(3)	(4)	(5)
Treat × post	-3.77	-3.79	-3.98	-3.13	-2.36
	(0.53)	(0.53)	(0.55)	(0.53)	(0.51)
Observations	19,193,084	19,193,084	19,192,744	19,192,744	19,193,084
Realization rate	0.85	0.85	0.89	0.70	0.53
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 prediction errors in hourly energy consumption – measured in kWh and averaged across "blocks" of three hours – as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on energy consumption by the average expected savings in the sample. All regressions include a control for being in the post-training period for the machine learning.

Table 5: Realization rate heterogeneity

	(1)	(2)	(3)	(4)	(5)	(6)
Any intervention	0.53	0.55	0.29	0.45	0.65	0.08
	(0.11)	(0.13)	(0.26)	(0.13)	(0.10)	(0.22)
HVAC interventions	0.75	0.69	1.04	0.40	0.72	0.29
	(0.36)	(0.40)	(0.33)	(0.22)	(0.14)	(0.17)
Lighting interventions	0.65	0.63	0.13	0.43	0.39	0.40
	(0.26)	(0.27)	(0.24)	(0.19)	(0.12)	(0.17)
Other interventions	0.42	0.46	-0.21	0.21	0.48	-0.25
	(0.23)	(0.28)	(0.52)	(0.13)	(0.31)	(0.34)
Observations	19,193,084	18,934,974	19,193,084	19,193,084	18,934,974	19,193,084
Savings regression				X	X	X
Expected savings trim		X			X	
Time-varying treatment			X			X

Notes: This table presents estimated realization rates for different intervention types, using several estimation procedures. The first panel presents results across all upgrades. The second panel displays realization rates for HVAC interventions, lighting interventions, and other interventions, estimated jointly. In the first three columns, we calculate realization rates by estimating a regression of prediction errors in energy consumption on a treatment indicator, equal to zero before any upgrade occurs and one otherwise, and school-by-hour-block and month-of-sample fixed effects, as in Equation (3.1). We include three separate treatment indicators - one for each intervention type - in the regression that generates the results in the bottom panel, as in Equation (4.2). We then divide the point estimates from this regression by the average expected savings (from this intervention type) in the sample, conditional on expected savings being greater than zero. We compute standard errors on the realization rates by scaling the regression standard error by the average expected savings. In Column (2), we repeat this exercise, but dropping schools with expected savings below the 1st or above the 99th percentile of the distribution. In Column (3), rather than a binary treatment indicator, we define the treatment variable as the number of upgrades (of a given category) that school i has installed by time t. In Column (4), we compute the realization rate by estimating Equation (4.4) - that is, we regress prediction errors in energy consumption on average expected and a school-by-hour-block and month-of-sample fixed effect, which recovers the correlation between school i's expected savings and school i's realized savings. In Column (4), we repeat this exercise, trimming on expected savings as in Column (2). Column (5) implements the same procedure, using time-varying measures of expected savings as the treatment variable(s) of interest. All standard errors are clustered at the school level.

Table 6: Machine learning results by size of expected savings

	(1)	(2)	(3)
Overall	0.53		
	(0.11)		
Below median		1.27	0.92
		(1.04)	(0.55)
Above median		0.53	0.47
		(0.10)	(0.09)
Observations	19,193,084	19,193,084	19,193,084
Median			
Absolute savings		X	
Relative savings			X

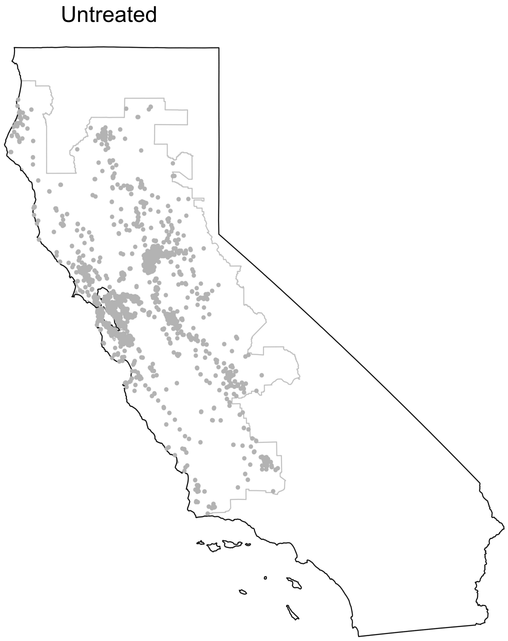
Notes: This table reports results from estimating Equation (3.1), with prediction errors in hourly energy consumption – measured in kWh and averaged across "blocks" of three hours – as the dependent variable. In Columns (2) and (3), we interact the treatment indicator with an indicator for whether a school's expected savings are above or below the median. Column (2) divides schools into above and below median in terms of total expected kWh saved, while Column (3) splits schools by kWh saved relative to average kWh consumed. All regressions include school-by-block and month-of-sample fixed effects. Realization rates are calculated by dividing the regression results on energy consumption by the average expected savings in the sample, separately for above and below median schools. Standard errors are calculated by dividing the regression standard errors (clustered at the school level) by the average expected savings in the sample, separately for above and below median schools. All regressions include a control for being in the post-training period for the machine learning.

Table 7: Predicting heterogeneous effects

Variable	(1)	(2)	(3)	(4)	(5)	(6)
Constant	0.80	0.61	0.63	0.93	0.99	1.11
	(0.12)	(0.23)	(0.22)	(0.21)	(0.22)	(0.23)
HVAC (0/1)		0.19	0.32	-0.03	0.12	0.10
		(0.24)	(0.23)	(0.23)	(0.24)	(0.25)
Lighting (0/1)		0.08	-0.05	-0.09	-0.25	-0.13
		(0.23)	(0.22)	(0.21)	(0.23)	(0.24)
Longitude			0.03	-0.06	-0.13	-0.24
			(0.25)	(0.25)	(0.27)	(0.28)
Latitude			0.29	0.32	0.33	0.23
			(0.18)	(0.18)	(0.19)	(0.20)
Average temperature (° F)			0.07	0.15	0.17	0.23
			(0.22)	(0.22)	(0.23)	(0.23)
Total enrollment				0.34	0.32	0.43
				(0.09)	(0.10)	(0.11)
Academic perf. index (200-1000)					-0.06	-0.14
					(0.13)	(0.14)
Poverty rate					0.00	-0.05
					(0.13)	(0.14)
Expected savings (kWh)						-0.20
						(0.08)
Number of schools	914	914	892	847	784	784

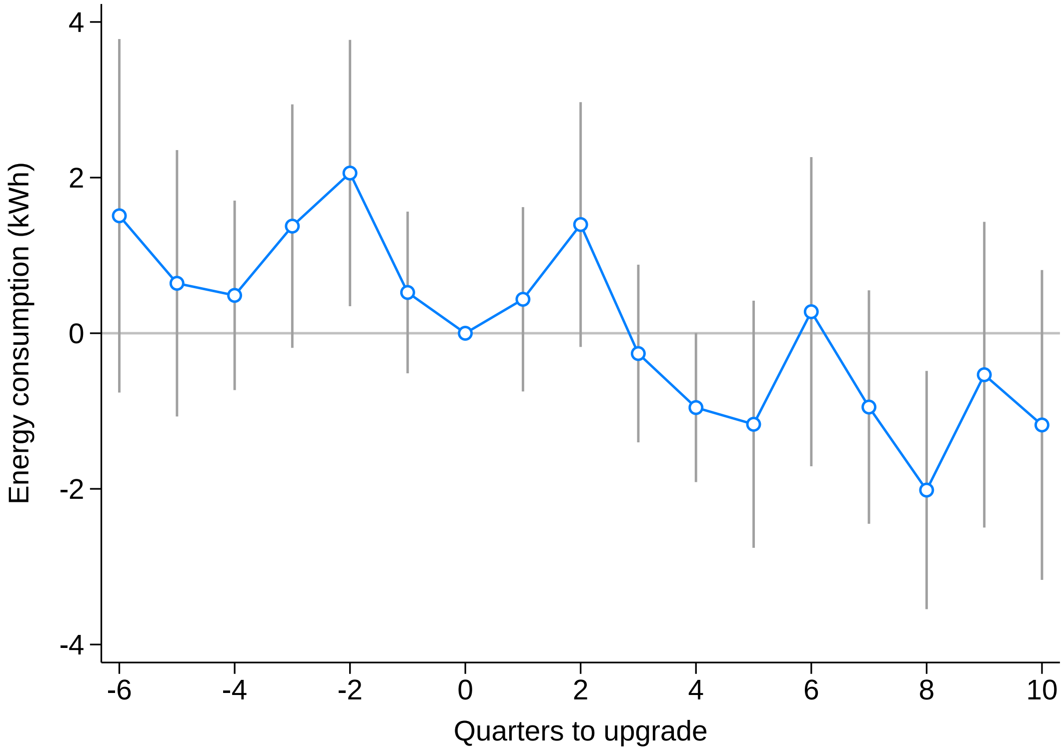
Notes: This table presents results from median regressions of school-specific realization rates on a variety of covariates. The school-specific realization rates are estimated from a regression of prediction errors (in kWh) on school-specific treatment indicators and school-by-block-by-month fixed effects. 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



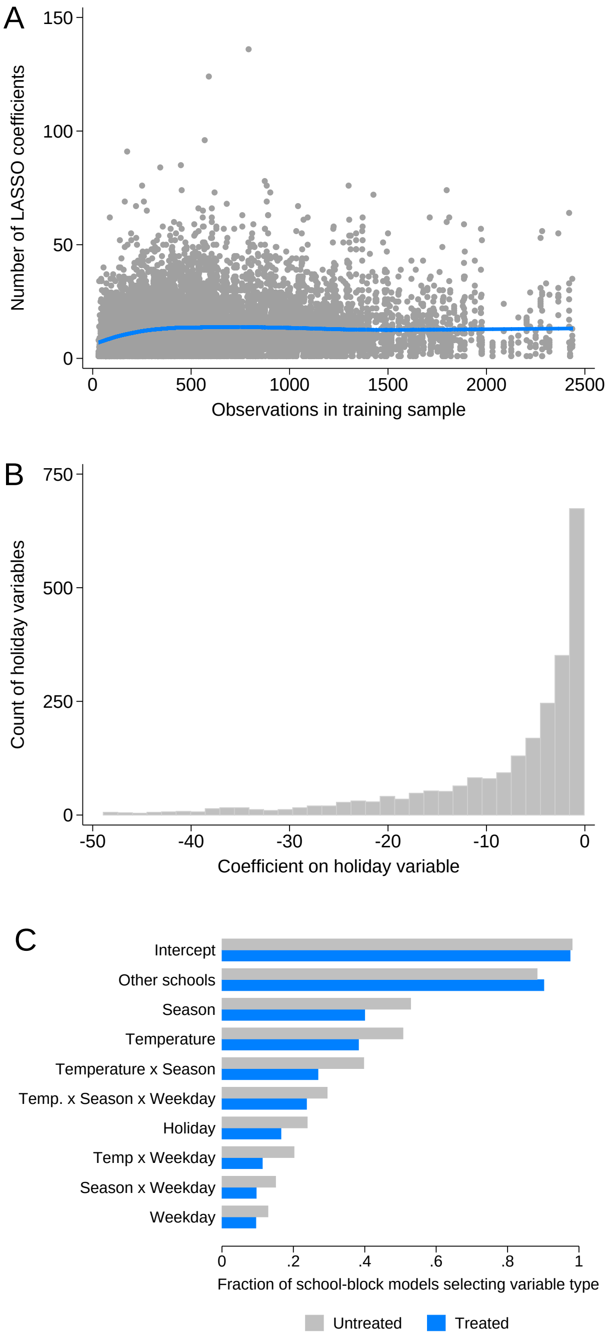
Notes: This figure displays the locations of schools in our sample. "Untreated" schools, in gray on the left, did not undertake any energy efficiency upgrades during our sample period. "Treated" schools, in blue on the right, installed at least one upgrade during our sample. There is substantial overlap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

Figure 2: Panel fixed effects event study



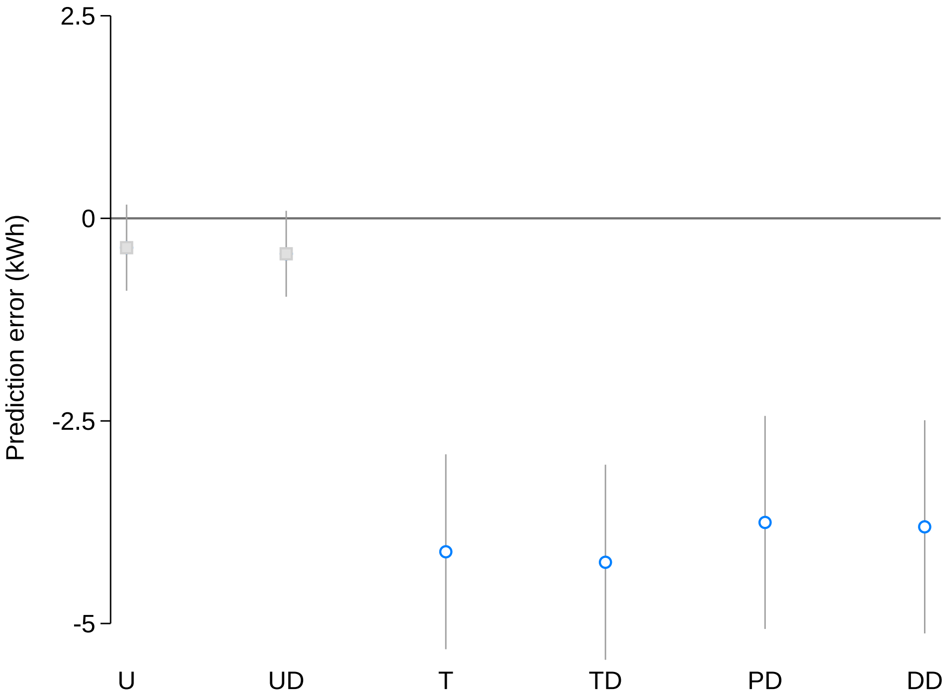
Notes: This figure displays point estimates and 95 percent confidence intervals from event study regressions of energy consumption before and after an energy efficiency upgrade. We estimate Equation (3.2) with hourly electricity consumption (in kWh, averaged by three hour block) as the dependent variable. 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 3: Machine learning diagnostics



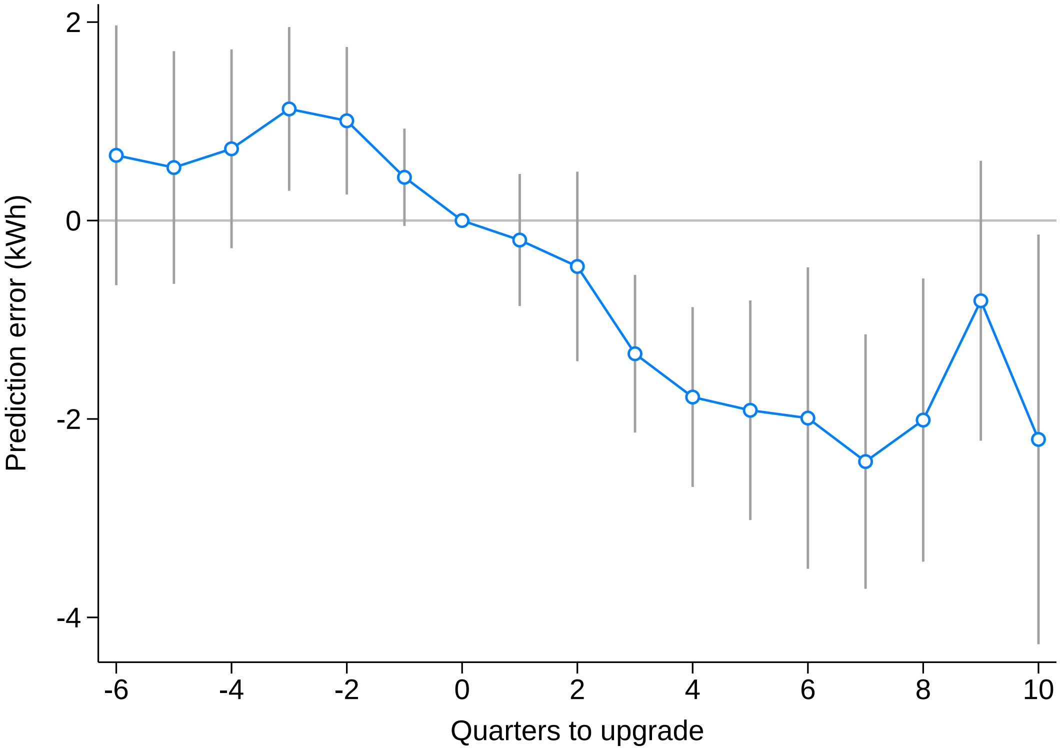
Notes: This figure presents three checks of our machine learning methodology. Panel A displays the relationship between the number of observations in the pre-treatment ("training") dataset and the number of variables LASSO selects to include in the prediction model for each school in the sample. Schools with very few training observations yield sparse models. As expected, the larger the training sample, the more flexible the prediction model becomes up to a point. This suggests that the LASSO is not overfitting, but that the underlying data generating process is relatively sparse, which is required for the LASSO to perform well. Panel B displays the marginal effect of holiday indicators in each school-specific prediction model. The majority of the coefficients on these models are negative and we do not observe large outliers, which suggests that the LASSO model is picking up patterns that we would expect to be present in the data and that will do well out of sample. Finally, Panel C displays the categories of variables selected by our preferred LASSO method for untreated and treated schools. Most models selected at least one untreated school's prediction for inclusion in the model.

Figure 4: Comparing machine learning estimators



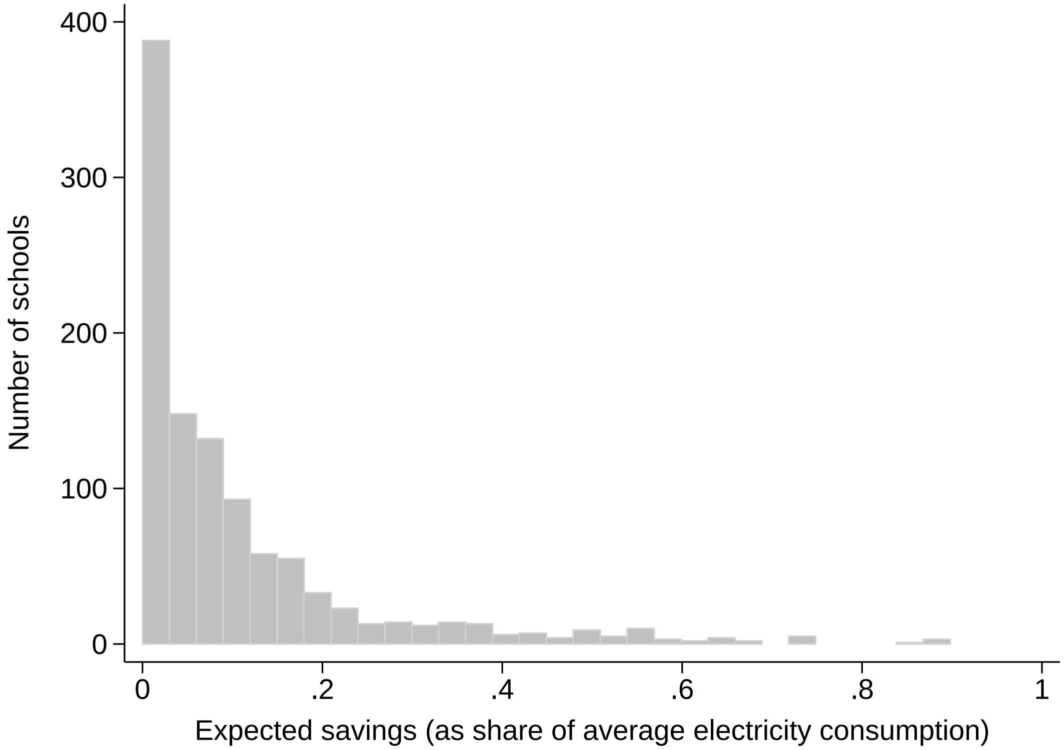
Notes: This figure shows average treatment effects, in the form of prediction errors based on hourly electricity consumption in kWh (averaged across three-hour ”blocks”), from a variety of different machine learning estimators. The effect marked U shows prediction errors (real energy consumption minus 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 5: Machine learning event study



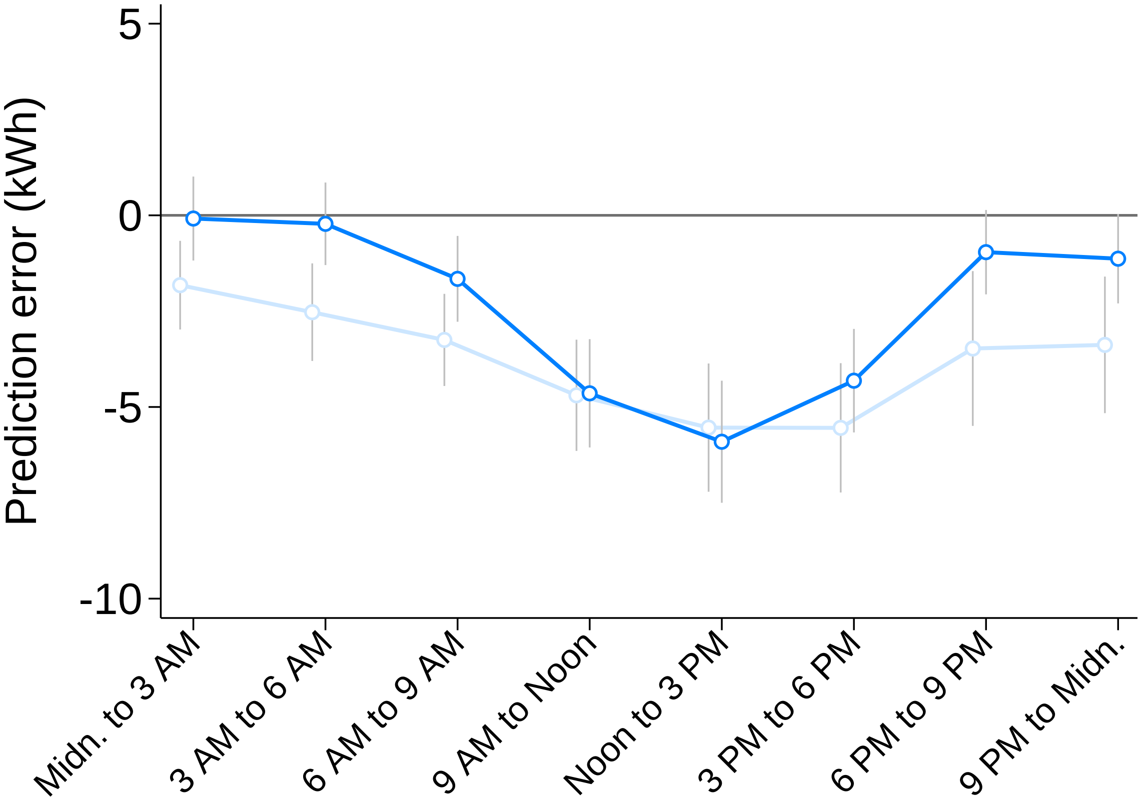
Notes: This figure displays point estimates and 95 percent confidence intervals from event study regressions of energy consumption before and after an energy efficiency upgrade. We estimate Equation (3.2) with prediction errors based on electricity consumption in kWh (averaged across three-hour "blocks") 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. Unlike the regression estimates displayed in Figure 2, there is a clear change in energy consumption after the installation of energy efficiency upgrades, which persists more than a year after the upgrade.

Figure 6: Expected savings relative to consumption



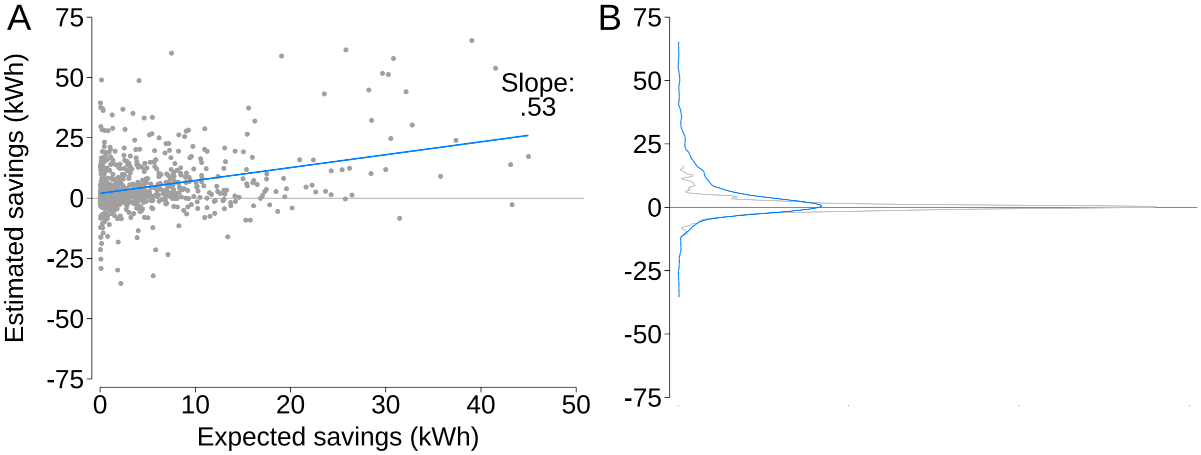
Notes: This figure shows the distribution of expected savings relative to average electricity consumption among the treated schools in our sample. While the median school is expected to reduce its energy consumption by 6 percent of its average consumption, the mean school has expected savings equal to 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 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 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 and month-of-sample fixed effects. Standard errors are clustered by school.

Figure 8: 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. 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 0.53 (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's distribution is narrow and centered around zero, the treated group appears shifted towards more savings.

Appendix: For online publication

A Supplemental tables and figures

Table A.1: Panel fixed effects results (alternative standard errors)

Clustering	(1)	(2)	(3)	(4)	(5)
	-3.24	-3.24	-3.88	-2.24	-1.30
School	(0.45)	(0.45)	(0.45)	(0.48)	(0.47)
School, month of sample	[1.58]	[1.58]	[0.72]	[0.53]	[0.52]
Observations	19,193,084	19,193,084	19,192,744	19,192,744	19,193,084
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 hourly energy consumption in levels (averaged across "blocks" of three hours) as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. This table shows two variations on clustered standard errors: errors clustered at the school level, as in the main text, in parentheses; and errors clustered at the school and month-of-sample level, in brackets.

Table A.2: Matching results

	(1)	(2)	(3)	(4)	(5)
Any district	-2.79	-2.70	-2.99	-0.66	-0.31
	(0.93)	(0.93)	(0.98)	(1.07)	(1.01)
Same district	-0.23	-0.17	-0.40	1.14	0.97
	(0.82)	(0.83)	(0.82)	(0.86)	(0.79)
Opposite district	-3.55	-3.55	-3.64	-0.44	-0.16
	(0.94)	(0.94)	(1.01)	(1.12)	(1.05)
Observations	6,043,052	6,043,046	6,042,653	6,042,653	6,043,046
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 hourly energy consumption in kWh (averaged across "blocks" of three hours) as the dependent variable. As above, the independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. 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.

Table A.3: Panel fixed effects results (trimming)

	(1)	(2)	(3)
Treat × post	-1.30	-0.37	-0.18
	(0.47)	(0.36)	(0.33)
Observations	19,193,084	18,808,392	18,425,268
Realization rate	0.29	0.09	0.04
Trimming			
Dependent variable (1, 99)		X	
Dependent variable (2, 98)			X

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in levels (averaged across "blocks" of three hours) as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. All regressions include school-by-block and month-of-sample fixed effects. This table presents three types of trimming of the dependent variable: Column (1) does not trim at all; Column (2) trims the sample to exclude observations below the 1st or above the 99th percentile, as in the main text; and Column (3) trims the sample to exclude observations below the 2nd or above the 98th percentile.

Table A.4: Machine learning results (alternative standard errors)

Clustering	(1)	(2)	(3)	(4)	(5)
	-3.77	-3.79	-3.98	-3.13	-2.36
School	(0.53)	(0.53)	(0.55)	(0.53)	(0.51)
School, month of sample	[0.66]	[0.66]	[0.64]	[0.56]	[0.52]
Observations	19,193,084	19,193,084	19,192,744	19,192,744	19,193,084
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 prediction errors in hourly energy consumption in levels (averaged across "blocks" of three hours) as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. This table shows two variations on clustered standard errors: errors clustered at the school level, as in the main text, in parentheses; and errors clustered at the school and month-of-sample level, in brackets. All regressions include a control for being in the post-training period for the machine learning.

Table A.5: Machine learning results (alternative prediction methods)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treat × post	-2.07	-1.95	-2.86	-2.36	-2.33	-2.24	-2.06	-2.02
	(0.59)	(0.51)	(0.60)	(0.51)	(0.53)	(0.51)	(0.51)	(0.51)
Realization rate	0.46	0.44	0.64	0.53	0.52	0.50	0.46	0.45
Method	LASSO	LASSO	LASSO	LASSO	LASSO	LASSO	RF	RF
Block-specific model	X	X	X	X	X	X	X	
Basic variables	X	X	X	X			X	X
Untreated schools $-i$			X	X	X	X		
Tuning parameter	Min	1SE	Min	1SE	Min	1SE		

Notes: This table reports results from estimating Equation (3.1), with prediction errors in hourly energy consumption in levels (averaged across "blocks" of three hours) as the dependent variable. All regressions include school-by-block and month-of-sample fixed effects. Each column displays results from a different prediction approach. Columns 1 through 6 display predictions generated via LASSO, while Columns 7 and 8 show predictions generated using a random forest algorithm. In all but Column 8, we generate prediction models for each school-hour-block separately. The "basic variables" include day of the week, a holiday dummy, a seasonal spline, a temperature spline, and all of their their multi-way interactions. In Columns 3, 4, 5, and 6, we include energy consumption at all (other) untreated schools as candidate variables. For the LASSO estimates, we report results for two tuning parameters: "Min," which minimizes the root mean squared error, or "1SE," which chooses a slightly more parsimonious model than Min, but which has a root mean squared error that remains within one standard error of Min. In all cases, the independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. All regressions include a control for being in the post-training period for the machine learning, and standard errors are clustered at the school level.

Table A.6: Machine learning results: Alternative estimation methods

	(1)	(2)	(3)	(4)	(5)
Trained on pre	-3.77	-3.79	-3.98	-3.13	-2.36
	(0.53)	(0.53)	(0.55)	(0.53)	(0.51)
Realization rate	0.85	0.85	0.89	0.70	0.53
Observations	19,193,084	19,193,084	19,192,744	19,192,744	19,193,084
Trained on post	-3.63	-3.60	-3.73	-2.72	-1.96
	(0.47)	(0.47)	(0.50)	(0.47)	(0.45)
Realization rate	0.82	0.81	0.84	0.61	0.44
Observations	19,998,558	19,998,556	19,998,180	19,998,180	19,998,556
Pooled	-3.70	-3.69	-3.85	-2.92	-2.15
	(0.49)	(0.49)	(0.52)	(0.49)	(0.46)
Realization rate	0.83	0.83	0.86	0.66	0.48
Observations	39,191,644	39,191,640	39,190,920	39,190,920	39,191,640
Double LASSO	-5.47	-5.46	-5.61	-4.15	-2.50
	(0.61)	(0.61)	(0.62)	(0.59)	(0.57)
Realization rate	1.23	1.23	1.26	0.93	0.56
Observations	19,076,036	19,076,036	19,075,702	19,075,702	19,076,036
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: In this table, we present four variations on our prediction procedure. The “Trained on pre” panel presents results where the pre-treatment period is used to train the machine learning model, which we then forecast into the post-treatment periods to estimate treatment effects. In the “Trained on post” panel, we reverse this procedure, training the model on the post-treatment period, and projecting into the pre-treatment period (scaled such that treatment effects have the same sign). The “Pooled” panel uses both trained-on-pre and trained-on-post predictions. Finally, we report results from a variant on the “double LASSO” procedure described by Belloni, Chernozhukov, and Hansen (2014), allowing for proper inference with model selection, which we adapt to the panel setting. We first estimate a LASSO to predict the timing of treatment, next estimate a second LASSO to predict electricity consumption, and finally estimate a third LASSO with time as the dependent variable, in order to accommodate trends. We then regress energy consumption on treatment timing and the union of the non-zero-coefficient variables from all three LASSOs. To make this computationally tractable, we apply the selection of variables on a school-by-school basis. Then we residualize each dependent variable by the full set of controls, and implement the final step, with all schools pooled, by regressing residualized prediction errors on residualized treatment date error and residualized time error. These procedures are mathematically equivalent, via Frisch-Waugh-Lovell. In all machine learning procedures, the independent variables are treatment indicators, equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors are clustered at the school level. All regressions include a control for being in the post-training period for the machine learning, with the exception of the double LASSO, for which this is not appropriate.

Table A.7: Machine learning results (trimming)

	(1)	(2)	(3)
Treat × post	-2.36	-2.33	-2.25
	(0.51)	(0.32)	(0.26)
Observations	19,193,084	18,809,216	18,425,356
Realization rate	0.53	0.55	0.54
Trimming			
Dependent variable (1, 99)		X	
Dependent variable (2, 98)			X

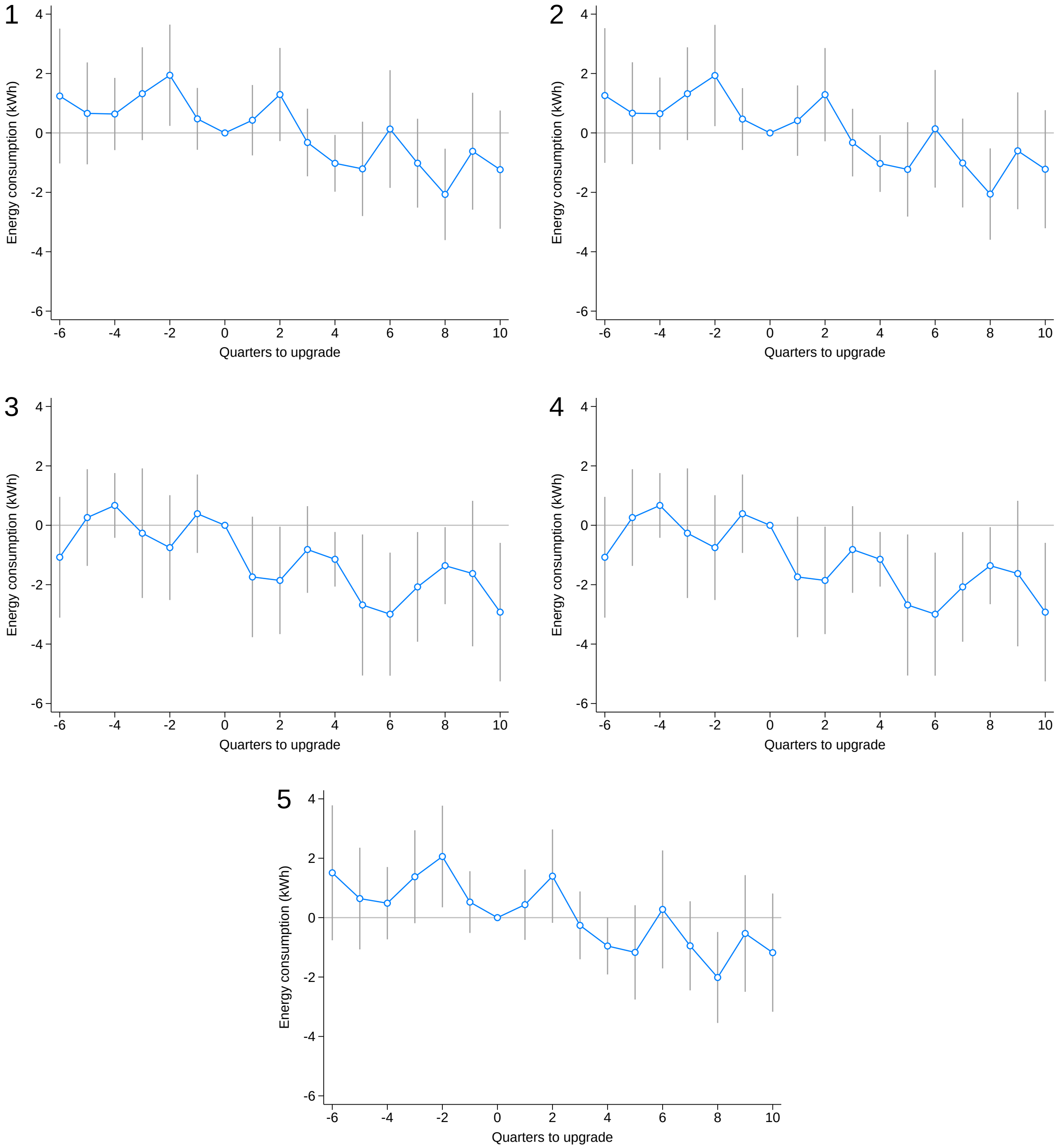
Notes: This table reports results from estimating Equation (3.1), with prediction errors in hourly energy consumption in levels (averaged across "blocks" of three hours) as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. All regressions include school-by-block and month-of-sample fixed effects. This table presents three types of trimming of the dependent variable: Column (1) does not trim at all; Column (2) trims the sample to exclude observations below the 1st or above the 99th percentile, as in the main text; and Column (3) trims the sample to exclude observations below the 2nd or above the 98th percentile. All regressions include a control for being in the post-training period for the machine learning.

Table A.8: Realization rate heterogeneity (Panel fixed effects)

	(1)	(2)	(3)	(4)	(5)	(6)
Any intervention	0.29	0.27	0.20	0.41	0.54	0.04
	(0.10)	(0.11)	(0.25)	(0.12)	(0.09)	(0.22)
HVAC interventions	0.33	0.23	0.83	0.36	0.56	0.14
	(0.35)	(0.39)	(0.32)	(0.16)	(0.12)	(0.15)
Lighting interventions	0.55	0.51	0.05	0.42	0.37	0.37
	(0.25)	(0.26)	(0.25)	(0.18)	(0.10)	(0.17)
Other interventions	0.27	0.25	-0.20	0.23	0.48	-0.23
	(0.22)	(0.27)	(0.53)	(0.14)	(0.30)	(0.35)
Observations	19,193,084	18,934,974	19,193,084	19,193,084	18,934,974	19,193,084
Savings regression				X	X	X
Expected savings trim		X			X	
Time-varying treatment			X			X

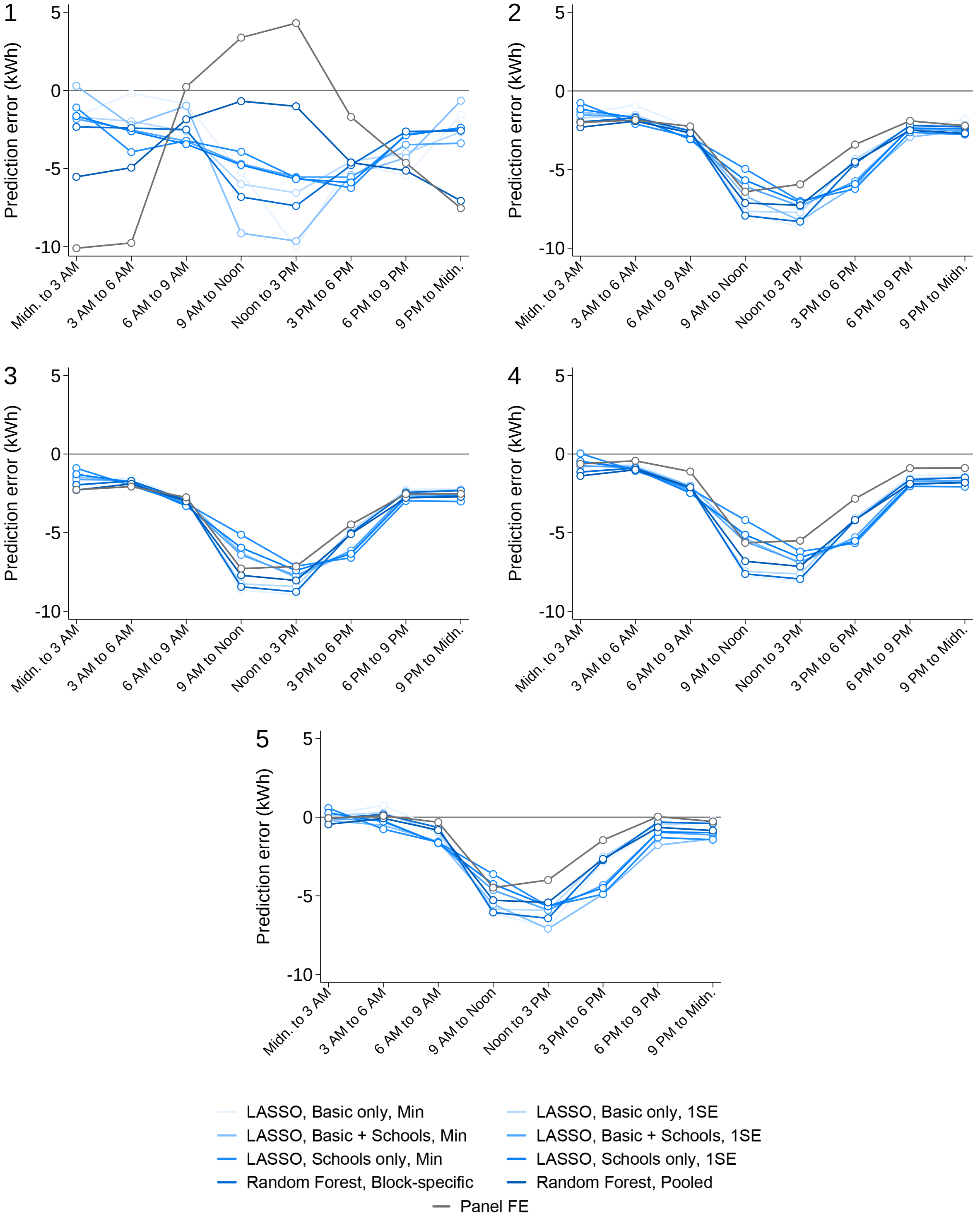
Notes: This table presents estimated realization rates for different intervention types, using several estimation procedures. The first panel presents results across all upgrades. The second panel displays realization rates for HVAC interventions, lighting interventions, and other interventions, estimated jointly. In the first three columns, we calculate realization rates by estimating a regression of energy consumption on a treatment indicator, equal to zero before any upgrade occurs and one otherwise, and school-by-hour-block and month-of-sample fixed effects, as in Equation (3.1). We include three separate treatment indicators - one for each intervention type - in the regression that generates the results in the bottom panel, as in Equation (4.2). We then divide the point estimates from this regression by the average expected savings (from this intervention type) in the sample, conditional on expected savings being greater than zero. We compute standard errors on the realization rates by scaling the regression standard error by the average expected savings. In Column (2), we repeat this exercise, but dropping schools with expected savings below the 1st or above the 99th percentile of the distribution. In Column (3), rather than a binary treatment indicator, we define the treatment variable as the number of upgrades (of a given category) that school i has installed by time t. In Column (4), we compute the realization rate by estimating Equation (4.4) - that is, we regress energy consumption on average expected and a school-by-hour-block and month-of-sample fixed effect, which recovers the correlation between school i's expected savings and school i's realized savings. In Column (4), we repeat this exercise, trimming on expected savings as in Column (2). Column (5) implements the same procedure, using time-varying measures of expected savings as the treatment variable(s) of interest. All standard errors are clustered at the school level.

Figure A.1: Panel fixed effects event study – all specifications



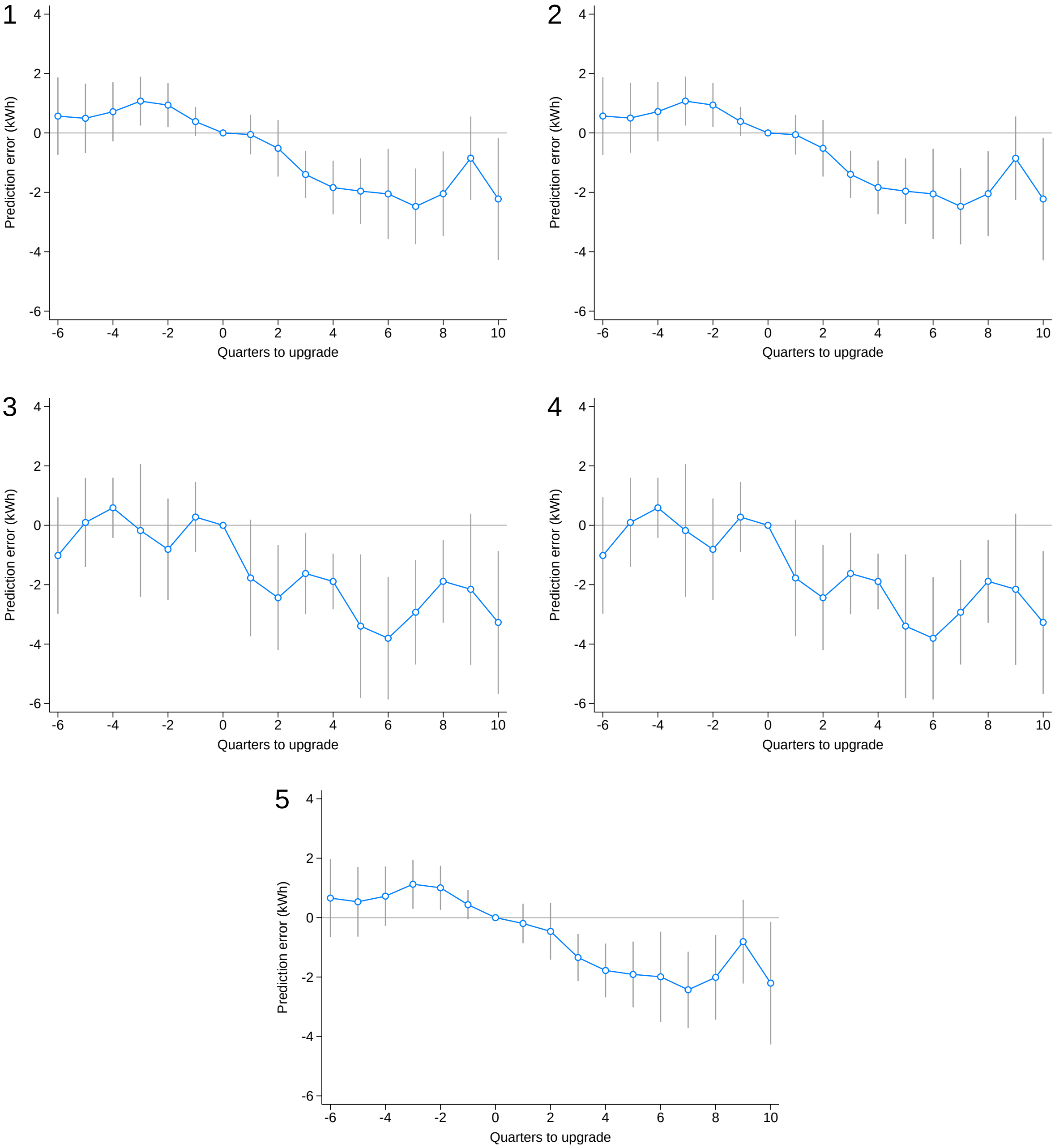
Notes: This figure displays point estimates and 95 percent confidence intervals from event study regressions of energy consumption before and after an energy efficiency upgrade. We estimate Equation (3.2) with hourly electricity consumption (in kWh, averaged by three hour block) as the dependent variable. We normalize time relative to the quarter each school undertook its first upgrade. Each panel corresponds to the like-numbered column of Table 2, and includes both treated and untreated schools. Standard errors are clustered by school. 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 A.2: Machine learning results by hour-block (alternative prediction methods)



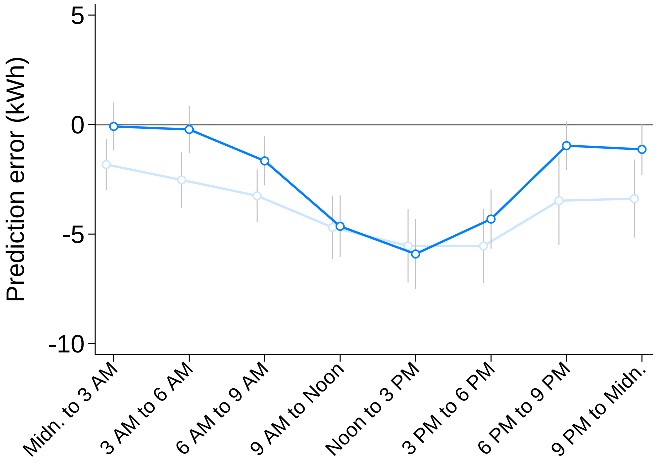
Notes: This figure presents treatment effects for each three-hour block of the day estimated using prediction errors based on electricity consumption in kWh (averaged across three-hour "blocks") as the dependent variable. Here, we present results from 9 different estimation procedures: LASSOs with, without, and exclusively using other schools' consumption as candidate variables using a larger and smaller tuning parameter; random forests with and without imposing hour-block-specific branches; and the panel fixed effects analogue. Each panel corresponds to one column of Table 4.

Figure A.3: Machine learning event study – all specifications



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 prediction errors in hourly 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. Each panel corresponds to the like-numbered column of Table 4, and includes both treated and untreated schools. Standard errors are clustered by school. Unlike the regression estimates displayed in Figure 2, there is a clear change in energy consumption after the installation of energy efficiency upgrades, which persists more than a year after the upgrade.

Figure A.4: 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 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. Standard errors are clustered by school.

MIT CEEPR

MIT Center for Energy and Environmental Policy Research

Since 1977, the Center for Energy and Environmental Policy Research (CEEPR) has been a focal point for research on energy and environmental policy at MIT. CEEPR promotes rigorous, objective research for improved decision making in government and the private sector, and secures the relevance of its work through close cooperation with industry partners from around the globe. Drawing on the unparalleled resources available at MIT, affiliated faculty and research staff as well as international research associates contribute to the empirical study of a wide range of policy issues related to energy supply, energy demand, and the environment.

An important dissemination channel for these research efforts is the MIT CEEPR Working Paper series. CEEPR releases Working Papers written by researchers from MIT and other academic institutions in order to enable timely consideration and reaction to energy and environmental policy research, but does not conduct a selection process or peer review prior to posting. CEEPR's posting of a Working Paper, therefore, does not constitute an endorsement of the accuracy or merit of the Working Paper. If you have questions about a particular Working Paper, please contact the authors or their home institutions.

MIT Center for Energy and Environmental Policy Research
77 Massachusetts Avenue, E19-411
Cambridge, MA 02139
USA

Website: ceepr.mit.edu

MIT CEEPR Working Paper Series is published by the MIT Center for Energy and Environmental Policy Research from submissions by affiliated researchers.

Copyright © 2019
Massachusetts Institute of Technology

For inquiries and/or for permission to reproduce material in this working paper, please contact:

Email ceepr@mit.edu
Phone (617) 253-3551
Fax (617) 253-9845