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 19, 2020

Abstract

We use high-frequency panel data on electricity consumption to study the effectiveness of energy efficiency upgrades in K-12 schools in California. Using a panel fixed effects approach, we find that these upgrades deliver between 12 and 86 percent of expected savings, depending on specification and treatment of outliers. Using machine learning to inform our specification choice, we estimate a narrower range: 52 to 98 percent, with a central estimate of 60 percent. These results imply that upgrades are performing less well than *ex ante* predictions on average, although we can reject some of the very low realization rates found in prior work.

JEL Codes: Q4, Q5, C4

Keywords: energy efficiency; machine learning; schools.

*Burlig: Harris School of Public Policy and Energy Policy Institute, University of Chicago and NBER, 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 several venues for helpful comments. Joshua Blonz and Kat Redoglio provided 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 (McKinsey & Company (2009); Tonn et al. (2014), Nadel and Ungar (2019)). 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); Gerarden et al. (2017)). Among other explanations, economists have suggested the possibility of market failures and behavioral biases (Fowlie et al. (2018)).

A second strand of literature seeks to better understand the real-world savings and costs of energy efficiency investments when compared to engineering projections. There are a variety of reasons why engineering estimates might overstate savings or understate the costs consumers face.[1] Economists have 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 et al. (2018), Allcott and Greenstone (2017)) and quasi-experimental (e.g., Levinson (2016a), Myers (2015), and Davis et al. (2014)) approaches to developing this counterfactual. These

[1]For example, engineering models do not take consumer behavior into account. If an energy efficiency upgrade lowers the effective price of energy services and consumers respond by demanding more energy services, the energy efficiency upgrade will look less effective than the engineering prediction – even if this prediction would have been correct in the absence of behavior change. Furthermore, *ex post* evaluation is relatively uncommon in the energy efficiency industry, so there is limited feedback between real-world outcomes and engineering models (Fowlie et al. (2018)).

studies, all of which estimate the effectiveness of energy efficiency upgrades in residential settings, find substantial underperformance, with upgrades delivering between 25 and 58 percent of *ex ante* expected savings.[2]

A more complete view of which energy efficiency opportunities are cost-effective requires more evidence from a variety of settings. While 37 percent of electricity use in the United States in 2014 was residential, over half is attributable to commercial and industrial uses (Energy Information Administration (2015)). Despite the large role of non-household sectors in energy use, however, the existing literature is largely focused on residential energy efficiency (Gerarden et al. (2017)).[3] We extend this work into a non-residential sector by estimating the impacts of energy efficiency upgrades in K-12 schools in California from 2008 to 2014. We match 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, leveraging high-frequency electricity consumption data generated from advanced metering infrastructure ("smart metering").[4]

We estimate two empirical models. The first is a panel data model that uses a rich set of fixed effects and controls to non-parametrically separate the causal effect of energy efficiency upgrades from other confounding factors. We find evidence that our panel fixed effects approach is sensitive to outliers and to specification. However, choosing the "correct" set of controls is difficult as there are many possible candidate covariates, especially once we allow for interactions between control variables and unit or time fixed effects. For example, one might want to allow flexible functions of temperature to impact electricity consumption in a granular manner that is specific to each school. The richness of the data makes it difficult for researchers to choose between a large set of plausible regression models both from a conceptual and computational point of view.

To overcome these challenges, we estimate a second empirical model based on new tech-

[2]An interesting counterpoint to these estimates is Blonz (2019), who finds that refrigerator replacements in Southern California delivered savings in line with engineering estimates when these upgrades were implemented according to program rules.

[3]A notable exception is Ryan (2018), who studies energy audits in Indian manufacturing firms, and finds evidence of substantial rebound: treated firms use 9.5 percent more electricity.

[4]Over 52 percent of US households had smart meters as of 2018 according to data from EIA Form-861.

niques in machine learning. We combine our high-frequency electricity consumption data with machine learning methods in order to select among the set of possible covariates in a disciplined and computationally feasible manner.[5] In particular, we use each individual school's pre-treatment data only to build a machine learning model of that school's energy consumption. We use LASSO, as well as a set of alternative algorithms, to flexibly build these prediction models while avoiding overfitting. We then use each school's model to forecast counterfactual energy consumption in the post-treatment period. These models provide us with a prediction of what would have happened in the absence of any energy efficiency investments in a flexible, data-driven way, allowing us to control parsimoniously for school-specific heterogeneity while enabling systematic model selection. In our setting, this allows us to algorithmically choose between over 12 million possible covariates. in a disciplined, computationally-feasible way. In order to account for common shocks, we then embed these school-by-school counterfactuals in a panel fixed effects model to estimate causal effects.

The identifying assumption for the standard panel fixed effects model and our machine learning augmented version are essentially the same.[6] Conditional on a chosen set of controls, treated schools would have continued on a parallel trajectory to untreated schools in the absence of treatment. We provide evidence in support of these assumptions by demonstrating that treated and untreated schools do not exhibit differential trends in school characteristics, and by showing that there is a trend break among treated schools at the time of treatment. The key difference is that our machine learning framework allows us to select a richer set of control variables in a systematic and computationally tractable manner.

Using our machine learning method, we find that energy efficiency investments installed in California's K-12 schools underperform relative to average *ex ante* engineering projections of expected savings, delivering approximately 60 percent of expected savings. Comparing our machine learning approach to standard panel fixed effects approaches yields two primary

[5]Machine learning methods are increasingly popular in economics and other social sciences. They have been used to predict poverty and wealth (Blumenstock et al. (2015), Engstrom et al. (2016), Jean et al. (2016)), improve municipal efficiency (Glaeser et al. (2016)), understand perceptions about urban safety (Naik et al. (2015)), improve judicial decisions to reduce crime (Kleinberg et al. (2017)), examine heterogeneous treatments in electricity critical peak pricing (Prest (2020)) and weatherization (Souza (2019)), and more.

[6]Varian (2016) provides an overview of causal inference targeted at scholars familiar with machine learning.

findings. First, we show that estimates from standard panel fixed effects approaches are quite sensitive to specification, outliers, and the set of untreated schools we include in our models, with estimated energy savings ranging from 12 to 86 percent of *ex ante* expectations. By contrast, our machine learning method yields estimates that are substantially more stable across specifications and samples: we estimate savings between 52 and 98 percent of *ex ante* expectations. In addition to enabling data-driven covariate choice, these results highlight another potential benefit of using machine learning.

We also explore the extent to which we are able to predict realization rates using easily-observable school and upgrade characteristics. We do not find statistically significant correlations between these observables and realization rates in this setting. With more extensive data collection, larger samples, or in contexts where the signal-to-noise ratio is stronger, policymakers may be able to make progress towards identifying schools where upgrades are more effective.

The remainder of this paper proceeds by describing our empirical setting and data (Section 2). We then describe the baseline panel fixed effects 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. We compare approaches in Section 3.3. In Section 4, we explore heterogeneity in realizations rates. Section 5 concludes.

2 Context and data

Unlike most of the existing literature, which focuses on the residential setting, our paper evaluates the effectiveness of energy efficiency upgrades in public buildings: K-12 schools. School buildings, 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.

We analyze schools that participated in Pacific Gas and Electric Company's (PG&E's) energy efficiency programs. School districts identified opportunities for improvements at

their schools and then applied to PG&E for rebates to help cover the costs of qualifying investments. In California, utility energy efficiency programs are funded by a small adder on electricity and gas customer bills, which provides over $1 billion per year for programs across the residential, commercial and industrial sectors.

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 spent on energy efficiency over the same time period (Barbose et al. (2013)). The upgrades we study in this paper largely predate the investments financed through Proposition 39, but are similar to the later projects, making our results 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 et al. (2014) provide descriptive evidence on differences in energy consumption across one utility's commercial buildings as a function of various observables, including incentives embedded in the occupants' leases, age, and other physical attributes of the buildings. In other work, Kok and co-authors analyze the financial returns to energy efficiency attributes, though many of the attributes were part of the building's original construction and not part of deliberate

retrofits, which are the focus of our work (Kok and Jennen (2012); Eichholtz et al. (2013)).

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. Recent studies explore the advantages of automating energy efficiency evaluations exploiting the richness of smart meter data and highlight the potential for the use of machine learning in this area (Granderson et al. (2017)). In this paper, we implement one of the first quasi-experimental evaluations of energy efficiency upgrades outside the residential sector. Our results indicate that energy efficiency upgrades performed better in this context than in the residential settings of prior work, but we can still reject 100% realization rates using our preferred approach.[7]

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.

Smart meter data We obtain hourly interval electricity metering data for the universe of public K-12 schools in Northern California served by PG&E. The data begin in January

[7]The results remain on the upper end of realization rates in the literature (with a few exceptions finding support for full realization rates), and it is plausible that these relatively higher realization rates are driven by differences between the K-12 schools of our study and the residential settings of prior work. It is possible that schools are less subject to rebound than households. Schools are likely to be limited in their ability to change their usage, including facing regulations requiring them to provide adequate environmental conditions for education. In addition, school districts, rather than individual schools, receive and are responsible for paying electricity bills, which could make schools less sensitive to price than households. On the other hand, results in schools may differ from other commercial settings. Schools are not profit-maximizing. Deferred maintenance and rebound effects, for instance, may lead energy efficiency upgrades in schools to perform less well than in other commercial contexts.

2008, or the first month after the school's smart meter was installed, whichever comes later.[8] 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 collaboration 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. 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.[9] 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. To avoid potential concerns about these clusters, we use only school-meter matches that did not need to be aggregated.[10] Our final sample includes 1,931 schools.

Energy efficiency upgrades The PG&E data include energy efficiency upgrades for districts that applied for utility rebates. The upgrades in our database are likely to comprise a majority of energy efficiency upgrades undertaken by schools, which we discuss in Section 2.4.[11] 2,484 upgrades occurred at 911 schools between January 2008 and December 2014. For each energy efficiency installation, our data include the measure code, the measure

[8]The raw PG&E interval data recorded consumption information every 15 minutes; we collapse these data to the hourly level because 15-minute level intervals are often missing. Similarly, we interpolate consumption at a given hour if consumption at no more than two consecutive hours is missing.

[9]To ensure the quality of the match, we dropped schools for which energy consumption seemed implausibly low as a precaution (less than 1.5 kWh per hour). When we looked at these matches by hand, we found that such meters were not necessarily mismatched, but they were matched to non-standard educational centers with extremely low consumption, such as centers for adult learning with limited opening hours. We also checked for match quality among high-consuming outliers, but did not find evidence of imperfect matches.

[10]In early work, we found that results including clustered meters were very similar.

[11]Anecdotally, PG&E reports making concerted marketing efforts to reach out to districts to induce them to make these investments; districts often lack funds to devote to energy efficiency upgrades in the absence of such rebates.

description[12], a technology family (e.g., "HVAC", "Lighting"), number of units installed, installation date, 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.[13] Expected savings are determined through a standardized database, the Database of Energy Efficient Resources (DEER).[14] Before measures enter this database, they undergo extensive review by the California Public Utilities Commission.[15] Many schools undertake multiple upgrades, either within or across categories.

Figure 1 displays summary statistics for the energy efficiency upgrades in our sample. HVAC and lighting upgrades make up the bulk of the installed upgrades, both by number and by expected savings. The mean expected savings per intervention is 15,321 kWh per year, with a median of 3,344 kWh per year. The mean expected savings from HVAC (lighting) interventions is 10,290 kWh (23,673) per year, with a median of 1,640 (10,742) kWh per year. We attempt to normalize these savings per installed unit (eg, 1 light bulb), although there is substantial heterogeneity within these categories. We find that median expected savings per unit installed for HVAC (lighting) are in the order of approximately 77 (88) kWh per year.[16] For the median school, expected savings are less than 6 percent of annual consumption. However, in some schools, expected savings appear to be unreasonably high.[17]

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

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

[14]DEER is based on a DOE model that produces common energy savings estimates by upgrade type. Expected savings in DEER are based on a set of baseline model assumptions on pre-upgrade building characteristics and a measure case, designed to predict energy use with a particular energy efficiency measure installed. While some measures used by PG&E are tailored to K-12 schools, the majority of measures in our data use generic commercial buildings as the base case. Expected savings in DEER do vary with building vintage and climate zones in principle, though we see little variation by climate zone in our data. Expected savings are independent of previously installed upgrades, and scale linearly with number of units installed.

[15]Utilities have limited scope to adjust the savings estimates, which must be externally validated, and face dynamic incentives to avoid inflating savings estimates. Since energy efficiency upgrades are ultimately financed by ratepayers, proposed measures are the subject of considerable scrutiny by the Public Utilities Commission.

[16]For lighting, this is equivalent to an 8-watts saving per hour for a bulb running eight hours per day all year. For HVAC, this is a relatively small number, reflective the fact that many of the HVAC measures are limited to tune-ups to increase efficiency.

[17]These outliers are most likely due to measurement error either in the match between savings and school electricity consumption or in the expected savings themselves. We show that our machine learning approach below is robust to removing outliers in expected savings.

Other data We also obtain 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 obtain information on whether school district voters had approved facilities bonds in the two to five years before retrofits began at treated schools.[18]

We download 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.[19] We match 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 for our entire sample period, for schools with and without energy efficiency projects. We construct the main variables for each school as the average during the whole sample period. Of the 1,931 schools in the sample, 910 undertook at least one energy efficiency upgrade. There are 1,021 "untreated" schools that did not install any energy efficiency upgrades during our sample period. Our main variable of interest is hourly electricity consumption. 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.[20]

Table 1 highlights measurable differences between treated and untreated schools. Treated schools consume substantially more electricity, appear in our sample earlier, are larger, and

[18] Bond data are from EdSource (www.edsource.org), a non-profit education journalism website.

[19] 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/.

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

tend to be located to the southeast of untreated schools.

2.3 Trends in school characteristics

Because 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. One reassuring feature, shown in Appendix Figure C.1, is that, in spite of the measurable differences across schools, there is substantial geographical overlap between them.

Because we have repeated observations for each school over time, we will employ a panel fixed effects approach, meaning that level differences alone do not constitute threats to identification. For our results to be biased, there must be *time-varying* differences between treated and untreated schools which correlate with the timing of energy efficiency upgrades. In order to examine the extent to which this is occurring, we examine patterns in five key school characteristics across treated and untreated schools over time using an event study specification. In particular, we examine the number of enrolled students, number of staff members, and the percentage of students performing "proficient" or better – the state standard – on California's Standardized Testing and Reporting (STAR) math and English/language arts exams, and energy consumption. Our estimating equation is:

$$Y_{it} = \sum_{y=-2}^{4} \beta^y \mathbf{1}[\text{Year to upgrade } = y]_{it} + \alpha_i + \gamma_t + \varepsilon_{it} \tag{2.1}$$

where Y_{it} is our outcome of interest for school i in year t, $\mathbf{1}[\text{Year to upgrade } = y]_{it}$ is an indicator defining "event time," such that $y = 0$ is the year of the energy efficiency upgrade, $y-2$ is 2 years prior to the upgrade, and $y+4$ is 4 years after the upgrade, etc. α_i are school fixed effects, γ_t are year fixed effects, and ε_{it} is an error term, which we cluster at the school level.[21] Figure 2 displays the results of this exercise.

Across the four demographic variables, we see that treated and untreated schools are behaving similarly before and after energy efficiency upgrades. The relatively flat pre- and

[21] Because we have richer data on electricity consumption, we include a school-by-hour-of-day fixed effect rather than a school fixed effect in this final regression.

post-treatment trends are evidence in favor of our identifying assumption that treated and untreated schools were and would have remained on parallel trends in the absence of energy efficiency upgrades. In particular, the results on the number of students and number of staff suggest that treated schools did not grow or shrink substantially at the same time as they installed energy efficiency upgrades, and the test score results provide evidence that schools' instructional quality did not change dramatically around energy efficiency upgrades. We can rule out even small changes in all four variables; we find precisely-estimated null results.

The final panel of Figure 2 provides suggestive evidence that treated and untreated schools had similar trends in energy consumption prior to energy efficiency upgrades. Furthermore, we find that these upgrades are associated with a marked decline in energy consumption at treated schools. This panel lends further support to our assumption that treated and untreated schools would have remained on similar trajectories in the absence of energy efficiency upgrades, and suggests that energy efficiency upgrades caused a substantial reduction in energy use.

2.4 Potential for unobserved energy efficiency measures

In our data, we can only measure energy efficiency upgrades that happened through a subsidized program. If control schools are also implementing energy efficiency measures, the treatment effect might be attenuated. Contrarily, if treated schools are complementing the subsidized upgrades with unsubsidized purchases, our treatment effects will be overestimated.

Whereas this can be a source of bias, we have performed background research into the institutional details of energy efficiency upgrades in California schools to assess the relevance of this threat to identification, including speaking to officials involved in California's energy efficiency sector. We find evidence that energy efficiency upgrades are often prohibitively costly for schools without outside financial assistance (Borgeson and Zimring (2013); Gordon and Barba (2012)). Indeed, one rationale for passing Proposition 39 in the first place was that schools were reporting being unable to invest in energy efficiency upgrades for lack of funds.

During our sample period, prior to Proposition 39, California schools had two main ways of financing energy efficiency upgrades: energy efficiency subsidies/utility incentives

and bonds (Borgeson and Zimring (2013)), which have historically been concentrated in wealthy school districts. The primary sources of energy efficiency subsidies were the utility incentives that we study in this paper (North Carolina Clean Energy Technology Center (2020); California Energy Commission (2014)). Our empirical exercise is on these subsidies, and captures the universe of subsidized energy efficiency upgrades undertaken by schools served by PG&E during our sample.

The other potential method used to raise money for energy efficiency upgrades is passing local bond measures. We use data on the universe of local bond measures passed in California during our sample period to understand whether bond-funded but unobservable energy efficiency upgrades are biasing our results. We estimate the effect of bond passage on energy consumption in schools that did not undergo subsidized energy efficiency upgrades, to test whether these bonds indeed led to energy efficiency upgrades which we do not observe in Appendix Table C.1. We find that schools that passed bond measures appear to be consuming slightly more energy, with the point estimate being relatively small and insignificant.

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 both specification and outliers. 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 and sample restrictions than our regression analysis, and enables us to select among the millions of possible covariates in a systematic way.

3.1 Panel fixed effects approach

3.1.1 Methodology

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:

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

where Y_{ith} is energy consumption in kWh at school i on date t during hour-of-day h. Our treatment indicator, D_{it}, is a dummy indicating that school i has undertaken at least one energy efficiency upgrade by date t.[22] The coefficient of interest, β, can be interpreted as the average savings in kWh/hour at a treated school. α_{ith} 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_{ith} is an error term, which we cluster at the school level to account for arbitrary within-school correlations.[23]

We present results from several specifications with increasingly stringent controls. In our most parsimonious specification, we control for school-by-hour-of-day fixed effects, accounting for hour-specific time-invariant characteristics at each school. Our preferred specification includes school-by-hour-by-month-of-year 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

[22] Though schools can and do undertake multiple upgrades, we use a binary treatment indicator here due to concern about mismeasurement of treatment dates. When we instead define D_{it} as the cumulative number of upgrades undertaken by school i by time t, we find smaller realization rates, further supporting our conclusion that energy efficiency upgrades deliver less than the expected savings. Discussions with the utility confirmed that there is substantial heterogeneity on how accurately the dates are recorded.

[23] Both here and in the machine learning approach described below, to speed computation time, we collapse the data to the school-by-month-of-sample-by-hour-of-day level to run the regressions. This collapse enables us to more easily include month-of-sample and school-hour-specific fixed effects while being able to consider a wide range of robustness checks. After collapsing the data, we weight our regressions such that we recover results that are equivalent to our estimates on the disaggregated data. The main results using the uncollapsed data are virtually the same and available in Table C.2 in the Appendix.

comes from within-school-by-hour-month-of-year and within-month-of-sample differences between treated and untreated schools.

Realization rates To assess the performance of these energy efficiency measures, we compare our estimates to average *ex ante* estimates of expected savings. We follow the existing energy efficiency literature in calculating realization rates.[24] Specifically, we calculate the realization rate as $\hat{\beta}$ divided by the average expected savings for upgrades in our sample. To ensure that the average savings are properly weighted to match the relevant regression sample, we compute these average savings by regressing expected savings for each school at a given time t (equal to average expected savings for treated schools during the post period, and zero otherwise) on the treatment time variable and the same set of controls and fixed effects as its corresponding regression specification. 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.

Additionally, we also consider the implied realization rates that result from a modified model:

$$Y_{ith} = -\beta \, \overline{S}_i \times D_{it} + \alpha_{ith} + \varepsilon_{ith} \tag{3.2}$$

where $\overline{S}_i$ are the expected average savings at school i that undergoes an energy savings intervention (and zero otherwise). A coefficient of one implies that expected savings explain the post-treatment gap between schools with energy efficiency upgrades and without them. Note that the realization rate estimated with this method has a different interpretation. Whereas our first approach (the method used throughout the literature) yields the average realization rate across the program, our second approach estimates the average school-specific realization rate. Because the second approach leverages school-specific savings estimates, it is more demanding. For example, to the extent that expected savings contain measurement error in a classical sense, which would be very natural as they are a prediction of future savings, one would expect estimates from the second approach to be attenuated. Because these ap-

[24] Davis et al. (2014), Fowlie et al. (2018), Levinson (2016b), Kotchen (2017), Novan and Smith (2018), and Allcott and Greenstone (2017) all use this method.

proaches estimate different objects, we do not necessarily expect the resulting realization rate estimates to be the same.[25]

3.1.2 Results

Panel A in 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 3.5 kWh/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-by-month-of-year and month-of-sample fixed effects, we find that energy efficiency upgrades caused a 1.81 kWh/hour reduction in energy consumption at treated schools. In column (6), we also control for temperature, and find a 1.60 kWh/hour reduction in energy consumption, and a realization rate of 0.39. These results are all precisely estimated; all energy savings estimates are statistically significant at the 1 percent level.[26]

Panel B in Table 2 reports results from estimating Equation (3.2) using the same sets of fixed effects.The implied realization rate is similar using this approach, with estimates between 0.41 and 0.59.

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: our estimated realization rates range from 0.31 to 0.81. This suggests that energy savings in schools are not as large as expected. In our most comprehensive specification, which includes a temperature control, the realization rate is 0.39, implying that only 39 percent of the expected savings are realized.

[25]Heterogeneity in treatment effects across schools is another reason that will cause these estimates to differ, eg Houde and Myers (2019).

[26]In Appendix Table C.3, we present standard errors 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.

3.1.3 Panel fixed effects robustness

Trimming We subject our panel fixed effects approach to a number of standard robustness checks. We begin by examining the sensitivity of our estimates to outliers. This is particularly important in our context, because we run our main specifications in levels to facilitate the computation of realization rates. Table 3 repeats the estimates from Table 2 with three different approaches to removing outliers. In Panel A, we trim observations below the 1st or above the 99th percentile of energy consumption. Doing so reduces the point estimates dramatically. We now estimate savings between 0.47 kWh/hour and 2.47 kWh/hour. This trimming also has substantial impacts on our realization rate estimates, which now range from 0.12 to 0.59.

In Panel B, we instead trim schools below the 1st and above the 99th percentile in terms of expected savings. We implement this trim because expected savings has an extremely skewed distribution in our sample.[27] We find that the results are less sensitive to this trim than the trim in Panel A; we now estimate point estimates between 1.02 kWh/hour and 3.25 kWh/hour, and realization rates between 0.29 and 0.86.

In Panel C, we implement both trims together, and the results are similar to those in Panel A. We again find much lower point estimates (ranging from 0.49 kWh/hour to 2.43 kWh/hour) and realization rates (ranging from 0.14 to 0.65) than in the full sample.[28]

Overall, the panel fixed effects estimates are sensitive to both specification and to outliers in the sample. This is concerning from a policy perspective; realization rates between 0.31 and 0.81 have substantially different implications than rates between 0.14 and 0.65, and is also cause for concern about the performance of the panel fixed effects estimator in this context. Controlling for temperature in specification (6) helps mitigate the effects of trimming somewhat, but the results remain sensitive to outliers, with estimated realization rates moving between 0.39 with no trimming to 0.23 when trimming outlier observations and upgrades.

[27] The median project was expected to save 16,663 kWh, while the average project was expected to save 46,050 kWh. We believe some of this to be measurement error; five percent of schools in the sample which are expected to reduce their energy consumption by 50 percent through energy efficiency upgrades, which seems unrealistic.

[28] Appendix Table C.4 presents the analogous results for the school-specific realization rate calculations.

Measurement error A concern in this setting is mis-measured treatment dates. We use two approaches to address this. First, we run "donut" specifications where we drop months immediately before and after treatment to account for possible mis-measurement of treatment dates. We present these results in Appendix Table C.5 and Table C.6. Our estimated average program realization rates rise somewhat, from 0.39 with the full sample to 0.46 dropping 3 months before and after, using our preferred specification; the school-specific realization rates rise from 0.45 to 0.50, suggesting that there may be some mis-measurement in treatment dates.[29] Second, our treatment variable is a binary indicator equal to 1 after a school undergoes its first energy efficiency upgrade, rather than a continuous timing measure, since the time of implementation is measured with substantial error. Appendix Table C.7 presents the results using a continuous timing variable. As expected, these effects are quite attenuated: in our main specification, we estimate a realization rate of 0.16.

The results presented above come from a fairly standard parsimonious specification. It bears pointing out that there are many possible variants on the panel fixed effects design (see for example the matching approach from Ferraro and Miranda (2017) and Cicala (2015) or the Abadie et al. (2010) synthetic control method).[30] Given the richness of our data, we also have a great deal of flexibility in our choice of control variables, fixed effects, and functional form. In order to add additional controls in an algorithmic fashion, we now turn to a machine learning approach.

3.2 Machine learning approach

Even with a large set of high-dimensional fixed effects, the standard panel approach performs poorly on basic robustness tests, and is extremely sensitive to specification. A natural next step would be to add additional controls. However, given the size of the dataset, a

[29] One important caveat to these donut results is that we have a relatively short panel for many schools, so we are losing some schools and inducing selection as we drop data, as shown by the decreasing number of observations in the donut tables, e.g., comparing Panel A (one month donut) to Panel C (three months donut).

[30] As one variant on our main approach, we conduct a limited nearest neighbor matching exercise, in which we use observable characteristics of treated schools to find similar untreated schools. Appendix Table C.8 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.

researcher interested in capturing heterogeneity could interact several variables with school and hour-of-day, generating millions of candidate interactions. This makes the process of model selection computationally expensive and ad hoc. In order to address some of these issues more systematically, we use a machine learning approach that leverages the richness of the data.[31]

3.2.1 Methodology overview

Our machine learning estimator proceeds in two steps. In a **first step**, we use machine learning tools to create unit-specific models of an outcome of interest. Our approach builds on a standard regression model, of the form:

$$Y_{it} = \beta D_{it} + \gamma_i X_{it} + \alpha_{ith} + \epsilon_{it}$$

Our estimation differs from this traditional approach in two ways. First, we use machine learning (rather than researcher choice) to choose the set of X_{it} variables.[32] Second, we are informing γ_i using pre-treatment observations only.[33] This allows us to separate the choice of school specific coefficients, γ_i, from the estimation of the treatment effect, β, which in our context is computationally appealing.[34]

[31] Machine learning is 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 out-of-sample prediction (Abadie and Kasy (2017)). These methods also enable researchers to allow for a substantially wider covariate space than would be feasible with trial-and-error.

[32] Machine learning methods have become increasingly popular in economics. Athey (2017) and Mullainathan and Spiess (2017) provide useful overviews. Other papers in this literature include McCaffrey et al. (2004), who propose a machine learning based propensity score matching method; Wyss et al. (2014), who force covariate "balance" by directly including balancing constraints in the machine learning algorithm used to predict selection into treatment; and Belloni et al. (2014) and Chernozhukov et al. (2018) propose a "double machine learning" 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.

[33] We take a similar approach to that proposed Varian (2016) to use pre-treatment data and machine learning to forecast a post-treatment counterfactual. Our approach is also similar in spirit to Athey et al. (2017), in which the authors propose a matrix completion method for estimating counterfactuals in panel data. Souza (2019) also uses a similar approach to our first step, in the context of the Weatherization Assistance Program, and highlights some of the appealing features of the approach to examine heterogeneity and treatment effects, as well as robustness to concerns about two-way fixed effects estimation raised in Borusyak and Jaravel (2018) and Goodman-Bacon (2019).

[34] By using only pre-treatment data, our estimator is consistent but not particularly efficient. In Section 3.2.4, we consider an alternative double machine learning procedure (Chernozhukov et al. (2018)) and find

We use these models to create (fully out-of-sample) predictions of our outcome of interest in the post-treatment period. The difference between the actual outcome and the prediction (i.e., the prediction error) already give us a rough idea of a treatment effect.[35] However, it does not properly control for time trends and other confounding factors that would be accounted for in a differences-in-differences setting.[36]

To address this concern, our regression specification in the **second step** is analogous to our panel fixed effects model, described in Equation (3.1), but we now use the prediction error from the first step as the dependent variable:

$$Y_{ith} - \hat{Y}_{ith} = \beta D_{it} + \alpha_{ith} + \gamma \text{posttrain}_{ith} + \varepsilon_{ith}, \tag{3.3}$$

where α_{ith} and ε_{ith} are defined as in Equation (3.1), $\hat{Y}_{ith}$ is the prediction in kWh from step one and posttrain_{ith} is a dummy, equal to one during the out-of-sample prediction period.[37] We include this dummy to account for possible bias in the out-of-sample predictions, by re-centering prediction errors in the untreated schools around zero.[38] We cluster our standard errors at the school level.[39] This combination of machine learning methods with panel fixed effects approaches enables us to control for confounding trends.

very similar results, both in point estimates and standard errors.

[35] Figure A.1 in Appendix A provides a graphical intuition of this raw comparison.

[36] Appendix A uses the machine learning predictions in the first step to construct treatment effects in a variety of ways, and shows how our results vary depending on what control group and time periods we include in the analysis. In contrast with step 2, these estimators do not flexibly control for month of sample or trends.

[37] As in the panel approach, we run these regressions using month-hour collapsed data, given that there is limited value in keeping the disaggregated hourly data. Table C.10 in the Appendix shows that the results are essentially the same when using the uncollapsed hourly data.

[38] As shown in Panel D of Figure 3 below, these prediction errors are centered around zero in our application, so in practice this has a minimal impact on the results. However, this correction could be important in other settings.

[39] These standard errors do not account for prediction errors from the first step. To our knowledge, there is no guidance from the econometrics literature on doing proper inference in this panel machine learning estimator, so we present an alternative bootstrap approach In Appendix Table C.12. We begin by sampling weeks of sample with replacement for each school independently. We then feed these bootstrapped data into the school-specific LASSO and compute 20 alternative prediction models per school, depending on the bootstrap sample. Finally, we sample these bootstrapped school predictions with replacement before running our ultimate regressions (to produce the block-bootstrap analogue of clustering by school). We use the standard deviation of the bootstrapped treatment effects as our bootstrapped standard errors. These standard errors are quantitatively similar to our conventional clustered standard errors. Because the bootstrap standard errors are very similar to the clustered ones, and because the bootstrap procedure is significantly more computationally intensive than clustering, we use the clustered standard errors throughout the remainder of the text.

As in the regression approach, we also estimate the complementary regression

$$Y_{ith} - \hat{Y}_{ith} = -\beta \ \overline{S}_i D_{it} + \alpha_{ith} + \gamma \text{posttrain}_{ith} + \varepsilon_{ith}, \tag{3.4}$$

in which we estimate the average school-specific realization rate.

Identification As with the standard panel fixed effects approach, we assume that, conditional on control variables, energy consumption at treated and untreated schools would have been trending similarly in the absence of treatment. In this specification, we require treated and untreated schools to be trending similarly in *prediction errors*, rather than in 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, computationally tractable, and systematic way.

It is important to note that our machine learning approach —just like the panel fixed effects approach— is not immune from bias stemming from energy consumption changes that coincide directly with the subsidized energy efficiency upgrades. If a school undertakes additional energy-saving behaviors or unsubsidized upgrades at the same time as an energy efficiency upgrade in our sample, we will overestimate energy savings and the resulting realization rates will be over-estimated.[40] Any remaining positive selection into treatment, for instance, based on the expected size of the treatment effect, will bias our estimates away from zero, leading us to estimate energy efficiency savings and realization rates that are more favorable. For a confounder to bias our results towards zero, a school would have to increase energy use at the same time as our upgrades.[41]

We continue by providing a more thorough discussion of our machine learning methodology and describing the results.

[40] As discussed in Section 2.4, unsubsidized upgrades are not likely in our context.

[41] As discussed in Section 2.3, we show in Figure 2 that school size, number of staff, and test scores do not change dramatically around the time of upgrade. This does not rule out the possibility of dramatic changes in energy usage that were coincident with energy efficiency upgrades, but it does appear unlikely that major schooling changes are driving our results.

3.2.2 Step 1: Predicting counterfactuals

In the first step, we use machine learning to construct school-by-hour-of-day specific prediction models. For treated schools, we define the pre-treatment period as the period before any intervention occurs. For untreated schools, we randomly assign a "treatment date," which we use to define the "pre-treatment" period.[42] We train these models using pre-treatment data only, as described above.[43]

There are many possible supervised machine learning methods that researchers could use in this step. 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.[44] We allow the LASSO to search over a large set of potential covariates, including the day of the week, a holiday dummy, a month dummy, a temperature spline, the maximum and minimum temperature for the day, and interactions between these variables. Because we are estimating school-hour-specific models, each covariate is also essentially interacted with a school fixed effect and an hour fixed effect—meaning that the full covariate space includes over 12 million candidate variables.[45,46] Having hourly data for each school enables us to build a rich model to effectively forecast electricity usage out of sample.[47] In

[42] We randomly assign this date between the 20th and 80th percentile of in-sample calendar dates in order to have a more balanced number of observations in the pre- and post-sample, similar to that in the treated schools.

[43] 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. For a treated school with a treatment date of July 16, 2012, we use only data prior to this date while to generate the prediction models.

[44] We also consider variants on the LASSO and two random forest approaches, as well as alternative tuning parameters. We use the correlation between the predicted and actual energy consumption for untreated schools in the post-training period as an out-of-sample check on the performance of these different models. Table C.9 displays the results of this exercise, showing the distribution of correlations between data and predictions across these six methods. Our chosen method, including basic variables and untreated schools, and using `glmnet's` default tuning parameter, performs slightly better than the other options. We also explore results using these different models in Appendix Figure C.2, which shows that hour-specific treatment effects are robust to the choice of method.

[45] To make the approach computationally tractable, we estimate a LASSO model one school-hour at a time. Therefore, each school-hour model has a few thousand variables at a time.

[46] Note that we do not include time trends in the prediction model, because we are generating predictions substantially out of sample and these trends could dramatically drive predictions. 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.

[47] In step 2, we aggregate the data to the school-by-month-by-hour level to speed regression computation time. However, we generate the underlying predictions using the highest-frequency data available, to enable the model to flexibly capture features which matter for energy use.

addition to these unit-specific variables, we also include consumption at untreated schools as a potential predictor, in the spirit of the synthetic control literature (Abadie et al. (2010)). The LASSO algorithm uses then cross-validation to parameterize the degree of saturation of the model and pick the variables that are included.[48]

Validity checks We perform several diagnostic tests to assess the performance of our predictions. Figure 3 presents four such checks. First, Panel A plots the number of selected covariates for each model against the size of the pre-treatment sample. LASSO penalizes extraneous variables, meaning that the optimal model for any given school will not include all of the candidate regressors.[49] Though the LASSO typically selects fewer than 100 variables, the joint set of variables selected across all schools and hours covers the majority of the candidate space (a total of 1,149 variables are selected), highlighting the importance of between-school heterogeneity.

We can also inspect the selected covariates individually. As an illustration, Panel B of Figure 3 shows the coefficient on the holiday dummy (and its interactions) in each school-hour-specific prediction model.[50] We find that, across models, holidays are negatively associated with energy consumption. This suggests that the LASSO-selected models reflect real-world electricity use. We also find substantial heterogeneity across schools: each of the candidate holiday variables is selected at least once, but the median school has no holiday variable, highlighting the importance of data-driven model selection.

Panel C of Figure 3 shows the variables selected by each of the school-hour models for treated and untreated schools separately. Nearly all of the models include an intercept, and around 70 percent of the models include consumption from at least one untreated school; the

[48] We use the package `glmnet` in R to implement the estimation of each model. To cross-validate the model, the 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 testing set. We tune the `glmnet` method to perform 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.

[49] The LASSO performs best when the underlying DGP is sparse (Abadie and Kasy (2017)). We find evidence in favor of this in our empirical context, as the number of chosen regressors does not scale linearly with the size of the training set.

[50] 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.

median school-hour model includes ten such covariates. Month and temperature variables are each included in nearly half of the models. Several 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, wherein we would estimate the same covariates for each unit.

Finally, we can perform a fully out-of-sample test of our approach by inspecting prediction errors at untreated schools in the post-treatment period. Because these schools do not experience energy efficiency upgrades, these prediction errors should be close to zero. Panel D of Figure 3 plots the distribution of average out-of-sample prediction error for each school-hour, trimming the top and bottom 1 percent. As expected, this distribution is centered around zero.[51]

Taken together, these four checks provide evidence that the machine learning approach is performing well in predicting schools' electricity consumption, even out-of-sample.

3.2.3 Step 2: Panel regressions with prediction errors

We now regress the prediction errors from the machine learning model on a treatment indicator and the rich set of fixed effects we use in the earlier panel fixed effects approach. Panel A in Table 4 reports results from estimating Equation (3.3) for five different fixed effects specifications. We find that energy efficiency upgrades resulted in energy consumption reductions of between 2.1 and 3.9 kWh/hour. In our preferred specification (Column (5)), which includes school-by-hour-by-month and month-of-sample fixed effects, we find that energy efficiency upgrades reduced electricity use by 2.4 kWh/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, and are highly statistically significant.[52]

We again compare these results to the *ex ante* engineering estimates to form realization rates. Our estimated realization rates range from 0.53 to 0.92. These realization rates are

[51] Because we see no trends in observable outcomes in Figure 2, the fact that this distribution is centered around zero provides further evidence that untreated schools are not undertaking unsubsidized energy efficiency measures that are not observed in our data.

[52] In Appendix Table C.11, we present results two-way clustering on school and month of sample. The results remain highly statistically significant using these alternative approaches.

statistically different than zero and larger than the estimates from our panel fixed effects approach. Some of the specifications imply that realized savings were close to expected savings, with our preferred specification implying a realization rate of 60 percent.

Panel B in Table 4 presents the alternative realization rates from equation (3.4). The estimated rates are stable across specifications ranging 0.50 to 0.58. As in regression case, the estimated rates using this alternative approach are smaller, consistent with measurement error in expected savings.

3.2.4 Machine learning robustness

Trimming As with the panel fixed effects approach, we test the extent to which our machine learning results vary as we exclude outlying observations in Table 5. In Panel A, we drop observations that are below the 1st or above the 99th percentile of the dependent variable – now defined as prediction errors in energy consumption. Unlike in the panel fixed effects approach, we find that this trimming has very limited impacts on the results. We now find point estimates ranging from -3.52 kWh/hour to -2.12 kWh/hour, and accompanying realization rates ranging from 0.55 to 0.86. These are similar to our estimates in Table 4. In Panel B, we again trim schools with expected savings below the 1st or above the 99th percentile. We find that this, too, neither meaningfully alters our point estimates nor our realization rates, which now range from -3.69 kWh/hour to -1.87 kWh/hour and 0.52 to 0.98, respectively. Finally, in Panel C, we trim on both dimensions, and again find remarkably stable point estimates and realization rates, ranging from -3.41 to -2.05 kWh/hour and 0.58 to 0.92. While the panel fixed effects results displayed in Table 3 were highly sensitive to these trimming approaches, the machine learning results are quite stable.[53]

Measurement error As with the panel approach, a concern in this setting is mis-measured treatment dates. We also run "donut" specifications where we drop months immediately before and after treatment. We present these results in Appendix Table C.14 and Table C.15. Our estimated average program realization rates rise somewhat, from 0.60 with the full sample to 0.68 dropping 3 months before and after, using our preferred specification; the

[53] Appendix Table C.13 presents the analogous results for the school-specific realization rate calculations.

school-specific realization rates rise from 0.50 to 0.56. Our main specification also uses a binary treatment variable, rather than a continuous measure. In Appendix Table C.16, we present results with a continuous timing variable. As with the panel approach, these are quite attenuated (our preferred realization rate estimate falls to 0.24), hence our preference for the binary timing of treatment variable.

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. 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. We also explore an alternative approach using double machine learning (Chernozhukov et al. (2018)), which has as key difference that all the data are used for the prediction, not just the pre-period.[54]

Appendix Table C.17 shows the results where we estimate (3.1) with different prediction algorithm approaches. We find energy savings between 2.20 kWh per hour and 2.46 kWh per hour. Using our preferred LASSO approach (Column (4)), we estimate savings of 2.42 kWh per hour. These estimates translate into realization rates of 0.54, 0.61, and 0.60, respectively.[55] Given that the double machine learning approach is the most different in spirit, as it takes advantage of the whole sample to estimate the model (as opposed to just pre-treatment data), we also include in Table C.18 in the Appendix the analogues of Tables 4 and 5 for the double machine learning approach. These estimates are generally not statistically distinguishable, suggesting that the machine learning approach is not highly sensitive to our chosen prediction algorithm.

[54]Note that the regression specification in this case is different, as the double machine learning approach needs to also partial out the timing of treatment. Therefore, we regress the predictions errors of electricity consumption on the *prediction errors* of the treatment variable. See Table C.18 in the Appendix for more details.

[55]Appendix Figure C.2 shows hour-specific treatment effects for all of the machine learning methods shown here. Both the hourly patterns and the levels are very similar across methods.

3.3 Comparing approaches

In contrast with the standard panel fixed effects approach, our machine learning method delivers results that are larger and substantially less sensitive to both specification and sample selection. This highlights one advantage of using machine learning approaches in panel settings: by controlling for confounding factors using a flexible data-driven approach, this method can produce results that are more robust to remaining researcher choices.

We explore this result further in Figure 4, which shows the distribution of estimated realization rates across several specifications and samples. Notably, the policy implications from the different panel fixed effects estimates vary widely, and are centered around a 40 percent realization rate, whereas the estimates using the machine learning approach are more stable around realization rates closer to 60 percent. As shown in the figure, controlling for temperature in the panel regressions does not meaningfully impact the comparison between the two approaches.

While researchers could attempt a variety of alternative specifications in an ad-hoc way in order to reduce sensitivity to specification and sample, this approach is impractical with high-frequency datasets. Doing model selection by hand is computationally expensive and arbitrary.[56] In contrast, our machine learning approach enables researchers to perform model selection in a flexible yet systematic way, while maintaining the identifying assumptions needed for causal inference in a standard panel fixed effects approach.

4 Heterogeneity in realization rates

Our preferred estimates imply that energy efficiency upgrades in public schools only delivered 60 percent of expected savings. What other lessons can we learn from the data? Unfortunately, we cannot perform a full cost-benefit analysis, which would require accounting for the full benefits of the energy efficiency upgrades as well as reliable cost data. First, energy efficiency upgrades may be associated with welfare benefits beyond reductions in electricity

[56] Given that we have an unbalanced panel, in which some schools are observed for longer periods than others, it is also unclear that saturating the model equally across schools is necessarily the best strategy.

consumption.[57] Second, the data we obtained from PG&E do not contain comprehensive information on costs.[58]

However, our methodology allows us to further explore heterogeneity in realization rates. Beyond estimating average realization rates, understanding whether these rates vary based on observable characteristics of upgrades or treated schools may be informative for policymakers deciding which upgrades to subsidize and which schools to target.[59]

Given the richness of our electricity consumption data, we start by estimating school-specific treatment effects, as a precursor to determining what drives heterogeneity in realization rates. These estimates should not be taken as precise causal estimates of savings at any given school, but rather as an input to projecting heterogeneous estimates onto school-specific and intervention-specific covariates for descriptive purposes.[60]

To compute these school-specific estimates, we regress prediction errors in kWh on a school-specific dummy variable, equal to one during the post-treatment period (or, for untreated schools, the post-training period from the machine learning model), as well as school-by-hour-by-month fixed effects to control for seasonality. The resulting estimates represent the difference between pre- and post-treatment energy consumption at each individual school. We follow Chandra et al. (2016) and use an empirical Bayes approach to shrink the school-

[57] As we discuss in Section 2, public funding was directed towards energy efficiency upgrades in K-12 schools in California in part because of the difficulty schools face in raising funds for capital upgrades, which could increase the marginal value of the upgrades. Additionally, more efficient air conditioning units could mitigate the negative impacts of high temperatures on human capital accumulation (Graff Zivin et al. (2017); Park (2017); Garg et al. (2018)). We provide suggestive evidence that energy efficiency upgrades do not improve standardized test scores in Figure 2, though aggregate test scores remain an imperfect and noisy proxy for human capital accumulation.

[58] The only cost information in our dataset is the "incremental measure cost," a measure of the difference in the cost of a "base case" appliance replacement versus an energy efficient version. We do not, however, have data on the total cost of the appliance replacement, nor on projected energy savings from the base case counterfactual.

[59] There can also be heterogeneity in the timing of savings. As Borenstein (2002) and Boomhower and Davis (2017) point out, however, the value of energy savings varies over time. We estimate hour-specific treatment effects, presented in Appendix Figure C.2, across several machine learning methods. We find evidence that the largest reductions occur during the school day, consistent with our results picking up real, rather than spurious, energy savings. Because our focus in this paper is on realization rates, which are determined by overall savings, we do not focus on these additional estimates.

[60] Naturally, the identifying assumptions required to obtain school-specific treatment effects are much stronger than when obtaining average treatment effects, as concurrent changes in consumption at each specific school will be confounded with its own estimated treatment effect (i.e., random coincidental shocks to a given school that might not confound an average treatment effect will certainly confound the school-specific estimate of that given school).

specific estimates. We can 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 5 displays the relationship between these school-specific savings estimates and expected savings for treated schools after implementing the shrinkage procedure. We find a positive correlation between estimated savings and expected savings, although there is substantial noise in the school-specific estimates. Once we trim outliers in expected savings, we recover a slope of 0.41.[61] As we expected, this is close to the estimates from our second realization rate approach, which estimates the average school-specific realization rate.

Panel B of Figure 5 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, in line with Panel D of Figure 3. 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.[62] These results are in line with our main finding that energy efficiency projects successfully deliver significant savings, although the relationship between the savings that we can measure and the *ex ante* predicted savings is noisy.

To explore heterogeneity in realization rates, we explore the correlation between our school-specific savings estimates and school and intervention covariates (latitude, longitude, school enrollment, type of intervention, etc.).[63] Ultimately, we uncover noisy correlations between school characteristics and realization rates, suggesting that finding "low-hanging fruit" to improve the success of energy efficiency upgrades in this setting is difficult. That said, several features of our setting make recovering these types of patterns particularly challenging. Our sample of treated schools is relatively small and each of the schools is subject to its idiosyncrasies, leading to concerns about collinearity and omitted variables bias. It is possible that, with more homogeneous energy efficiency projects and a larger

[61] Before shrinking the estimates, we recover a slope of 0.44.

[62] Appendix Figure C.3 presents the results using the double machine learning procedure rather than our standard approach. We find quantitatively similar estimates (a slope of 0.40) using this alternative method.

[63] See Appendix B for a detailed presentation of the regressions and their results. Importantly, we only have one observation per treated school, for a total sample size of fewer than 1,000 units. In addition to these regressions not being causally identified, the limited sample means they should be interpreted with caution.

pool of treated units, policymakers could identify covariates that better predict realization rates. This information could be used to target the most effective interventions and improve average performance.

5 Conclusion

We leverage high-frequency data on electricity consumption and develop a machine learning method to estimate the causal effect of energy efficiency upgrades at K-12 schools in California. We use two main approaches to do this, both of which leverage cross-sectional and temporal variation to separate the causal effect of energy efficiency upgrades from other confounding factors. We begin with a panel fixed effects approach. Using this method, we estimate that energy efficiency upgrades saved 40 percent of *ex ante* estimated savings. However, these estimates are sensitive to specification and outliers, and range from 12 to 86 percent. Given the richness of our setting, there are millions of possible covariates we could include as controls. In order to parsimoniously select among these control variables, we implement a second approach, using tools from machine learning.

In our machine learning approach, we use untreated time periods in high-frequency panel data to generate school-specific predictions of energy consumption that would have occurred in the absence of treatment for both treated and untreated schools. We generate prediction errors by comparing these predictions to realized energy consumption, and estimate the causal effect of energy efficiency upgrades by estimating a panel fixed effects model using prediction errors as the outcome variable. This approach allows us to select among covariates in a parsimonious way, while still accounting for common shocks. Our approach is computationally tractable, and can be applied to a broad class of applied settings where researchers have access to relatively high-frequency panel data.

Using this method, we find that energy efficiency upgrades deliver 60 percent of *ex ante* expected savings on average. As compared to our panel fixed effects approach, we see that the machine learning approach delivers a narrower range of estimates: energy efficiency upgrades deliver between 52 and 98 percent of expected savings, depending on outliers and specification, allowing us to reject the very low realization rates suggested by some of the

panel specifications. This highlights the potential benefits of using machine learning to select among a large set of exogenous control variables. We explore heterogeneity in realization rates but we ultimately find it difficult to identify school characteristics that systematically predict higher realization rates. This suggests that without collecting additional data, improving realization rates via targeting may prove challenging.

This paper extends the energy efficiency literature to a non-residential sector. We demonstrate that energy efficiency upgrades deliver lower savings than expected *ex ante*, although in some specifications we cannot reject full realization rates, and are able to reject some of the extremely low realization rates of the prior literature. 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, we show how machine learning tools can help with specification choice, leading to results that are robust to the machine learning algorithm of choice, varying sets of fixed effects, and the treatment of outliers.

References

Abadie, A., A. Diamond, and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program.

Abadie, A. and M. Kasy (2017). The risk of machine learning. *Working paper*.

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

Allcott, H. and M. Greenstone (2017, May). Measuring the welfare effects of residential energy efficiency programs. Technical report. National Bureau of Economic Research Working Paper No. 23386.

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

Athey, S., M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi (2017). Matrix completion methods for causal panel data models. Working Paper 1710.10251, arXiv.

Barbose, G. L., C. A. Goldman, I. M. Hoffman, and M. A. Billingsley (2013, 01/2013). 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, A., V. Chernozhukov, and C. Hansen (2014). Inference on treatment effects after selection amongst high-dimensional controls. *The Review of Economic Studies 81*(2), 608–650.

Blonz, J. A. (2019). The welfare costs of misaligned incentives: Energy inefficiency and the principal-agent problem. Working paper.

Blumenstock, J., G. Cadamuro, and R. On (2015). Predicting poverty and wealth from mobile phone metadata. *Science 350*, 1073–1076.

Boomhower, J. and L. Davis (2017). Do energy efficiency investments deliver at the right time? National Bureau of Economic Research Working Paper No. 23097.

Borenstein, S. (2002). The trouble with electricity markets: Understanding california's restructuring disaster. *Journal of Economic Perspectives 16*(1), 191–211.

Borgeson, M. G. and M. Zimring (2013). Financing energy upgrades for k-12 school districts: A guide to tapping into funding for energy efficiency and renewable energy improvements. Lawrence berkeley national laboratory report 6133e.

Borusyak, K. and X. Jaravel (2018). Revisiting event study designs. Working paper.

California Energy Commission (2014). Energy-related resources for schools. Cec report no. cec-400-2014-003.

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.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review 106*(8), 247–274.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018, 01). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Cicala, S. (2015). When does regulation distort costs? lessons from fuel procurement in us electricity generation. *American Economic Review 105*(1), 411–444.

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

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

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

Engstrom, R., J. Hersh, and D. Newhouse (2016). Poverty in hd: What does high resolution satellite imagery reveal about economic welfare? *Working Paper*.

Ferraro, P. J. and J. J. Miranda (2017). Panel data designs and estimators as substitutes for randomized controlled trials in the evaluation of public programs. *Journal of the Association of Environmental and Resource Economists 4*(1), 281 – 317.

Fowlie, M., M. Greenstone, and C. Wolfram (2018). Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program. *Quarterly Journal of Economics 133*(3), 1597–1644.

Garg, T., M. Jagnani, and V. Taraz (2018). Temperature and human capital in india. Working paper, UCSD.

Gerarden, T. D., R. G. Newell, and R. N. Stavins (2017, December). Assessing the energy-efficiency gap. *Journal of Economic Literature 55*(4), 1486–1525.

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

Glaeser, E., A. Hillis, S. D. Kominers, and M. Luca (2016). Crowdsourcing city government: Using tournaments to improve inspection accuracy. *American Economic Review: Papers & Proceedings 106*(5), 114–118.

Goodman-Bacon, A. (2019). Difference-in-differences with variation in treatment timing. Working paper.

Gordon, K. and J. Barba (2012). Proposition 39: Investing in california's future. Center for the next generation white paper.

Graff Zivin, J., S. M. Hsiang, and M. Neidell (2017). Temperature and human capital in the short and long run. *Journal of the Association of Environmental and Resource Economists 5*(1), 77–105.

Granderson, J., S. Touzani, S. Fernandes, and C. Taylor (2017). Application of automated measurement and verification to utility energy efficiency program data. *Energy and Buildings 142*, 191 – 199.

Houde, S. and E. Myers (2019, April). Heterogeneous (mis-) perceptions of energy costs: Implications for measurement and policy design. Working Paper 25722, National Bureau of Economic Research.

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

Itron (2017a, May). 2015 custom impact evaluation industrial, agricultural, and large commercial: Final report. Technical report.

Itron (2017b, March). 2015 nonresidential espi deemed lighting impact evaluation: Final report. Technical report.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016). Combining satellite imagery and machine learning to predict poverty. *Science 353*, 790–794.

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

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

Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2017). Human decisions and machine predictions. *Working Paper*.

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

Kotchen, M. 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.

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

Levinson, A. (2016b, October). How much energy do building energy codes save? evidence from california houses. *American Economic Review 106*(10), 2867–94.

McCaffrey, D., G. Ridgeway, and A. 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, S. and J. Spiess (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives 31*(2), 87–106.

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

Nadel, S. and L. Ungar (2019). Halfway there: Energy efficiency can cut energy use and greenhouse gas emissions in half by 2050. Report u1907 american council for an energy-efficient economy.

Naik, N., R. Raskar, and C. 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.

North Carolina Clean Energy Technology Center (2020). Dsire california programs.

Novan, K. and A. 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.

Park, R. J. (2017). Hot temperature and high stakes cognitive assessments. Working paper, UCLA.

Prest, B. C. (2020, jan). Peaking Interest: How Awareness Drives the Effectiveness of Time-of-Use Electricity Pricing. *Journal of the Association of Environmental and Resource Economists 7*(1), 103–143.

Ryan, N. (2018). Energy productivity and energy demand: Experimental evidence from indian manufacturing plants. National Bureau of Economic Research Working Paper No. 24619.

Souza, M. (2019). Predictive counterfactuals for event studies with staggered adoption: Recovering heterogeneous effects from a residential energy efficiency program. Working paper.

Tonn, B., D. Carroll, S. Pigg, M. Blasnik, G. Dalhoff, J. Berger, E. Rose, B. Hawkins, J. Eisenberg, F. Ucar, I. Bensch, and C. Cowan (2014). Weatherization works - summary of findings from the retrospective evaluation of the u.s. department of energy's weatherization assistance program. Oak ridge national laboratory report ornl/tm-2014/338.

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

Wyss, R., A. Ellis, A. Brookhart, C. Girman, M. Funk, R. LoCasale, and T. 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

Characteristic	Untreated	Treated	T-U
Hourly energy use (kWh)	33.3	57.4	24.1
	(34.1)	(72.7)	[<0.01]
First year in sample	2012	2010	-2
	(1.7)	(1.8)	[<0.01]
Total enrollment	544	727	184
	(365)	(484)	[<0.01]
Acad. perf. index (200-1000)	789	794	5
	(99)	(89)	[0.28]
Bond passed, last 2 yrs (0/1)	0.3	0.2	-0.0
	(0.4)	(0.4)	[0.24]
Bond passed, last 5 yrs (0/1)	0.4	0.4	-0.0
	(0.5)	(0.5)	[0.69]
High school graduates (%)	23.4	23.3	-0.1
	(12.3)	(11.7)	[0.87]
College graduates (%)	20.1	20.3	0.2
	(12.3)	(12.0)	[0.76]
Single mothers (%)	20.4	19.3	-1.1
	(19.2)	(18.4)	[0.22]
African American (%)	5.8	6.1	0.4
	(9.4)	(8.0)	[0.37]
Asian (%)	9.3	11.6	2.4
	(13.4)	(16.1)	[<0.01]
Hispanic (%)	41.9	43.5	1.6
	(28.4)	(26.8)	[0.21]
White (%)	34.6	30.8	-3.8
	(26.8)	(24.5)	[<0.01]
Average temp. (° F)	60.0	60.8	0.8
	(4.1)	(3.5)	[<0.01]
Coastal (0/1)	0.3	0.2	-0.1
	(0.5)	(0.4)	[<0.01]
Latitude	37.7	37.5	-0.2
	(1.1)	(1.0)	[<0.01]
Longitude	-121.6	-121.2	0.4
	(1.0)	(1.1)	[<0.01]
Number of schools	1021	910	

Notes: This table displays average characteristics of the treated and untreated schools in our sample for the entirety of our sample period. 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. 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.

Table 2: Panel fixed effects results

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Average program estimates						
Realization rate	0.68	0.81	0.52	0.31	0.43	0.39
Point estimate	-2.88	-3.47	-2.15	-1.26	-1.74	-1.60
	(0.44)	(0.44)	(0.46)	(0.45)	(0.47)	(0.45)
Observations	57,481,920	57,480,360	57,480,360	57,481,920	57,480,360	57,480,360
Panel B: Average school-specific estimates						
Realization rate	0.53	0.59	0.51	0.41	0.46	0.45
	(0.12)	(0.13)	(0.13)	(0.12)	(0.12)	(0.13)
Observations	55,818,652	55,817,256	55,817,256	55,818,652	55,817,256	57,480,360
School-Hour FE	Yes	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes	Yes
Time trend	No	No	Yes	No	No	No
Month of Sample FE	No	No	No	Yes	Yes	Yes
Temp Ctrl	No	No	No	No	No	Yes

Notes: Panel A in this table reports results from estimating Equation (3.1), with hourly energy consumption in kWh 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 a complementary regression of *ex ante* engineering energy savings where expected (and zero otherwise) on our treatment variable, where we include the same set of controls and fixed effects. Panel B reports results from estimating Equation (3.2), in which the independent variable equals (the negative of) average expected savings for treated schools after their first upgrade, and 0 otherwise.

Table 3: Sensitivity of panel fixed effects results to outliers

	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Trim outlier observation						
Realization rate	0.47	0.59	0.31	0.12	0.23	0.21
Point estimate	-1.96	-2.47	-1.25	-0.47	-0.89	-0.82
	(0.37)	(0.37)	(0.35)	(0.35)	(0.35)	(0.33)
Observations	56,323,212	56,321,525	56,321,525	56,323,212	56,321,525	56,321,525
Panel B: Trim outlier schools						
Realization rate	0.72	0.86	0.51	0.29	0.41	0.36
Point estimate	-2.69	-3.25	-1.87	-1.02	-1.47	-1.32
	(0.41)	(0.41)	(0.40)	(0.40)	(0.41)	(0.39)
Observations	56,737,632	56,736,096	56,736,096	56,737,632	56,736,096	56,736,096
Panel C: Trim observations and schools						
Realization rate	0.51	0.65	0.34	0.14	0.25	0.23
Point estimate	-1.92	-2.43	-1.24	-0.49	-0.90	-0.85
	(0.37)	(0.37)	(0.35)	(0.35)	(0.36)	(0.33)
Observations	55,689,089	55,687,427	55,687,427	55,689,089	55,687,427	55,687,427
School-Hour FE	Yes	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes	Yes
Time trend	No	No	Yes	No	No	No
Month of Sample FE	No	No	No	Yes	Yes	Yes
Temp. Ctrl	No	No	No	No	No	Yes

Notes: This table reports results from estimating Equation (3.1), with hourly energy consumption in kWh 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 a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. In Panel A, we drop observations below the 1st or above the 99th percentile of the dependent variable: energy consumption. In Panel B, we drop schools below the 1st or above the 99th percentile of expected savings. In Panel C, we drop both.

Table 4: Machine learning results

	(1)	(2)	(3)	(4)	(5)
Panel A: Average program estimates					
Realization rate	0.86	0.92	0.75	0.53	0.60
Point estimate	-3.64	-3.92	-3.17	-2.10	-2.42
	(0.50)	(0.52)	(0.49)	(0.47)	(0.49)
Observations	57,481,920	57,480,360	57,480,360	57,481,920	57,480,360
Panel B: Average school-specific estimates					
Realization rate	0.57	0.58	0.55	0.50	0.50
	(0.13)	(0.14)	(0.14)	(0.13)	(0.13)
Observations	57,481,920	57,480,360	57,480,360	57,481,920	57,480,360
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

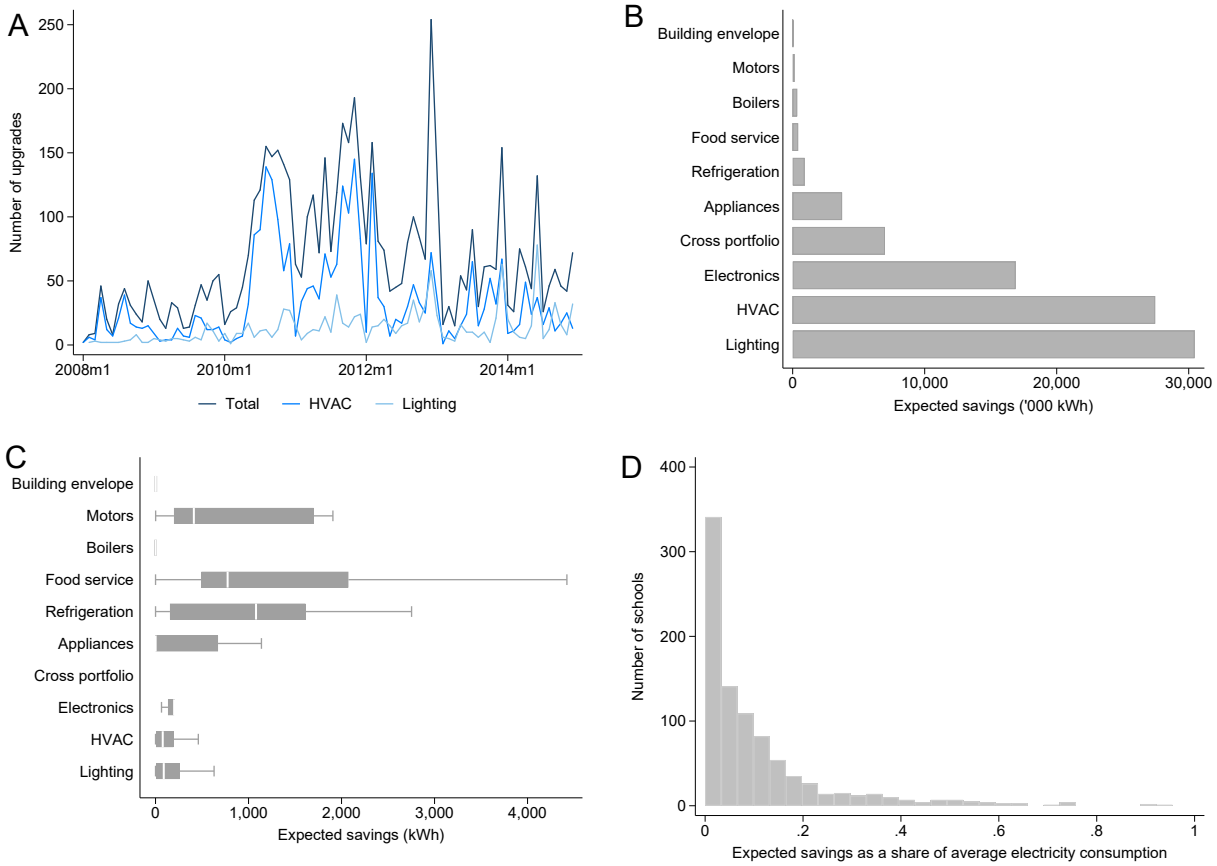
Notes: Panel A in this table reports results from estimating Equation (3.3), with prediction errors in hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. Panel B reports results from estimating Equation (3.4), in which the independent variable equals (the negative of) average expected savings 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.

Table 5: Sensitivity of machine learning results to outliers

	(1)	(2)	(3)	(4)	(5)
Panel A: Trim outlier observations					
Realization rate	0.82	0.86	0.71	0.55	0.61
Point estimate	-3.34	-3.52	-2.90	-2.12	-2.35
	(0.34)	(0.35)	(0.32)	(0.30)	(0.32)
Observations	56,332,278	56,330,677	56,330,677	56,332,278	56,330,677
Panel B: Trim outlier schools					
Realization rate	0.91	0.98	0.78	0.52	0.61
Point estimate	-3.42	-3.69	-2.92	-1.87	-2.18
	(0.47)	(0.50)	(0.47)	(0.44)	(0.47)
Observations	56,737,632	56,736,096	56,736,096	56,737,632	56,736,096
Panel C: Trim observations and schools					
Realization rate	0.87	0.92	0.76	0.58	0.64
Point estimate	-3.22	-3.41	-2.80	-2.05	-2.27
	(0.33)	(0.35)	(0.32)	(0.30)	(0.32)
Observations	55,673,654	55,672,077	55,672,077	55,673,654	55,672,077
School-Hour FE	Yes	Yes	Yes	Yes	Yes
School-Hour-Month FE	No	Yes	Yes	No	Yes
Time trend	No	No	Yes	No	No
Month of Sample FE	No	No	No	Yes	Yes

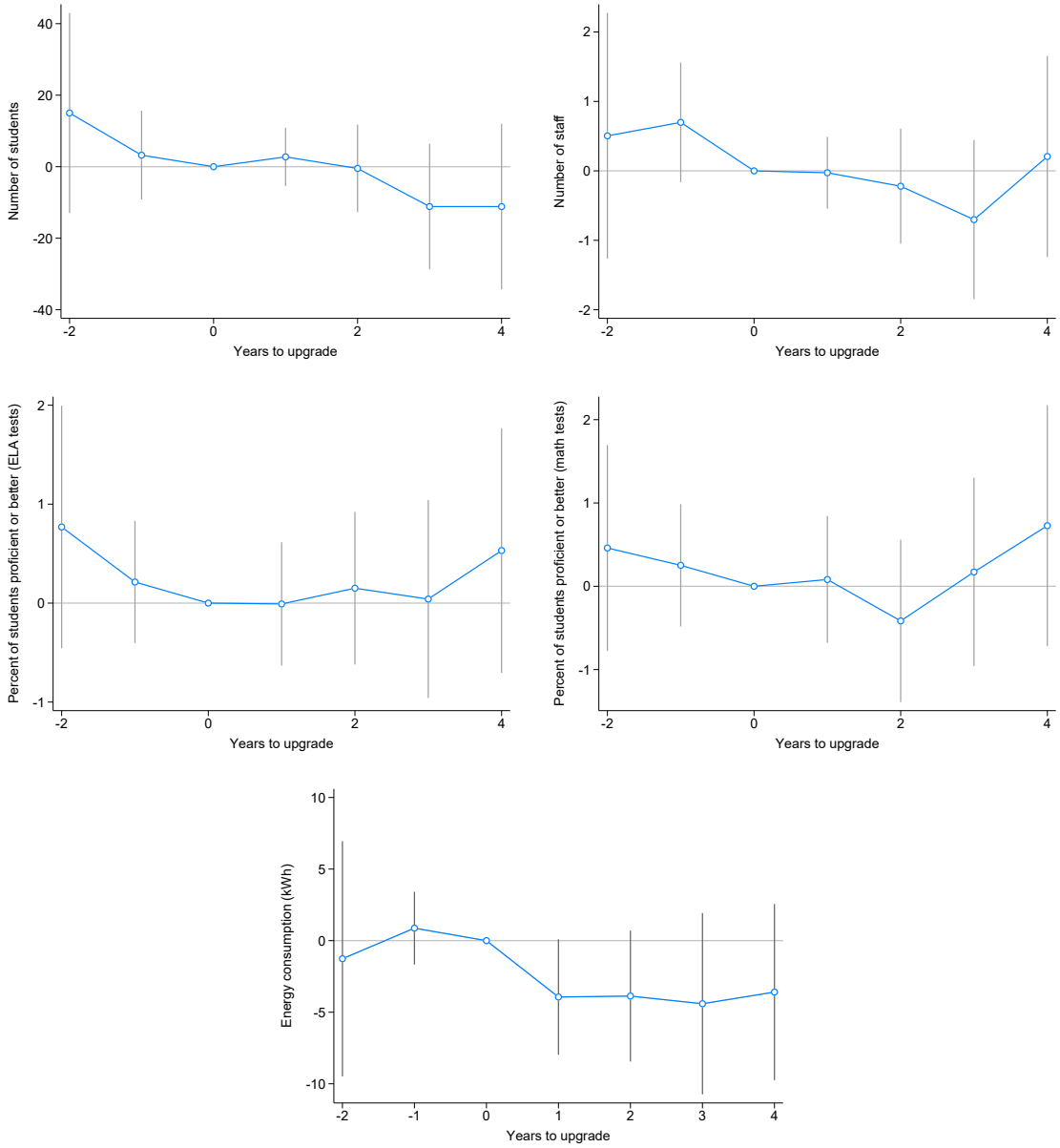
Notes: This table reports results from estimating Equation (3.3), with prediction errors in hourly energy consumption in kWh as the dependent variable. The independent variable is a treatment indicator, set equal to 1 for treated schools after their first upgrade, and 0 otherwise. Standard errors, clustered at the school level, are in parentheses. Realization rates are calculated by dividing the regression results on a complementary regression of ex-ante engineering energy savings where expected (and zero otherwise) on our treatment variable, also including the same set of controls. All regressions include a control for being in the post-training period for the machine learning. In Panel A, we drop observations below the 1st or above the 99th percentile of the dependent variable: energy consumption. In Panel B, we drop schools below the 1st or above the 99th percentile of expected savings. In Panel C, we drop both.

Figure 1: Energy efficiency upgrades



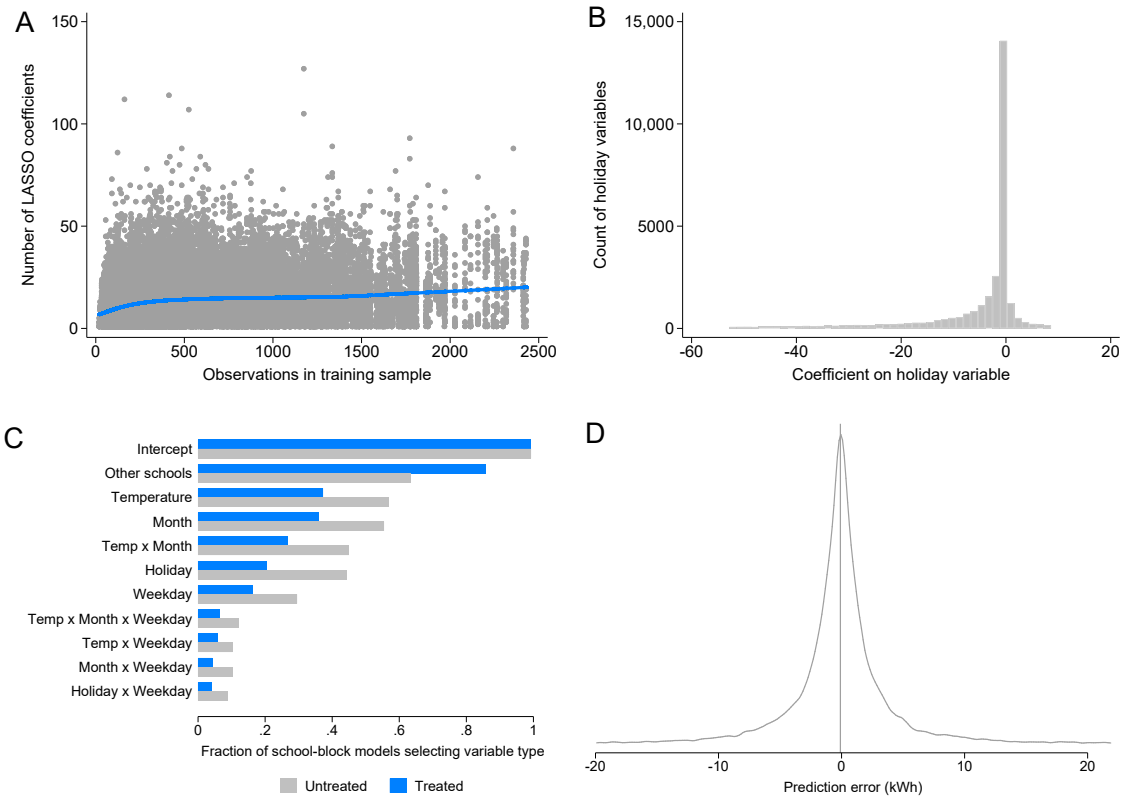
Notes: This figure shows the timing of upgrades in our sample (Panel A), the total expected savings by category of upgrade (Panel B), a box plot of savings by category (Panel C), and expected savings as a share of annual consumption (Panel D).

Figure 2: School characteristics before and after treatment



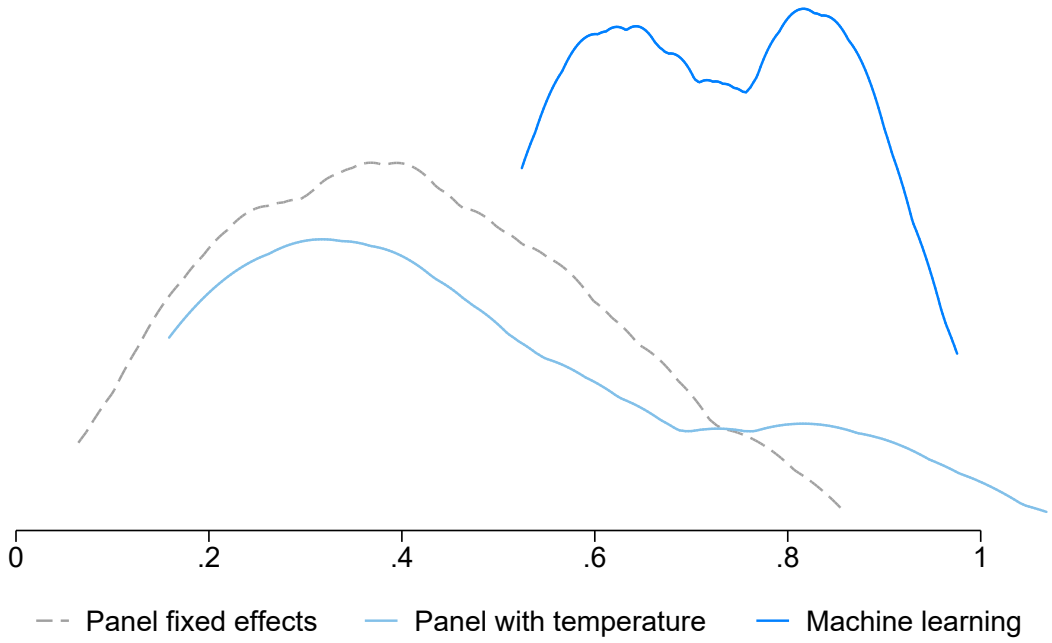
Notes: This figure shows point estimates and 95 percent confidence intervals from event study regressions of school demographics and test scores before and after an energy efficiency upgrade using a balanced panel of schools to facilitate interpretation. We normalize time relative to the year each school undertook its first upgrade. Standard errors are clustered by school. The top left panel displays results for number of students enrolled in school; the top right panel shows results for number of staff members; the middle left panel shows results for the percent of students scoring proficient (the state standard) or better on California's Standardized Testing And Reporting (STAR) math tests; and the middle right panel shows results for the percent of students scoring proficient or better on California's STAR English and Language Arts (ELA) tests.

Figure 3: Machine learning diagnostics



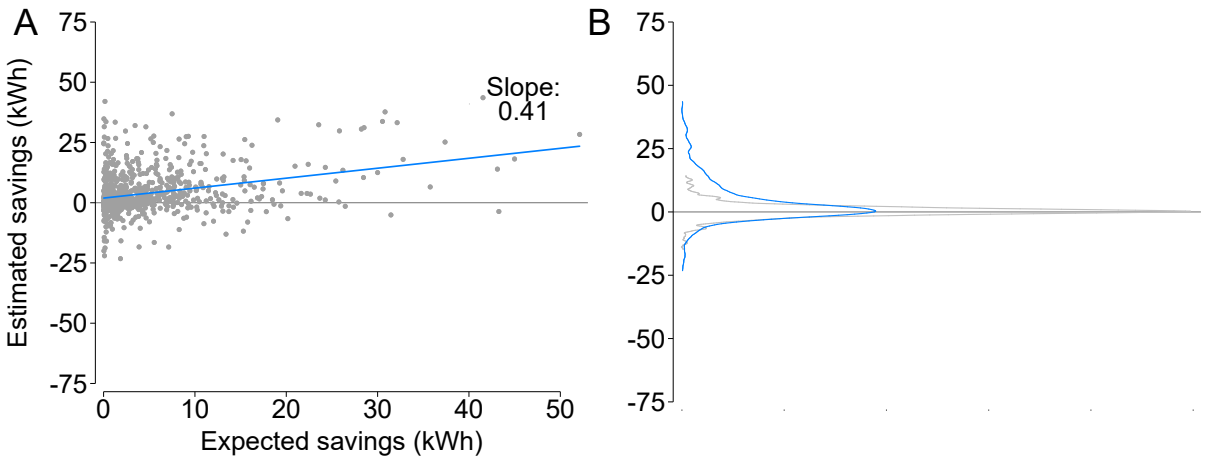
Notes: This figure presents three checks of our machine learning methodology. Panel A displays the relationship between the number of observations in the pre-treatment ("training") dataset and the number of variables LASSO selects to include in the prediction model for each school in the sample. Panel B displays the marginal effect of holiday indicators in each school-specific prediction model. Panel C displays the categories of variables selected by our preferred LASSO method for untreated and treated schools. Finally, Panel D shows the distribution of average prediction errors out-of-sample for untreated schools (trimming the top and bottom 1 percent).

Figure 4: Comparison of methods across specifications and samples



Notes: This figure shows the distribution of implied realization rates using three alternative approaches: a panel fixed effects regression, a panel fixed effects regressions with school-specific temperature controls, and a machine learning approach. The results include five specifications per method (with the fixed effects as described in Columns (1) - (5) of the main Tables 2 and 4). For the panel with temperature, we include the same fixed effects, and in all specifications, also include school-specific temperature controls. For all three curves, each of the five specifications is estimated on five different samples: no trimming; trimming observations below the 1st (2nd) and above the 99th (98th) percentile of the dependent variable; trimming the schools with smallest and largest 1 percent of interventions; and a combination of the latter two 1 percent trims. Each kernel density is computed from a total of 25 estimates.

Figure 5: 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. Panel B displays kernel densities of estimated savings in the untreated group (gray line) and estimated savings in the treated group (blue line).