Decomposing the Wedge Between Projected and Realized Returns in Energy Efficiency Programs

Peter Christensen, Paul Francisco, Erica Myers, and Mateus Souza

University of Illinois at Urbana-Champaign

February 15, 2021

Abstract

Evaluations of energy efficiency programs reveal that realized savings consistently fall short of projections. We decompose this 'performance wedge' using data from the Illinois Home Weatherization Assistance Program (IHWAP) and a machine learning-based event study research design. We find that bias in engineering models can account for up to 41% of the wedge, primarily from overestimated savings in wall insulation. Heterogeneity in workmanship can also account for a large fraction (43%) of the wedge, while the rebound effect can explain only 6%. We find substantial heterogeneity in energy-related benefits from IHWAP projects, suggesting opportunities for better targeting of investments.

Key words: Energy Efficiency, Weatherization Assistance Program
JEL Classification: Q51, Q53, R310

Christensen (pchrist@illinois.edu), Francisco (pwf@illinois.edu), Myers (ecmyers@illinois.edu), and Souza (nogueir2@illinois.edu). We are grateful for the generous support from the Alfred P. Sloan Foundation. Thanks to Mick Prince and Chad Wolfe from the Illinois' Department of Commerce & Economic Opportunity, without whom this research would not be possible. We also thank Bryan Parthum and others from the University of Illinois' Big Data and Environmental Economics and Policy (BDEEP) group for outstanding research assistance. We are grateful to Stacy Gloss and the Indoor Climate Research & Training center for liaison with state officials and utility providers. Souza is also thankful for the financial support from CAPES (Coordination for the Improvement of Higher Education Personnel - Brazil). Finally, we acknowledge the excellent feedback and comments from Josh Blonz, Meredith Fowlie, Dave Rapson, Sam Stolper, Catherine Wolfram and seminar participants at the UC Berkeley Energy Institute, the Midwest Energy Fest, the Heartland Workshop in Environmental and Resource Economics, Camp Resources, and the Association of Environmental and Resource Economists Summer Meeting. All errors are our own.

1 Introduction

Residential energy efficiency is widely considered to be one of the most cost-effective strategies for reducing greenhouse gas emissions (International Energy Agency, 2019; McKinsey & Co, 2009). As such, it has become central to climate policy around the world, with billions of dollars invested annually to unlock this potential (EEA, 2018; ARB, 2017; Barbose et al., 2013). Independent of climate change policy, many energy efficiency programs focus on other benefits such as reduced energy costs for low-income households and reduced pressure on energy generation capacity. However, these programs will be less cost-effective than anticipated if realized savings from energy efficiency fall short of expectations. Recent analyses have found that ex ante projections overestimate energy savings in home retrofit programs (Fowlie, Greenstone, and Wolfram, 2018; Allcott and Greenstone, 2017; Zivin and Novan, 2016), appliance rebate programs (Houde and Aldy, 2014; Davis, Fuchs, and Gertler, 2014), and building codes (Levinson, 2016).[1]

This paper examines this performance wedge – the discrepancy between ex ante projections and ex post savings – in the Illinois Home Weatherization Assistance Program (IHWAP), which is the Illinois implementation of the largest residential energy efficiency program in the United States. The program is intended to reduce energy bills for thousands of low-income households, thus contributing to equity, energy security, and resiliency against price fluctuations.[2] A wide range of utility and governmental programs, including IHWAP, rely on ex ante projections of energy reductions to allocate funding across homes and retrofits. Since many of these programs use similar tools, decomposing the wedge may be critical for increasing allocative efficiency of IHWAP and other large public programs.[3]

Despite increasing interest in understanding the performance wedge in energy effi-

[1] While not always recognized in climate policy reports such as McKinsey & Co (2009), internal evaluations from a subset of energy efficiency programs have found evidence of discrepancies between projected and realized savings since the early 1990s (e.g. Berry and Gettings, 1998; Dalhoff, 1997; Sharp, 1994; Nadel and Keating, 1991). Recognizing those discrepancies, improvements to the projections have been proposed and implemented in some cases (e.g. Ternes and Gettings, 2008; Ternes, 2007).

[2] Additional benefits from the program may include improved health, through better indoor air quality and comfort (Tonn, Rose, and Hawkins, 2018; Pigg, Cautley, and Francisco, 2018).

[3] Many programs rely on tools that use a common set of accepted structural engineering equations, thus are prone to similar biases (Edwards et al., 2013; Sentech, 2010).

ciency programs, the underlying factors remain unclear. This paper provides the first estimates of the impact of three primary channels that have been proposed to explain the wedge: 1) systematic bias in ex ante engineering measurement and modeling of savings, 2) workmanship, and 3) the rebound effect.[4] Policy implications can vary depending on which factors are most important. For example, the allocation of IHWAP funds (across and within homes) relies heavily on physical/structural models for ex ante projections, such that correcting systematic biases and improving projections through the use of utility data could improve the allocation of funds. Significant differences in workmanship in the program may warrant enhanced training, oversight protocols, or contractor incentives. Given a sufficiently large rebound effect, energy efficiency models may benefit from incorporating these effects in order to accurately account for the combined impact of treatment on energy use and household welfare.

We employ tree-based machine learning methods in an event study framework to predict per-home counterfactual energy usage. Burlig et al. (2017) and Souza (2019) find that the ML-based event study estimators increase robustness to model misspecification and bias from cohort effects. Our models make use of comprehensive administrative data for more than 9,800 homes served by 34 local weatherization agencies in Illinois between program years 2009 and 2016. The richness of our data and the ML-based approach enable a departure from prior work to examine fine-grained heterogeneity in the performance wedge and to estimate the home-specific cost-effectiveness of IHWAP energy-related investments. On average, we find that approximately 51% of projected *savings* are realized for the average home in the IHWAP program. This estimate falls within the range reported in prior literature: 38% for the WAP in Michigan (Fowlie, Greenstone, and Wolfram, 2018); 58% for a utility sponsored residential energy efficiency program in Wisconsin (Allcott and Greenstone, 2017); and 73% for insulation and air sealing measures from the Home Energy Services program in Massachusetts (MA-EEAC, 2018). We then use conditional averages of individual-specific wedge estimates to quantify the

[4]Workmanship refers to the quality of the implementation of retrofits, which results from the efforts of individual contractors, auditors, and agency quality control inspectors (QCI). The rebound effect refers to increased household consumption of energy services when they become less expensive due to efficiency improvements.

effects of model bias and workmanship.

Our results suggest that bias in model projections and workmanship are two major factors contributing to the performance wedge. Our estimates reveal that up to 41% of the wedge can be explained by discrepancies between ex ante projections and ex post savings in five major retrofits that collectively account for the vast majority of projected savings in IHWAP: air sealing, furnace replacement, wall insulation, attic insulation, and windows. In particular, the wedge is approximately 20 percentage points larger in homes that receive large expenditures on wall insulation (over $1,500), indicating an outsized impact from this measure. This suggests that improvements to model projections could be important for the optimal allocation of funds across measures.

To estimate heterogeneity in workmanship, we use lagged (prior years') contractor-specific effects on mean savings to create a prediction of current years' quality that is purged of the idiosyncratic unobservable features that may make homes easier or more difficult to weatherize. Our estimates indicate that if all workers performed at the level of the top 5%, then the wedge could be reduced by up to 43%. This is in line with prior work that has demonstrated that moral hazard and incentive problems can affect contractor efficacy in energy efficiency programs (Giraudet, Houde, and Maher, 2018; Blonz, 2018). It suggests a non-trivial role for reforms to improve workmanship, which could include changes to worker incentives, training, or other changes.

While the rebound effect is often touted as a potentially important contributor to the wedge, our results suggest that its effects are relatively modest and may not warrant any specific changes to program implementation. Using ex post data on the realized relationship between outdoor air temperature and energy consumption, we estimate that a standard rebound effect can account for up to 6% of the wedge.

Each of our estimates reflect upper bounds for the contributions of an individual channel, since the three mechanisms that we investigate can interact. Nevertheless, our results suggest that interactions between channels do not account for a large fraction of the wedge. For example, while we find that the top quintile of contractors may have a smaller performance wedge in wall insulation than the bottom quintile, that difference is

an order of magnitude smaller (2.1%) than the average wedge estimates for wall insulation (30%).

In the final section of the paper, we compare the disaggregated estimates of energy savings and home-specific expenditures to construct a measure of *net benefits* for each home in our sample. While recent advances in the program evaluation literature have emphasized heterogeneity in treatment effects, to our knowledge this is the first study to use unit-specific treatment effects to trace out the marginal benefits curve for expenditures made in a large public program (Davis and Heller, 2020; Wager and Athey, 2018; Athey and Imbens, 2017; Athey and Imbens, 2016). Our findings indicate that this is especially important in programs like the WAP, where costs and benefits vary substantially across treated households. This method allows us to identify the point where the energy-related benefits are equal to the costs of additional expenditures in the IHWAP sample. We also examine the relationship between home-specific net benefits and the performance wedge. This is important, since the presence of a wedge between projected and realized savings does not necessarily imply that an investment is not cost-effective.

This analysis reveals three key findings: (1) Approximately 42% of homes generate positive energy-related net benefits with substantial marginal benefits for top homes, suggesting an opportunity for improving the allocation of funds on the basis of energy or climate goals; (2) Homes falling in the lower tail of the marginal benefits curve have high net costs – these projects substantially reduce the overall cost-effectiveness of energy-related investments in the IHWAP; (3) We find no evidence of a correlation between net benefits and expenditure on wall insulation, which is consistent with lower-than-expected returns at high levels of spending on that measure.

The remainder of the paper is organized as follows. Section 2 describes the institutional background and the available data. Section 3 introduces the ML-based estimation strategy. Results on each of the channels affecting the wedge are discussed in Section 4. Section 5 reports results on heterogeneity in program cost-effectiveness. Section 6 concludes.

2 Institutional Setting and Data

2.1 Weatherization Assistance Program

The Weatherization Assistance Program (WAP) is the largest and most ambitious energy efficiency program in the United States. It is designed to lower energy bills and increase equity, energy security, and resilience to price fluctuations for low income families by providing energy efficiency improvements to their homes. It has provided weatherization assistance to over 7 million households since it began in 1976, many of whom live in "energy poverty." Prior evaluations report the success of the program along many of these important dimensions.[5] Unlike many energy efficiency programs, congressional funding allocations to the WAP are made on the basis of benefit-cost estimates derived from internal impact evaluations.

Through these internal evaluations, the WAP has long recognized that realized savings fall short of projections (e.g. Berry and Gettings, 1998; Dalhoff, 1997; Sharp, 1994). DOE has also sponsored qualitative studies to explore reasons for falling short of projections and identify opportunities for improved savings (e.g. Pigg, 2014; Berger, Lenahan, and Carroll, 2014) and then has implemented changes in response. For example, both Pigg (2014) and Berger, Lenahan, and Carroll (2014) identified work quality as a contributor to the gap, and in 2013 the WAP initiated a Quality Work Plan that mandated clear state standards requiring that every WAP job receive a final inspection by a Quality Control Inspector who had gone through accredited training and was certified by the Building Performance Institute (DOE, 2013). Other major contributors to the gap that were identified by these studies included malfunctioning heating systems and failure to use ex ante billing data to "true up" model estimates.

2.2 Weatherization Assistance Program Administrative Data

Through a partnership with Illinois' Department of Commerce & Economic Opportunity, we obtained comprehensive information from the program, including demograph-

[5]See, for example: Tonn, Rose, and Hawkins (2018), Pigg, Cautley, and Francisco (2018), Francisco et al. (2017), Tonn, Carroll, et al. (2014), Dalhoff (2013), Khawaja et al. (2006), Schweitzer (2005).

ics, housing characteristics, and IHWAP upgrades made for close to seventy thousand single-family homes in Illinois between program years 2009 and 2016. Households apply for the program by contacting one of the state's 35 'Local Administering Agencies' that are responsible for managing IHWAP within a given set of counties. Demographics collected during the IHWAP application include: household income and size (with a distinction between number of children or elderly occupants), applicant's age, sex, race, and tenancy. Age, sex, race, and tenancy are not used to determine eligibility, but are required to complete the application. To be eligible for IHWAP, a household's income must be less than 200% of the poverty line.[6] Further, a household is automatically eligible for WAP if: they qualified for the Low Income Home Energy Assistance Program (LIHEAP) within 1-year prior to the WAP application; if anyone in the household collects Social Security Disability (SSD) or Supplemental Security Income (SSI); or if they receive Temporary Assistance for Needy Families (TANF). Finally, households with elderly occupants (60 and over), young children (5 or below), or persons with disabilities are prioritized to receive WAP treatment earlier within a program year.

The home of a successful WAP applicant receives an extensive pre-treatment energy audit. Measurements collected during the energy audit include (but are not limited to): building airtightness; attic, wall and foundation insulation; window types and sizes; wall types; building orientation (to account for solar gains); floor area and height; and foundation type. These data also include the characteristics of mechanical systems such as furnaces, water heaters, and air conditioners, including: heating/cooling capacity, fuel type, draft type (for combustion appliances), efficiency rating, appliance location, duct location, duct type, duct leakage metrics, and ventilation flow rates. Data for smaller appliances include lighting type, number of light bulbs, and refrigerator age. Health and safety information (e.g. presence of ground cover in crawl spaces) and incidental repair information (e.g. updating gas piping to a furnace potentially slated for replacement) are

[6]During the American Recovery and Reinvestment Act (ARRA), the program was expanded such that households below 200% of the poverty line were eligible. Once ARRA funds were exhausted, the eligibility cutoff temporarily reverted to 150% before increasing again to 200%. ARRA mostly affected WAP in Illinois during program years 2010 to 2013.

also recorded.[7]

Data from the pre-treatment home energy audit become inputs to a DOE-approved engineering model, which generates estimates of the energy use of the home as is (pre-treatment) and with potential weatherization measures (post-treatment). The engineering model does not incorporate any energy billing data. The model's optimization scheme is designed to target program funds to the most cost-effective of all potential energy efficiency measures that are used to weatherize homes in the program. In practice, the optimization routine: (1) runs the engineering model to determine the most cost-effective measure that can be implemented in a home with characteristics recorded in the audit, (2) re-runs the model assuming that measure is implemented to determine the next-most cost-effective measure, (3) re-runs the model until all options are explored. Savings-to-investment ratios (SIRs) are determined for each measure using the cost of the measure, the cost of fuel, a uniform present worth factor provided by the federal government, the first-year projected savings of the measure, and the expected lifespan of the measure.[8] The DOE-approved system that is specific to Illinois is called "WeatherWorks."[9] The output from the WeatherWorks model consists of a list of measures in order from highest to lowest SIR. Measures with SIR values of 1.0 or greater are eligible for installation using DOE funds. Incidental repairs must be included in the SIR. Health and safety measures are often required to ensure that measures intended to save energy do not cause a health concern and are not subject to SIR eligibility since they are not targeting energy savings. Addressing health and safety in this way is part of the WAP mandate. Quantifying the benefits of those measures on health and safety outcomes is beyond the scope of this paper, thus their costs are excluded from cost-benefit analyses in section 5.

WAP policy requires the selection of measures from highest to lowest SIR until either

[7]Other examples of health and safety repairs include: expenditures on ventilation fans in kitchen and bathrooms; installing CO alarms in the home; power venting a water heater that is no longer drafting properly after HVAC upgrades.

[8]The uniform present worth factor assumes a 3% discount rate and adjusts for fuel price escalation, as proposed by the National Institute of Standards and Technology (Rushing, Kneifel, and Lippiatt, 2012). Expected lifespans range from 25 years for insulation measures, to 5 years for fluorescent lamps.

[9]Examples of other systems include the National Energy Audit Tool (NEAT) and Mobile Home Energy Audit (MHEA), developed by the DOE's Oak Ridge National Laboratory. There are also a number of other commercial and custom packages in use (Sentech, 2010).

(1) the available funding is exhausted or (2) there are no more remaining measures with SIRs of 1.0 or greater. WeatherWorks directly converts this list into a work order, which is provided to the contractor or in-house crew hired to implement the weatherization. The work order includes the *projected* labor and materials costs for the measure. After work is completed for a home, the *actual* costs (including labor and materials) for each measure are documented in the WeatherWorks database. Finally, every WAP job is subjected to a quality control (QC) inspection by a certified quality control inspector (QCI). The QCI can call back the contractor/crew to rectify problems if they find that measures were not installed correctly. In that case, WeatherWorks data are updated to reflect final costs of the job.

2.3 Energy Consumption Data

To test for the effects of WAP upgrades on energy usage, we obtained natural gas and electricity consumption data for over 9,800 WAP homes served by a major utility in Illinois. The utility primarily serves homes in the central and southwest regions of the state. Figure A.1 of the Appendix illustrates the geographic distribution of homes with available energy data. We focus on residences that use either natural gas or electricity as their main heating fuel.[10]

2.4 Weather Data and Supplementary Variables

Using geocoded addresses and daily weather variation from the PRISM Climate Group (2018), we extracted daily minimum temperature, maximum temperature, and precipitation for all homes in our energy data sample.[11] Daily weather observations were then aggregated to match the (monthly) billing cycle of each home.[12] We use minimum and maximum temperatures to construct measures of heating degree days (bases 60 and 65) and cooling degree days (base 75). We obtained state-level residential electricity and

[10] Data from the American Housing Survey from 2011 suggest that close to 64% of Illinois homes use natural gas as their primary heating fuel, while 22% use electricity (US Census Bureau, 2013).

[11] Addresses were geocoded with Google's geocoding API (Google, 2018). The PRISM Climate group compiles climate observations from various monitoring stations, validates those through rigorous quality control methods, and develops spatial interpolation climate models to produce estimates of weather variation at a 4km grid cell resolution for the U.S.

[12] We calculate average daily maximum temperature, minimum temperature and precipitation for all homes' 30-day billing cycles.

natural gas prices from the Energy Information Administration (EIA, 2017), which we use to compute average prices from 2009-2016 for cost-benefit calculations.

2.5 Summary Statistics for WAP Sample

Table 1 presents descriptive statistics for the main variables collected during the WAP application process and pre-treatment home energy audits. The top panel reports demographic data for the sample, which illustrate that treated households consist primarily of low-income families ($16,400 in average income). They are mostly middle-aged (52 years) homeowners (92%). The second panel of data on housing structure characteristics reveals that the program weatherized a highly diverse set of homes, characterized by wide variation in floor area, pre-treatment blower door tests, attic R-values, number of bedrooms, and vintage. Given this variation in housing type, it is not surprising that there is also considerable heterogeneity in retrofit-specific spending as indicated by the third panel. The bottom panel of Table 1 presents statistics for monthly natural gas and electricity consumption from homes served by IHWAP. Sample averages reveal that natural gas accounts for most of the energy demand in our sample. Nevertheless, there is significant variation of consumption from both fuels. Finally, the table also illustrates the significant variation in temperatures recorded during the billing cycles that we analyze.

Figure 1 compares the average energy consumption obtained from utility data with projections of pre- and post-treatment energy usage from the WeatherWorks model. The figure plots ratios between modeled usage and actual usage as a function of yearly pre-treatment usage averages (usage includes natural gas and electricity billing data). Ratios above 1 indicate that the engineering model overestimates a home's energy consumption. This overestimation can be severe in homes with lower usage, where the modeled usage may be overestimated by a factor of 5.5 pre-treatment and 4 post-treatment. The plot illustrates that modeled energy usage tends to differ greatly from raw consumption data in both the pre- and post-treatment periods.

Overestimation of usage will not *necessarily* result in bias in energy savings projections. Accurate savings estimates may be obtained in a special case where both pre- and post-treatment consumption are overestimated by the same amount, such that er-

rors cancel out. In the following sections, we show that that is not the case for IHWAP. We investigate the sources of error that result in biased projections of savings. We also examine the implications of this model bias.

3 Empirical Strategy

3.1 Machine Learning Estimates of Energy Savings

Our first goal is to estimate heterogeneity in the performance wedge, which we can then use to evaluate the importance of several mechanisms of interest. To that end, we employ a machine learning (ML)-based estimator which offers two advantages over traditional fixed-effects regression techniques. First, simulations from Souza (2019) show that the ML estimator that we employ recovers fine-grained treatment effect heterogeneity more efficiently than traditional regression techniques. This is because the ML approach is more flexible in the modeling of how our many fine-scale inputs, and complex interactions between them, affect energy consumption.

The second advantage is that the ML estimator does not suffer from biases identified in recent econometric literature on two-way fixed effects (TWFE) estimators (see, for example: Borusyak and Jaravel, 2017; Athey and Imbens, 2018; Goodman-Bacon, 2018; Strezhnev, 2018; Abraham and Sun, 2019; de Chaisemartin and D'Haultfœuille, 2020; Callaway and Sant'Anna, 2020). de Chaisemartin and D'Haultfœuille (2020) show that TWFE can be biased in the presence of significant heterogeneity of treatment effects across time or groups of treated units. That type of heterogeneity is expected within the context of this study since, for example, weatherization measures can differ substantially across homes. Also, specifically for the context of event studies with staggered rollout, Goodman-Bacon (2018) shows that TWFE estimators may be near-term biased because they place more weight on portions of the sample with higher variance of the treatment indicator variable (i.e. typically at the middle of the panel). Because the final step of our ML approach compares post-treatment predicted counterfactual usage to post-treatment realized usage, our estimates weigh observations equally so that they better reflect average effects throughout the whole post-treatment period.

Consistent with other recent applied work, we employ a machine learning model to predict counterfactual outcomes (energy usage) (e.g. Burlig et al., 2017; Poulos, 2019). Our ML algorithm is trained exclusively using *pre*-treatment observations. The model uses covariates related to housing structure, demographics and weather variation (presented in Table 1), as well as indicators for month- and year-of-sample to predict what the post-treatment consumption of homes would have been had they not received treatment. Yet-to-be treated WAP applicant homes are used to account for time-varying relationships between usage and all covariates to predict the counterfactual consumption in a home's post-treatment period. To select the algorithm/model with best predictive performance, we employ 5-fold *cross-validation*. First, we split our pre-treatment sample into 5 random partitions. We then recursively use 4 of those partitions to train/fit the ML algorithm, while the 5th (holdout) partition is used to assess prediction accuracy, measured by root-mean squared errors (RMSE). This process allows us to obtain proxy *out-of-sample* performance metrics of candidate algorithms/models. We find that the gradient boosted trees model (XGBoost, by Chen and Guestrin, 2016) has the lowest cross-validated RMSE among five competing models in predicting pre-treament energy use.[13] We therefore use the gradient boosted trees model to construct counterfactual estimates.

Once we obtain counterfactual predictions, we subtract them from realized post-treatment usage to obtain home-by-month treatment effects: $Y_{it}(1|D_{it}=1) - \hat{Y}_{it}(0|D_{it}=1) = \hat{b}_{it}^{ml}$.[14] The average of this measure reflects our estimate of the average treatment effect on the treated (ATT) (i.e. for households that have opted into the program). For our ATT estimate to be unbiased, it is essential that the energy consumption of treated versus not-yet-treated homes are on parallel trends, conditional on controls. This requires that there is nothing unobservable or uncontrolled for that affects energy use and

[13] Tree-based methods capture nonlinearities in the data through branch splits and automatically perform interactions between all variables, which is important for predicting energy use in a heterogeneous housing stock. The pool of models considered includes: ridge regression, elastic net, lasso, random forest, and XGBoost. Details on the characteristics, cross-validated (out-of-sample) performance, hyperparameter tuning, and prediction errors of each of the five models are provided in Appendix B.

[14] Here Y_{it} denotes energy consumption for home i in month t. D_{it} is a treatment indicator variable equal to one for all homes already exposed to WAP, zero otherwise. $Y_{it}(0)$ represents potential outcomes under control, while $Y_{it}(1)$ represents potential outcomes under treatment.

is correlated with the timing of treatment. For example, our estimates of program effects would be biased if households' decisions to upgrade were correlated with anticipated demand shocks that could occur at the time of weatherization. However, a feature of our setting is that it is difficult for a household to predict or control the exact timing of the upgrades. Once a house has been approved for the program and the auditor has selected the measures to be performed, the house goes into a job queue. The average wait-time between application and treatment is about four months. However, there is significant variation in wait-time such that some households wait no more than 20 days, while others wait more than a year.

Another concern may be that households served at different points in the program year are unobservably different from one another, such that those that apply later in the year are not adequate counterfactuals for those that apply earlier. We explore the possibility of this or other violations of the parallel trends assumption in Figure 2. This event-study graph depicts the average observed household consumption and the average ML predicted energy consumption in the months one year prior to and one year post treatment. We remove "work in progress" months during which upgrades are still being performed (months between audit and final inspection date), which causes the predicted usage not to be perfectly smooth across the treatment threshold. Nevertheless, the plot illustrates that the realized and counterfactual predicted usage are on parallel trends. If the yet-to-be treated homes were unobservably different than the treated homes in a way that biased the counterfactual estimates, then the post-treatment gap between realized and predicted usage would likely widen or narrow. Given that the gap is stable during the post-treatment period, it is unlikely that there are any major violations of the parallel trends assumptions.

Additionally, Figure 2, reveals that the difference between realized and predicted usage is negligible before treatment, indicating that the ML model provides accurate predictions on average. However, given that heterogeneity is the focus of our analysis, it is also important that the prediction errors are not correlated with measures performed, housing structure, or demographics. There would be systematic bias in the ML predic-

tions if, for example, there were unobservable determinants of energy consumption that were correlated with spending on particular measures. In Appendix B, we report the performance metrics we used to assess accuracy and rule out measure-specific bias for the ML algorithm. Notably, we plot out-of-sample prediction errors across all (binned) categories of program spending, as well as covariates related to housing structure and demographics. We find that errors are small and not significantly correlated with those factors. Importantly, this implies that the heterogeneity that we recover in subsequent sections is driven by systematic differences in treatment effects rather than systematic differences in ML prediction errors.

3.2 Estimates of the Performance Wedge

The outcome of interest in this paper is the performance wedge. We recover estimates of the wedge (in percentage points) by comparing the estimated treatment effects on the treated ($\hat{b}^{ml}_{it}$) to the engineering projections ($\hat{b}^{p}_{it}$):

$$\%\text{WEDGE}_{it} = \frac{\hat{b}^{p}_{it}}{\hat{Y}^{p}_{it}} - \frac{\hat{b}^{ml}_{it}}{\hat{Y}^{ml}_{it}} \quad \forall\, t > t_i \,, \tag{1}$$

where $\hat{Y}^{p}_{it}$ are counterfactual outcomes according to the engineering models, $\hat{Y}^{ml}_{it}$ are counterfactual estimates from machine learning predictions, and $t > t_i$ denotes all post-treatment dates for home i.[15]

Table 2 reports estimates of average program effects from engineering projections ($\hat{b}^{p}$) and from our machine learning approach ($\hat{b}^{ml}$).[16] Average savings according to engineering projections are close to 29%, which are almost double the ex post estimates (15%). That implies a realization rate of approximately 51%. In Appendix D we examine the robustness of these estimates to model specification (i.e. models in levels rather than

[15] The engineering model provides 2 predictions, representative of a full year of a home's energy usage: one year pre- and one year post-weatherization. From those, we calculate engineering projected savings. We divide by 12 to get monthly projections. More details about the engineering projections can be found in Appendix C.

[16] We estimate the average program savings as the sample average of $\hat{b}^{ml}_{it}$ obtained by the machine learning approach. For this analysis, we restrict the sample to billing data from at most two years post-treatment. Results are not sensitive to using only 1 year of post-treatment data. We focus on near-term effects and do not intend to capture depreciation of appliances or WAP-unrelated consumption changes that may happen long after treatment.

logs) and provide a comparison to standard two-way fixed effects (TWFE) estimates.

4 Decomposing the Performance Wedge

In this section, we explore three main factors that have been hypothesized to explain of the wedge between projected and realized savings in the WAP: (1) systematic errors in upgrade-specific projections of savings; (2) under-performance (workmanship) of WAP workers; (3) the rebound effect for treated households. We use our house-month specific wedge estimates to provide insight about which factors are most relevant in this context.

4.1 Wedge heterogeneity by upgrade-specific spending

We begin by focusing on whether we can identify systematic errors in upgrade-specific savings projections. Since prioritization of measures often depends on their ex ante savings-to-investment ratios (SIRs), modeling errors can lead to over(under)-investment in certain upgrades and result in misallocation of the budget available for a given home. For the purposes of this exercise, we focus on the measures that are considered by WAP to account for the vast majority of energy savings (air sealing, attic insulation, wall insulation, and furnace replacement) and reflect the primary costs in the program (attic insulation, wall insulation, furnace repair or replacement, and window replacement). Those measures collectively account for almost 68% of expenditures in the average home.

We analyze heterogeneity in the wedge by spending on each measure. A benefit of our empirical setting is that we observe all information that is documented by IHWAP, including a large set of parameters that the WeatherWorks model uses to determine expenditures including: housing structure, household demographics, weather variation, energy prices, and the contractor serving each home. This allows us to recover the mean wedge by spending category conditional on this rich set of factors using the following

regression:

$$\%WEDGE_{ijt} = \alpha_0 + \eta_j + \sum_{k=1}^{K}\sum_{b=1}^{B_k} \beta_{kb} \mathbb{1}[Category = k]_{it} \cdot \mathbb{1}[Bin = b]_{it} + \sum_{g=1}^{G} \gamma_g X_{it}^g + \varepsilon_{it} \quad \forall\, t > t_i\,, \quad (2)$$

where the performance wedge $\%WEDGE_{ijt}$ for home i, served by contractor j in the post-treatment billing cycle $t > t_i$, is defined by equation (1); α_0 is a constant; η_j are contractor-specific fixed effects. The coefficients of interest, β_{kb}, indicate the marginal change in the wedge associated with spending the amount in bin b on each measure k. The expression $\mathbb{1}[Category = k]_{it}$, indicates spending on measure k; $\mathbb{1}[Bin = b]_{it}$ indicates if the level of spending falls within a given bin b.[17] Other controls (X_{it}^g) include the complete set of factors in the WeatherWorks database that determine allocation of spending across measures (covariates from Table 1 related to housing structure and demographics), natural gas prices, electricity prices, and weather controls, including monthly average minimum/maximum temperatures, and monthly average precipitation. We introduce flexible controls for each of these variables (X_{it}^k). For this regression, as in section 3, we restrict the sample to observations no longer than two years post-treatment.

Given that our ML prediction errors do not systematically vary with the observable factors in model (2), the coefficients capture systematic bias in the projections from the WeatherWorks model at different levels of spending on major classes of retrofit. This bias can be driven by anything unobservable or uncontrolled for that systematically varies by spending in a given category. For the purposes of this analysis, we broadly define "model bias" as bias arising from: (1) errors in structural parameters in the WeatherWorks model, which may be optimistic; (2) systematic errors in the inputs to the model from mis-measurement in pre-weatherization audits; or (3) failure to capture the effects of unobservable features of the home that reduce the effectiveness of retrofits. Patterns in estimates from model (2) may also capture the effects of poor workmanship or household

[17] There are twelve categories of spending, and up to twelve bins for each category. Bins can vary in size, depending on the category being considered.

behavior, if the effects of either of these on the wedge systematically vary with spending on a particular category. In sections 4.2 and 4.3, we investigate how workmanship and the rebound effect contribute to the patterns we observe here. To the extent that both of those factors may be positively correlated with spending on a measure, the results in this section provide upper bound estimates of the contribution from systematic ex ante engineering measurement and modeling errors.

Figure 3 presents the coefficient estimates from equation (2), focusing on the five major program measures: air sealing, attic insulation, wall insulation, furnace repair or replacement, and windows.[18] The omitted category is zero spending on each measure, such that positive (negative) estimates indicate that the wedge increases (decreases) relative to the wedge at zero spending. Note that we employ stratified (by home) bootstrapping to obtain standard errors for all analyses in this paper. That implies carrying out all the steps of the analyses, including machine learning, for 200 bootstrap iterations. Then, the standard deviations for each point estimate over all of the bootstrap iterations will represent the standard errors of those estimates. Since we are interested in interpreting many coefficients from equation (2), we also apply false discovery rate (FDR) corrections from Benjamini and Hochberg (1995) to the p-values associated with estimated coefficients. Estimates that are significant after the FDR corrections are plotted in red with a square marker, while those that are non-significant after the corrections are plotted in blue with a triangle marker. The grey bars represent the number of homes in a given category of spending.

Our estimates reveal a substantial discrepancy between projected and realized savings in one category in particular: wall insulation. All estimated coefficients are positive and statistically significant, which indicates systematic upward bias in engineering projections on this measure. Point estimates suggest that the wedge in homes receiving investments of more than $300 in wall insulation is 13-20 percentage points higher than

[18] We present coefficients for demographics and housing controls in Appendix E.

in comparable homes with no investments on that measure.[19] Note that while homes with zero spending on wall insulation are included in the regression, we omit that bin from the histogram plot to better illustrate variation among the 30% of homes with non-zero spending on this measure.

The distribution of furnace spending is bimodal: furnace repairs, cleaning, and tuning usually costs less than $900, while complete furnace replacements start at about $1,800. Interestingly, the bias associated with furnace replacements, is somewhat negative (ranging from -2% to -4%). While the individual coefficients are not all statistically different from zero, taken together, the estimates for these higher-spending bins suggest that ex ante projections may *underestimate* the energy benefits from new furnaces. Coefficients on bins corresponding to furnace tune-ups, on the other hand, suggest the opposite: point estimates trend upward in spending. Estimates in the bins corresponding to spending between ($300–$900) are statistically and economically significant. However, we note that the density of homes falling in bins with the largest wedge ($900–$1,200) is relatively small. Furthermore, according to program guidelines, furnace tune-ups are often justified for health and safety reasons, rather than for energy-efficiency.

We also find a positive trend between the level of the performance wedge and spending on windows. Homes with high spending (above $1,400) on windows exhibit performance wedge that is 5 to 8 percentage points greater than the wedge at zero spending. Furthermore, the number of homes with large window spending is substantial (around 1,200). The patterns for air sealing and attic insulation are not as stark. The relationship for air sealing is almost flat, and, while the plot for attic insulation suggests that the wedge becomes larger at higher levels of spending, only the highest spending bin is associated with a statistically significantly higher wedge (5 percentage points). Because fewer homes fall into the higher spending bins of attic insulation, the effects on overall program performance are likely negligible.

We simulate the fraction of the total wedge that can be attributed to spending on

[19] This does not necessarily imply that wall insulation is associated with nonsignificant energy savings or that it is not cost-effective, as it is possible for measures to fall short of expected savings and still be both cost-effective and important drivers of savings. For example, Souza (2019), in the same context of this paper, finds that wall insulation can lead to natural gas savings ranging from 4 to 9 percent.

four of the most important program measures. The thought experiment behind the simulation asks what would happen to the wedge if any positive spending on each of these categories had no marginal effect on the wedge. We implement this by estimating the mean wedge for the sample if each of the coefficients associated with these spending categories were zero. We display the results of this simulation in Table 3 with bootstrapped standard errors in parentheses. The column labeled "Baseline" indicates that the mean performance wedge for the sample is 14.7%. Each of columns 1-5 shows the effect on the average performance wedge if the coefficients for positive spending were zero for each of the 5 measures in turn. Spending on all of the considered measures, except furnace replacements, is associated with a larger wedge relative to no spending on that measure, all else equal. The final column of the table shows the combined effect of doing this exercise for all of the measures. While fewer than 1/3 of homes receive wall insulation, the performance wedge would be 26% to 34% smaller for the full sample if there was no change in the wedge associated with implementing that measure. Our estimates suggest that the 5 major measures collectively account for up to 41% of the wedge, with a confidence interval of 24% to 57%.

4.2 Workmanship

We next explore the proportion of the wedge that can be explained by heterogeneity in workmanship. Our definition of "workmanship" includes not only contractors' performance, but also aspects of the pre-weatherization audits and the quality-control inspections, since either of these could systematically influence program outcomes.[20] We begin by estimating the contribution of workmanship to energy savings in each program year. Then, we run a simulation for a thought exercise that asks what would happen to the wedge if all workmanship was performed at the level of the "best" in the sample.

Estimating Heterogeneity in Workmanship

To construct a measure of workmanship, first we consider a variation of equation (2) with energy savings $\hat{b}_{it}^{ml}$ as the dependent variable. The differences among the coefficient

[20] Whereas the estimates reported in Figure 3 and Table 3 capture the effects of systematic mismeasurement in pre-weatherization audits (such as the level of pre-existing wall insulation), workmanship includes other potential unobservable effects of auditors on the performance of specific contractors.

estimates on the contractor fixed effects ($\hat{\eta_j}$) reflect mean contractor-specific differences in the energy savings. If there were no unobserved or uncontrolled for determinants of energy savings, then these coefficients could be interpreted as a measure of contractor j's workmanship quality (q_j). However, there may be unobserved factors that affect savings. Therefore, some of the differences in the coefficient estimates could be due to unobservable variation (ϵ_j) across the homes that to which contractors were assigned, rather than true differences in contractor quality as follows: $\hat{\eta_j} = q_j + \epsilon_j$.

In order to estimate the role of workmanship on the wedge, we would ideally isolate variation in performance driven by the quality of workers assigned to each job rather than systematic differences across homes' potential for energy savings. There were 97 unique contractors who served homes in our study sample. To isolate the contribution of each one, we create a contractor quality measure for each program year, based on contractors' mean savings from the previous program year. To estimate those mean savings, we use a variant of equation (2) with energy savings as the dependent variable and contractor fixed effects interacted with program year indicators. The associated coefficients give us fixed effect estimates $\hat{\eta}_{jy}$ for each contractor j and in each program year y. Then, to create a measure of contractor quality that is purged of idiosyncratic unobservable effects due to contractors' assignments to particular homes (ϵ_{jy}), we predict a contractor's year-specific quality using the coefficients from the following regression:

$$\hat{\eta}_{jy} = \alpha_0 + \delta\hat{\eta}_{jy-1} + \sum_{k=1}^{K}\sum_{b=1}^{B_k}\beta_{kb}\mathbb{1}[Category = k]_{it} \cdot \mathbb{1}[Bin = b]_{it} + \sum_{g=1}^{G}\gamma_g X_{it}^g + \varepsilon_{it} \quad \forall\, t > t_i\,, \quad (3)$$

with notation as defined in equation (2). This can be thought of as a "first-stage" instrumental variables regression where lagged quality is used as an instrument for contemporaneous quality. We use this regression to predict contemporaneous quality, $\tilde{\eta}_{jy}$, which will then be based solely on the observable factors about the job and the part of workmanship that can be explained by last year's performance.

To obtain an unbiased measure of quality, the component of a contractor's outcome that is not attributable to quality must not be correlated across years. We include a rich set of controls that contains all measurable components of a household/home that are available to the assigning agency. Further, since our sample is limited to the part of the state that is outside of Chicago, there is good overlap in the characteristics of the housing stock across contractors: single-family homes in areas that are outside of the urban metro area and that heat with piped natural gas rather than propane. Given this comparable housing stock, and our comprehensive set of controls, the component of a contractor's outcome that is not attributable to quality is likely idiosyncratic in any given year.[21] We provide several tests of this assumption below.

In Table 4, we provide results from the estimation of equation 3 (first column of panel A) as well as two robustness checks to rule out that systematic unobservable differences in housing stock may be driving the differences in our contractor quality measure. The first column indicates that a contractor's mean savings in a particular program year is strongly correlated with the previous program year. The second column of panel A adds flexible interactions among all the different categories of spending and between those categories and pre-treatment air-tightness. Including more flexible fixed effects helps to control for any systematic unobserved differences in contractor performance across years that may be driven by interaction effects between spending in particular categories and the wedge. If unobserved interaction effects were important and across years, housing stock differences across contractors could be a large biasing factor in our quality measure. However, as shown in Table 4, the point estimate on lagged quality changes little with the addition of these controls, suggesting that these are not a significant source of bias.

In Panel B of Table 4, we implement the approach developed by Oster (2019) to further assess the robustness of our quality estimates to unobservables. The first column of Panel B presents bias-corrected estimates of the coefficient on lagged contractor quality,

[21] Most jobs are assigned in sequential order rather than any characteristics about the home, measures assigned, or the contractors themselves. However, especially in smaller agencies, jobs may be assigned to a particular contractor on the basis of home characteristics. Our identification strategy addresses this concern in two ways: (1) any assignment outside of the queue is made based on observable characteristics that are in our WeatherWorks dataset and are included as controls; (2) we examine the overlap in the support of the distribution of measures performed across contractors and find that it is substantial.

with varying levels of "R-squared Max."[22] We find that the point estimate is stable, irrespective of assumptions. The second column reports a coefficient of proportionality, δ, which captures the relative importance of unobservables in our setting. For example, for our most conservative specification, we find that $\delta = 3$, which means that unobservables would need to be three times as important as observables to drive our estimates to zero. Thus, it is unlikely that unobserved factors correlated across program years are important drivers of our measure of contractor quality.

Finally, in Appendix F we assess balance of top versus bottom performing contractors in terms of spending on the five major upgrade categories. If our quality measure was driven by top and bottom contractors being systematically assigned to different types of homes, then one may also expect systematic differences in the distributions of spending on upgrades. However, figures F.1 through F.5 provide evidence against that. They reveal significant overlap and similarity between top and bottom contractors in terms of spending distributions. This suggests that our quality measure is not being driven by systematic unobserved differences in the types of homes to which contractors are assigned.

Simulations of the Effect of Workmanship on the Wedge

We use these estimates of contractor quality to quantify the effects of workmanship heterogeneity on the wedge. We implement a simulation in which we quantify what would happen to the average observed wedge if all contractors were assigned the predicted quality, $\tilde{\eta}_{jy}$, of the 95th, 90th, 75th, or 50th percentile of performers rather than their own predicted quality. Results are displayed in Table 5. The first column shows that the mean wedge would drop from 15.36% to 8.81% if all contractors performed at the 95th percentile of quality – a 43% decrease in the wedge, with a confidence interval of 28% to 57%. Likewise, the wedge would decrease by 32% or 16% if all contractors performed at the 90th or 75th percentiles of quality, respectively. These results indicate that workmanship is a key contributor to the wedge and suggest a role for policies that can improve worker performance by restructuring incentives or through training and inspection standards. We note that these large effects are consistent with an emerging

[22]The "R-squared Max" proposed by Oster (2019) denotes the maximum fit that a researcher may achieve for a model given the setting and data available.

body of evidence on the impact of contractor performance in other contexts (Giraudet, Houde, and Maher, 2018; Blonz, 2018).

Interactions Between Workmanship and Model Bias

We examine interaction effects between our measure of contractor quality and spending on wall insulation, which is the primary driver of the performance wedge in our context. It is possible, for example, that lower quality workmanship results in disproportionately negative effects at higher levels of spending on wall insulation. In that case, the pattern revealed in Figure 3 would reflect the bias in ex ante measurement and modeling as well as workmanship. To estimate the magnitude of the interaction, we add indicators for top and bottom quintile of contractor quality interacted with binned spending on wall insulation to equation (2). Figure F.6 from Appendix F reports the results, which show that the relationship between the wedge and spending on wall insulation remains unchanged for median contractors. Further, while bottom contractors are associated with a somewhat larger wedge, we do not find evidence that these effects increase with spending on wall insulation. Panel B in Table 5 reports results from simulations that remove those interaction effects by replacing the estimated coefficients with zero to recalculate the impact of contractor quality on the wedge. Contractor quality cannot explain a large fraction of the observed relationship between spending on wall insulation and the wedge – the overall effect of contractor performance only changes by about two percentage points. We consider interaction effects for other measures as well (not reported), but do not find any economically or statistically significant effects.

4.3 Household Energy Consumption Behavior

In this section, we examine how consumer behavior affects energy usage and the performance wedge. The focus of the WAP is to reduce energy needed for heating by implementing measures that are designed to improve furnace efficiency and the tightness of the building envelope. The primary behavioral channel through which occupants affect energy used for heating is through their choice of indoor air temperatures, or thermostat

set points.[23]

Looking at Figure 1, one might wonder how much of the observed bias is attributed to errors in the modeling of occupant heating/cooling demand versus to errors in modeling the contributions of given upgrades or the structural energy efficiency of a home. The WeatherWorks model assumptions about initial occupant set points are quite in line with recent observational studies and, if anything, would likely lead to under- rather than over-estimation of energy consumption. The model has two separate inputs for daytime and nighttime thermostat set points. For the early part of our data (2009–2013), the model assumed the indoor air temperature was set to 68^{o}F both during the day and at night. This is somewhat lower than observed pre-weatherization set points in a national study of homes served by WAP of 70.3 +/- 0.4 ^{o}F (Pigg, Cautley, Francisco, et al., 2014), suggesting that the assumption would likely lead to an underestimate of energy consumption. For the latter years (2013–2016), the daytime and nighttime set points were directly recorded by enumerators during the pre-treatment energy audit and used as inputs to the model.

The WeatherWorks model assumes that treatment has no effect on thermostat set point behavior. Therefore, 68 degrees is assumed to be the post-weatherization day and night time set point in the early years, and the measured pre-weatherization day and night time set points are assumed to be the post-weatherization set points in the latter years. But, households may change their behavior in response to weatherization. For example, households may choose to increase their indoor air temperature in response to reduced heating costs. This rebound effect would reduce net savings and thus contribute to the wedge between projected and realized savings. This section examines how changes in thermostat behavior affect realized energy savings and, in turn, the performance wedge.

To quantify the role of the changes in thermostat settings on the wedge, we take advantage of the fact that the amount of heat (ϕ_h) required to maintain a particular

[23]Secondary channels through which households could increase heat consumption include: opening doors or windows for extended periods during winter; or failing to maintain the furnace. We assume that households rarely open windows for extended periods during winter. Also, to impact the performance wedge, that behavior would have to change as a result of weatherization. Failing to replace furnace filters, while problematic for a furnace's functioning, has only a small impact on efficiency.

indoor thermostat setting is linear in outdoor air temperature (T_o) as follows:

$$\phi_h = H(T_b - T_o) \ , \tag{4}$$

where the slope, H, is a function of the surface area of the house, the thermal resistance of the wall and the furnace efficiency (See Appendix G). The "balance point" (T_b)–the outdoor air temperature at which the heating systems must be turned on to maintain a household's desired indoor temperature–has a direct mapping to indoor air temperature (T_i), such that: $T_b = T_i - (\phi_i + \phi_s)/H$.

The indoor air temperature is chosen by the household, while H, internal (ϕ_i) and solar (ϕ_s) gains are structural features of the house. Although all the terms vary over time within and across homes, the equation may be applied to mean values.[24] Figure 4 Panel A shows this relationship between average natural gas usage and outdoor air temperature for our full sample, both before and after weatherization. As engineering models predict, there is a linear relationship between energy consumption and outdoor air temperature, especially for colder months. The figure also reveals a clear change in slope (comparing pre- and post-treatment) during those months, which reflects the increased heating efficiency for homes after treatment. Post-treatment usage is generally lower than pre-treament at all temperature ranges, which suggests that baseload fuel-efficiency might have also increased.

The "kinks" in the curves from Figure 4 Panel A represent balance points. In order to estimate pre- and post-weatherization the balance points, we employ the PRInceton Scorekeeping Method (PRISM) (Fels, 1986).[25] Assuming that homes follow a linear relationship in outdoor temperature as described above, the PRISM method identifies the balance point for a given sample by regressing home-by-month energy usage on a

[24] In Appendix G, we present an engineering structural model which derives the relationship between indoor air temperature and residential space heating requirements (Johannesson et al., 1985).

[25] PRISM is a method primarily used in engineering to assess whether a home's usage pattern fits with an expected physical relationship between outdoor temperatures and energy usage. One intermediate step of this method estimates the home's heating balance point.

constant plus heating degree days (HDD), iterating through different HDD bases:

$$Y_{it} = \alpha + \beta \text{HDD}_{it}^{s} + \sum_{g=1}^{G} \gamma_g X_{it}^{g} + \varepsilon_{it} \,, \tag{5}$$

where Y_{it} is natural gas usage for home i in billing cycle (month) t; α is a constant; HDD_{it}^{s} are heating degree days for iteration s; X_{it}^{g} are housing and demographic controls; and ε_{it} is an error term. We run the regression (5) separately for pre- and post-weatherization samples, and we iterate through HDD bases from 55 to 65. Results are presented in Figure 4 Panel B, which plots PRISM regression R-squares for each HDD base iteration. The bases with highest R-squared are selected as the balance points for the samples. The figure reveals a slight increase in balance point of 0.6 degrees from 61.2^{o}F pre-weatherization to 61.8^{o}F post-weatherization.

Since the increase in balance point is a combination of both behavioral and structural factors, it cannot be used to directly assess the rebound effect. Ideally, to quantify the rebound effect, the researcher would like a measure of the average indoor temperature for each household before and after weatherization. While we do not have measures of indoor air temperature for homes in our sample, Pigg, Cautley, Francisco, et al. (2014) performed direct pre- and post-treatment measurements for a closely related WAP population. They hung indoor temperature (and relative humidity) data loggers from the main thermostat in homes served by the WAP across 35 states. They took snapshots of indoor conditions every 10 minutes for a study period that included both pre- and post-weatherization dates. They found a small but statistically significant mean increase of 0.3±0.2 ^{o}F across all hours of the day for weatherized homes.[26] Based on this work, we consider increases in mean monthly indoor temperature from 0.2 to 0.6^{o}F to quantify the rebound effect on the performance wedge. Given that the relationship in equation 4 applies to mean values aggregated across homes and over time, we must assume that set points are not a function of outdoor air temperature, and we can be agnostic as to how exactly occupants

[26] Fowlie, Greenstone, and Wolfram (2018) measure indoor air temperatures for both weatherized and unweatherized homes in southeastern Michigan and also found a small and statistically non-significant increase in daytime set points of 0.67^{o}F.

are changing their set points across the hours, days, and bill-months.[27]

To quantify the rebound effect, we simulate the average performance wedge assuming that households had not increased their indoor air temperature after weatherization. The simulations consist of using the PRISM model to predict *post-weatherization* energy consumption for homes in our sample, which are then subtracted from the ML counterfactual predictions to obtain home-specific savings, which in turn are used to estimate the performance wedge as described in section 3.2. For the baseline simulation, we employ a PRISM model with a balance point of 61.8^{o}F. Our alternative scenarios consist of lowering that balance point by 0.2, 0.4, and 0.6^{o}F. That exercise can be used to estimate the impact of "removing the rebound effect," since changes in indoor air temperature map to changes in balance point one-for-one. We note that lower balance points will translate into decreased energy consumption, which in turn decreases the performance wedge. Results from these simulations are presented in Table 6. We find that removing the rebound effect would decrease the performance wedge by 6.35% $\pm$ 0.35% for an indoor air temperature decrease of 0.4^{o}F.[28]

The results in this section suggest that the rebound effect can reduce savings by almost 8.5%, explaining a small but non-trivial fraction of the wedge – approximately 6%. As a point of comparison, in Appendix G.2, we estimate that roughly 83% of total program savings come solely through improvements in furnace efficiency and building envelope tightness, i.e. the change in slope from the temperature response function in Figure 4A. The behavioral effects examined here have a small enough effect on the wedge that even if the rebound effect occurred exclusively for homes that received wall insulation, it could not explain the bias that we observe in the projected savings for that category. Therefore, even after controlling for behavioral factors, ex ante engineering modeling errors likely play an important role in a significant portion of the wedge.

[27] Figure 4 Panel A suggests that the relationship between consumption and outdoor air temperature remains linear during cold months, providing strong support that set points are not a function of outdoor air temperature.

[28] This 0.4^{o}F reduction in rebound associated with a 1 percentage point reduction in the wedge is consistent with a 1^{o}F reduction in indoor temperature is associated with a 3 percentage point decrease in natural gas usage from Fowlie, Greenstone, and Wolfram (2018). Similar effects have also been found by the American Council for an Energy Efficient Economy.

In Appendix G.3, we investigate two other behavioral mechanisms which may affect the performance wedge: (1) the existence of non-working furnaces prior to WAP treatment; (2) prolonged periods of building vacancy. We add indicator variables for both of those factors to our wedge specification (2). Results suggest that neither significantly affect our estimates of the wedge.

5 Heterogeneity in Cost-Effectiveness

In this final section, we develop home-specific estimates to better understand heterogeneity in overall program performance and to examine the relationship between cost-effectiveness and the performance wedge. To estimate home-specific energy savings, we take the mean of our monthly *post-weatherization* savings predictions ($\hat{b}_{it}^{ml}$), conditional on the rich set of factors we observe for that home: structural characteristics, spending on each upgrade category, household demographics, weather variation, and energy prices.[29] We estimate a version of equation (2) that uses savings ($\hat{b}_{it}^{ml}$) rather than the wedge as the outcome and excludes contractor fixed effects. We then use the estimated coefficients to predict a home's monthly savings. For each home, we take the mean of these predictions for each calendar month and then sum those means across all 12 calendar months, resulting in estimates of home-specific annualized savings. The exact specification is shown in Appendix H.

This approach offers two advantages for our cost-benefit analysis. First, it provides home-specific estimates of the energy savings that are purged of unobserved variation due to household-specific behavior. Note that the home-specific savings estimates in the procedure above rely only on observed variation in covariates, such that they will not reflect circumstances where the residing household uses more energy than would be "typical" or happened to change their behavior substantially over the study period. A second advantage is that policymakers can use these estimates to target future program

[29] In this section, we restrict the sample to homes with at least a full year of post-treatment data to allow for a comparison across homes. Homes for which we only observe winter months mechanically exhibit higher savings, since the effects of the program are stronger during winter. Appendix H presents results without sample restrictions, and also results without homes for which there is a poor correlation between energy use and weather, as determined by the PRISM method.

spending based on the standard data they collect on the house, household, contractor and projected weather.

We take our estimates of home-specific annualized savings and use the following assumptions to compute home-specific net present benefits from IHWAP expenditures, $\widehat{NPV(b)}_i^{ml}$. Benefits accrue through a stream of monthly energy savings attributed to IHWAP measures.[30] Based on analyses presented in Appendix D, we assume that 83% of savings accrue from natural gas and the remainder from electricity. We then estimate the monetized value of energy savings based on social marginal costs of gas and electricity as described in Davis and Muehlegger (2010) and Borenstein and Bushnell (2018), respectively.[31] The resulting social marginal benefits of reductions are, on average, \$6.76 per MMBtu for natural gas and \$34.35 per MMBtu for electricity. We also provide a calculation of the private marginal benefits to IHWAP-served households using retail energy prices, which is the metric used by the WAP program and in the WeatherWorks model. The retail residential energy prices for Illinois over our sample period (2009-2016) were, on average, \$10.47 per MMBtu for natural gas and \$34.66 per MMBtu for electricity (EIA, 2017).[32]

The expected lifespan of a given upgrade determines the total number of months across which benefits accumulate. For our baseline scenario, expected lifespans for most individual upgrades were obtained from the WeatherWorks documentation. They range from 5 years for fluorescent lamps to 20 years for furnace replacements. However, for insulation we consider longer lifespans. Whereas WeatherWorks assumed a 25-year lifespan for insulation, recent engineering literature suggests that insulation measures have substantially longer lifespans, such as 50 years for cellulose fiber (ISOCELL GmbH, 2014),

[30] Our estimates of net present benefits must be viewed as being applicable to IHWAP's implementation and not necessarily to other states. WAP in other states may have other funding sources and potentially different program practices.

[31] We calculate marginal private costs of natural gas for each month-of-year, based on month-of-year average citygate prices in Illinois from 2009-2016. We then add to those prices a social cost of carbon of \$40 per ton. Emissions factors were obtained from EPA (1998). For electricity, we use data provided by Borenstein and Bushnell (2018) to estimate the difference between retail prices and social marginal costs for the areas of the state which we analyze. We then apply that difference to the month-of-year averages of residential electricity prices, again for 2009-2016.

[32] For both approaches, we also apply price escalation based on indices from Rushing, Kneifel, and Lippiatt (2012), which increase yearly after the first year since treatment.

35-50 years for expanded polystyrene (EPS) (EUMEPS, 2017; IVH, 2015), or the full building lifetime for extruded polystyrene (XPS) (50-150 years) (EXIBA, 2019).

To account for the fact that homes receive unique bundles of upgrades, we calculate weighted averages of those lifespans using the expenditures made on each retrofit. The resulting weighted average of retrofit lifespans for an average home in our sample is approximately 20 years, after which upgrades are assumed to fully depreciate. To obtain the present value of benefits, we assume a baseline discount rate of 3%, which is the rate recommended by the US Department of Energy and used by WeatherWorks (Rushing, Kneifel, and Lippiatt, 2012). Finally, we subtract costs of all weatherization upgrades, excluding measures that are implemented specifically for health and safety reasons, to obtain estimates of net present benefits for energy-related expenditures.[33]

Figure 5 ranks homes according to net present benefits evaluated at the social marginal cost of energy (a) and at retail energy prices (b). The figure illustrates enormous heterogeneity in the net benefits of energy-related expenditures across homes, demonstrating the importance of considering the marginal returns of IHWAP investments. Vertical black lines identify the point where marginal benefit from an additional IHWAP investment in energy benefits equals zero (MC=MB), which occurs at the 58th percentile in panel A and at the 47th percentile in panel B. Investments in homes performing above these percentiles collectively have a benefit-cost ratio of 1.36 for Panel A, and 1.46 for Panel B. The figure also highlights the importance of outcomes in the tails, where IHWAP projects generate substantial gains and losses. Whereas a dollar of spending in the highest-performing quartile returns approximately $1.55 in energy benefits, the same dollar returns three times less ($0.45) when allocated to homes in the lowest quartile. Homes in the top 10th percentile each generate more than two thousand dollars in net benefits from energy reductions, while homes in the bottom 10th percentile each generate a net cost of more than two thousand dollars. The average IHWAP project generates net energy-related social benefits of $-325 when estimated using baseline assumptions,

[33] Approximately $550 or 10.5% of total expenditures are allocated to on health and safety measures in the average home in our sample. Health and safety expenditures are generally not expected to produce energy savings and are omitted from our cost-benefit analysis since we are not able to measure or account for any benefits from health and safety measures.

though the energy-related benefits for the sample become positive when excluding just 10% percent of the worst-performing homes. Using retail energy prices, our estimates indicate that the average IHWAP-treated household receives $232 in private (energy-related) benefits.

We evaluate the sensitivity of net benefits estimates to assumptions regarding retrofit lifespans (10-40 years) and discount rates (0-6%) (See Appendix Table H.1). The sign of *total net present benefits* in the IHWAP sample is sensitive to these ranges of values for both parameters when holding others at baseline values. In terms of social net benefits, we find: (a) 6% versus 0% discount rates result in estimates of $-7.52 versus $+9.58 Million; and (b) 10 versus 20 year retrofit lifespans result in estimates of $-12.76 to $-5.80 Million. Sensitivity to retrofit lifespans is particularly important to consider, given documented uncertainty in the lifespan of long-lived materials such as insulation. An 80-year lifespan for insulation increases the combined retrofit lifespan for the average home in our sample to just under 40 years, which corresponds to a total net benefits estimate of $+2.07 Million.

Overall, these findings caution against using a single sample average to draw conclusive statements regarding the cost-effectiveness of the IHWAP. While subject to uncertainty, the home-specific estimates indicate that certain types of projects are highly cost-effective, and that there is a potential role for targeting energy/climate investments based on marginal benefits. When considered as part of a greenhouse gas abatement strategy, we find that net benefits imply abatement costs of $7.65 per ton of CO2 for the average home in the IHWAP sample (see Appendix Table H.5 for more details). Net benefits for homes in the top quartile of the IHWAP sample imply abatement costs of $-39.4 per ton of CO2, which is among the most cost-effective investments available today (Gillingham and Stock, 2018).

Finally, we dig deeper into some factors associated with home-specific net benefits (see Appendix H.2). We find that the most cost-effective homes, on average, have lower expenditures, particularly on windows. Conversely, top homes spend more on insulation measures, with the exception of wall insulation. For this measure, we find no evidence

of a correlation between performance and expenditures, which is consistent with lower-than-expected returns at high levels of spending. The top-performing homes also still exhibit a substantial performance wedge. Results in Appendix H.2 also demonstrate that average net benefits were higher for homes served by IHWAP for program years 2013 and beyond. During those years, multiple quality improvement changes were implemented to the program, which may have yielded substantial benefits. While our research design does not allow us to explain these differences, this could be a fertile area for future research.

6 Conclusion

Evaluations of a wide range of energy efficiency programs consistently find a wedge between ex ante projected and ex post realized savings. This paper examines the role of three hypothesized channels: 1) systematic bias in ex ante engineering modeling of savings, 2) workmanship, and 3) occupant behavioral responses. To quantify the effects of each of these channels, we employ novel machine learning techniques that allow us to recover home-specific estimates of both realized savings and the performance wedge in the Illinois Home Weatherization Assistance Program.

We explore bias in projected savings for the five investments that combine to account for the vast majority of expected energy savings in IHWAP: air sealing; attic insulation; furnace repair or replacement; wall insulation; and window replacement. Taken together, we estimate that ex ante engineering measurement and modeling bias across these five measures can explain up to 41% of the wedge, a large fraction of which can be attributed to overestimated savings in one retrofit class: wall insulation. Further, we find significant heterogeneity in workmanship. If all workmanship were performed at the level of the top 5th percentile in terms of quality, then the wedge could be reduced by up to 43%. Finally, our results suggest that the rebound effect is a relatively modest contributor to the wedge–up to 6%.

We then evaluate the cost-effectiveness of investments made on each home in our sample. While other studies have recovered heterogeneous effect of energy efficiency programs, to the best of our knowledge this is the first study to use estimates of home-

specific treatment effects to trace-out a marginal benefits curve. This methodological advance has important implications. Our results indicate that the energy-related social benefits of investments in the top 42% of homes, and the private benefits in the top 53% of homes, exceed their costs to the program and collectively have a highly attractive cost-benefit ratio close to 1.4. While WAP does not prioritize treatment on the basis of energy-related benefits alone, this result suggests a key role for targeting investments when funds are allocated on the basis of expected energy/climate benefits. Our estimates reveal that investments in the highest performing homes have lower abatement costs than most available greenhouse gas mitigation strategies. They further indicate that performance in the lower tail significantly reduces the overall cost-effectiveness of the IHWAP.

These findings have the following policy implications: First, while prior literature has mostly emphasized systematic bias in accounting properly for engineering relationships, our results suggests that workmanship is a significant contributor to the existence of a performance wedge. This suggests an important role for re-structuring worker incentives or improving performance through training programs. Second, even though the rebound effect is often considered to be a potentially important contributor to the wedge, we find that its effects are relatively modest and may not warrant any specific changes to program implementation. In addition, our results reveal areas where focused improvements in ex ante models may lead to better allocation of IHWAP program funds. The majority of model bias appears to be explained by a single measure: wall insulation. Efficiency programs like the WAP may therefore need to improve ex ante measurement of existing wall insulation or better calibrate the model of predicted savings from wall insulation retrofits. Finally, while IHWAP already aims to target funds to the more cost-effective measures, our heterogeneity analysis suggests that spending in some measures may not be at their optimal levels. Therefore, reevaluating measure selection practices could systematically improve overall program performance.

References

Abraham, Sarah and Liyang Sun (2019). "Estimating Dynamic Treatment Effects in Event Studies with Heterogeneous Treatment Effects". *Working Paper*.

Allcott, Hunt and Michael Greenstone (2017). "Measuring the Welfare Effects of Residential Energy Efficiency Programs". *NBER Working Paper*(23386).

Athey, Susan and Guido Imbens (2016). "Recursive partitioning for heterogeneous causal effects". *Proceedings of the National Academy of Sciences* 113(27), pp. 7353–7360.

Athey, Susan and Guido Imbens (2018). "Design-based Analysis in Difference-In-Differences Settings with Staggered Adoption". *arXiv Working Paper*.

Athey, Susan and Guido W Imbens (2017). "The state of applied econometrics: Causality and policy evaluation". *Journal of Economic Perspectives* 31(2), pp. 3–32.

Barbose, Galen L, Charles A Goldman, Ian M Hoffman, and Megan Billingsley (2013). "The future of utility customer-funded energy efficiency programs in the USA: projected spending and savings to 2025". *Energy Efficiency* 6(3), pp. 475–493.

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". *Journal of the Royal Statistical Society. Series B (Methodological)* 57(1), pp. 289–300.

Berger, Jacqueline, Tim Lenahan, and David Carroll (2014). "National Weatherization Assistance Program Process Field Study: Findings from Observations and Interviews at 19 Local Agencies Across the Country". *ORNL/TM-2014/304, Oak Ridge National Laboratory, Oak Ridge, Tennessee*.

Berry, Linda G. and Michael B. Gettings (1998). "Realization Rates of the National Energy Audit". *Thermal Performance of the Exterior Envelopes of Buildings VI Conference*.

Blonz, Joshua (2018). "The Welfare Costs of Misaligned Incentives: Energy Inefficiency and the Principal-Agent Problem". *University of California, Berkeley. Energy Institute. Working Paper* 297.

Borenstein, Severin and James B Bushnell (2018). *Do Two Electricity Pricing Wrongs Make a Right? Cost Recovery, Externalities, and Efficiency*. Working Paper 24756. National Bureau of Economic Research.

Borusyak, Kirill and Xavier Jaravel (2017). "Revisiting Event Study Designs". *SSRN Working Paper*.

Burlig, Fiona, Christopher Knittel, David Rapson, Mar Reguant, and Catherine Wolfram (2017). "Machine Learning from Schools about Energy Efficiency". *NBER Working Paper*(23908).

California Air Resources Board (2017). *California's 2017 Climate Change Scoping Plan*. [Online; accessed in 2020].

Callaway, Brantly and Pedro H. C. Sant'Anna (2020). "Difference-in-Differences with Multiple Time Periods". *Journal of Econometrics*.

Chen, Tianqi and Carlos Guestrin (2016). "XGBoost: A Scalable Tree Boosting System". *arXiv:1603.02754*.

Council, Massachusetts Energy Efficiency Advisory (2018). "Massachusetts Home Energy Services: Impact Evaluation (RES 34)". *Technical Report, produced in collaboration between Navigant Consulting and Cadeo*.

Dalhoff Associates (2013). "Report on the Impacts and Costs of the Iowa Low-Income Weatherization Program: Calendar Year 2013". *Technical Report*.

Dalhoff, Gregory K. (1997). "An Evaluation of the Performance of the NEAT Audit for the Iowa Low-Income Weatherization Program". *1997 Energy Evaluation Conference*.

Davis, Jonathan MV and Sara B Heller (2020). "Rethinking the benefits of youth employment programs: The heterogeneous effects of summer jobs". *Review of Economics and Statistics* 102(4), pp. 664–677.

Davis, Lucas W., Alan Fuchs, and Paul Gertler (2014). "Cash for Coolers: Evaluating a Large-Scale Appliance Replacement Program in Mexico". *American Economic Journal: Economic Policy* 6(4), pp. 207–38.

Davis, Lucas W. and Erich Muehlegger (2010). "Do Americans consume too little natural gas? An empirical test of marginal cost pricing". *The RAND Journal of Economics* 41(4), pp. 791–810.

de Chaisemartin, Clément and Xavier D'Haultfœuille (2020). "Two-Way Fixed Effects Estimators with Heterogeneous Treatment Effects". *American Economic Review* 110(9), pp. 2964–96.

Edwards, J., D. Bohac, C. Nelson, and I. Smith (2013). "Field Assessment of Energy Audit Tools for Retrofit Programs". *Technical Report, National Renewable Energy Laboratory*.

European Extruded Polystyrene Insulation Board Association (2019). "Extruded Polystyrene (XPS) Foam Insulation with halogen free blowing agent". *Institut Bauen und Umwelt e.V. (IBU)*.

European Manufacturers of Expanded Polystyrene (2017). "Expanded Polystyrene (EPS) Foam Insulation (shape moulded, density 25 kg/m^3)". *Institut Bauen und Umwelt e.V. (IBU)*.

Executive Office of Energy and Environmental Affairs (2018). *Massachusetts Global Warming Solutions Act: 10-Year Progress Report*. [Online; accessed in 2020].

Fels, Margaret F. (1986). "PRISM: An introduction". *Energy and Buildings* 9(1), pp. 5–18.

Fowlie, Meredith, Michael Greenstone, and Catherine Wolfram (2018). "Do Energy Efficiency Investments Deliver? Evidence from the Weatherization Assistance Program". *The Quarterly Journal of Economics* 133(3), pp. 1597–1644.

Francisco, P. W., D. E. Jacobs, L. Targos, S. L. Dixon, J. Breysse, W. Rose, and S. Cali (2017). "Ventilation, indoor air quality, and health in homes undergoing weatherization". *Indoor Air* 27(2), pp. 463–477.

Gillingham, Kenneth and James H Stock (2018). "The cost of reducing greenhouse gas emissions". *Journal of Economic Perspectives* 32(4), pp. 53–72.

Giraudet, Louis-Gaëtan, Sébastien Houde, and Joseph Maher (2018). "Moral hazard and the energy efficiency gap: Theory and evidence". *Journal of the Association of Environmental and Resource Economists* 5(4), pp. 755–790.

Goodman-Bacon, Andrew (2018). *Difference-in-Differences with Variation in Treatment Timing*. Working Paper 25018. National Bureau of Economic Research.

Google (2018). *Google Maps Platform: Geolocation API*. [Online; accessed in 2018].

Houde, Sébastien and Joseph E. Aldy (2014). "Belt and Suspenders and More: The Incremental Impact of Energy Efficiency Subsidies in the Presence of Existing Policy Instruments". *NBER Working Paper*(20541).

Industrieverband Hartschaum e.V. (2015). "EPS rigid foam (Styropor®) for ceilings/floors and as perimetric insulation (in German)". *Institut Bauen und Umwelt e.V. (IBU)*.

International Energy Agency (2019). *World Energy Outlook*. Tech. rep. Organization for Economic Co-operation and Development.

ISOCELL GmbH (2014). "Blown insulation made of cellulose fibre". *Bau EPD GmbH*.

Johannesson, G., L. Agnoletto, G. Anderlind, B.R. Anderson, R.D. Godfrey, K. Kimura, O. Lyng, C. Roulet, H.C. Sorensen, and H. Werner (1985). "The Calculation of Space-Heating Requirements for Residential Buildings". *Buildings III Conference Proceedings* 89.

Khawaja, M. S., A. Lee, M. Perussi, and E. Morris (2006). "Ohio Home Weatherization Assistance Program Impact Evaluation". *Technical Report, Quantec, LLC*.

Levinson, Arik (2016). "How Much Energy Do Building Energy Codes Save? Evidence from California Houses". *American Economic Review* 106(10), pp. 2867–94.

McKinsey & Co (2009). "Pathways to a low-carbon economy: Version 2 of the global greenhouse gas abatement cost curve". *Technical Report, McKinsey & Company*, pp. 1–165.

Nadel, Steven and Kenneth Keating (1991). "Engineering Estimates vs. Impact Evaluation Results: How do they compare and why?" *American Council for an Energy-Efficient Economy, Research Report*.

Oster, Emily (2019). "Unobservable Selection and Coefficient Stability: Theory and Evidence". *Journal of Business & Economic Statistics* 37(2), pp. 187–204.

Pigg, S., D. Cautley, and P. W. Francisco (2018). "Impacts of weatherization on indoor air quality: A field study of 514 homes". *Indoor Air* 28(2), pp. 307–317.

Pigg, Scott (2014). "National Weatherization Assistance Program Impact Evaluation: A Field Investigation of Apparent Low and High Savers". *ORNL/TM-2014/308, Oak Ridge National Laboratory, Oak Ridge, Tennessee*.

Pigg, Scott, Dan Cautley, Paul Francisco, Beth A Hawkins, and Terry M Brennan (2014). "Weatherization and Indoor Air Quality: Measured Impacts in Single Family Homes Under the Weatherization Assistance Program". *ORNL/TM-2014/170, Oak Ridge National Laboratory, Oak Ridge, TN (United States)*.

Polley, Eric, Erin LeDell, Chris Kennedy, Sam Lendle, and Mark van der Laan (2018). "SuperLearner: Super Learner Prediction". *The Comprehensive R Archive Network (CRAN)*.

Poulos, Jason (2019). "State-Building through Public Land Disposal? An Application of Matrix Completion for Counterfactual Prediction". *Working Paper*.

PRISM Climate Group, Oregon State University (2018). *PRISM Climate Data*. [Online; accessed in 2018].

Rushing, Amy S., Joshua D. Kneifel, and Barbara C. Lippiatt (2012). "Energy Price Indices and Discount Factors for Life-Cycle Cost Analysis – 2012: Annual Supplement to NIST Handbook 135 and NBS Special Publication 709". *NIST Interagency/Internal Report (NISTIR)* 15(n29).

Schweitzer, Martin (2005). "Estimating the National Effects of the U.S. Department of Energy's Weatherization Assistance Program with State-Level Data: A Metaevaluation Using Studies from 1993 to 2005". *ORNL/CON-493, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

Sentech Inc (2010). "Review of Selected Home Energy Auditing Tools: In Support of the Development of a National Building Performance Assessment and Rating Program". *Technical Report, DOE's Office of Energy Efficiency and Renewable Energy.*

Sharp, T.R. (1994). "The North Carolina Field Test: Field Performance of the Preliminary Version of an Advanced Weatherization Audit for the Department of Energy's Weatherization Assistance Program". *ORNL/CON-362, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

Souza, Mateus (2019). "Predictive Counterfactuals for Treatment Effect Heterogeneity in Event Studies with Staggered Adoption". *SSRN Working Paper Series.*

Strezhnev, Anton (2018). "Semiparametric weighting estimators for multi-period difference-in-differences designs". *Working Paper.*

Ternes, Mark P. (2007). "Validation of the Manufactured Home Energy Audit (MHEA)". *ORNL/CON-501, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

Ternes, Mark P. and Mike B. Gettings (2008). "Analyses to Verify and Improve the Accuracy of the Manufactured Home Energy Audit (MHEA)". *ORNL/CON-506, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

Tonn, B., E. Rose, and B. Hawkins (2018). "Evaluation of the U.S. department of energy's weatherization assistance program: Impact results". *Energy Policy* 118, pp. 279–290.

Tonn, Bruce, David Carroll, Scott Pigg, Michael Blasnik, Greg Dalhoff, Jacqueline Berger, Erin Rose, Beth Hawkins, Joel Eisenberg, Ferit Ucar, Ingo Bensch, and Claire Cowan (2014). "Weatherization Works–Summary of Findings from the Retrospective Evaluation of the U.S. DOE's Weatherization Assistance Program". *ORNL/TM-2014/338, Oak Ridge National Laboratory, Oak Ridge, Tennessee.*

US Census Bureau (2013). *American Housing Survey for the United States: 2011.* [Online; accessed in 2018].

US Department of Energy (2013). "Weatherization Program Notice 14-4". *Published by the National Association for State Community Services Programs.*

US Energy Information Administration (2017). *Residential Electricity and Natural Gas Prices.* [Online; accessed in 2019].

US Environmental Protection Agency (1998). "AP 42, Fifth Edition Compilation of Air Pollutant Emissions Factors, Volume 1: Stationary Point and Area Sources". *Technical Report.*

Wager, Stefan and Susan Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests". *Journal of the American Statistical Association* 113(523), pp. 1228–1242.

Zivin, Joshua G. and Kevin Novan (2016). "Upgrading Efficiency and Behavior: Electricity Savings from Residential Weatherization Programs". *The Energy Journal* 37(4).

Tables

Table 1: Descriptive Statistics for Main Variables in the Study

	Average	Standard Deviation	Min	Max
Demographics				
Income($/1000)	16.36	10.12	0.00	50.53
N Occupants	2.67	1.63	1.00	9.00
Householder Age	51.83	16.19	22.00	89.00
Female Householder (%)	68.33	46.52	0.00	100.00
Renter(%)	8.03	27.17	0.00	100.00
White(%)	72.32	44.74	0.00	100.00
Black(%)	20.57	40.43	0.00	100.00
Hispanic(%)	4.44	20.61	0.00	100.00
Native American (%)	0.44	6.58	0.00	100.00
Other Race (%)	2.23	14.76	0.00	100.00
Seniors 65+ (%)	32.48	46.83	0.00	100.00
Children Under 18 (%)	18.79	39.07	0.00	100.00
LIHEAP(%)	5.96	23.68	0.00	100.00
Housing Structure				
Blower Door Pre (CFM50)	3990.78	2133.10	990.00	13662.00
Blower Door Post (CFM50)	2527.63	1136.47	746.00	7529.00
Blower Door Reduced (CFM50)	1389.31	1208.18	-1311.00	6527.00
Pct. Blower Door Reduced (%)	31.74	17.55	-114.10	83.31
Attic R-Value	15.82	9.31	0.50	40.00
Floor Area (sqft)	1455.71	592.13	570.00	3768.00
N Bedrooms	2.76	0.98	1.00	6.00
Has Multiple Stories (%)	32.17	46.72	0.00	100.00
Built Pre-1900 (%)	5.53	22.85	0.00	100.00
Built 1900-1929 (%)	22.39	41.69	0.00	100.00
Built 1930-1959 (%)	48.93	49.99	0.00	100.00
Built 1960-1989 (%)	20.38	40.29	0.00	100.00
Built 1990-Present (%)	2.77	16.42	0.00	100.00
Spending per Home or Upgrade (in US$)				
Total	5312.62	1541.85	54.15	11220.26
Air Conditioning	13.90	131.89	0.00	1827.00
Air Sealing	323.02	341.30	0.00	2020.52
Attic	943.44	742.42	0.00	3426.13
Baseload	190.12	243.72	0.00	982.79
Door	321.76	332.47	0.00	2020.34
Foundation	323.54	508.10	0.00	2988.15
Furnace	1348.72	1185.75	0.00	4664.21
General	54.60	288.05	0.00	5121.10
Health and Safety	558.34	365.70	0.00	1708.42
Wall Insulation	348.67	693.05	0.00	3467.00
Water Heater	128.88	238.06	0.00	1553.93
Window	640.59	852.15	0.00	4411.01
Number of Homes	9,881			
Energy Consumption and Weather				
Monthly Gas Usage (MMBtu)	6.33	6.69	0.00	44.10
Monthly Elec. Usage (MMBtu)	2.79	1.83	0.00	27.19
Min. Outdoor Temperature (C)	6.64	9.53	-19.83	24.37
Max. Outdoor Temperature (C)	17.61	10.36	-7.05	37.10
Precipitation (in)	3.08	1.91	0.00	18.36
Number of Observations	277,167			

Notes: This table presents averages, standard deviations, minimum, and maximum values for the main variables used in this study. All monetary values are in real terms, adjusted to 2017 dollars.

Table 2: WAP Average Treatment Effects on the Treated

Specification:	Engineering Projections	Machine Learning
Outcome: *Percent Energy Savings*	(1)	(2)
WAP Treatment	-29.03***	-14.83***
	(0.20)	(0.37)
Realization Rate		51.08%
Observations	22,394	142,327

Notes: This table presents the average projected savings for our sample and our estimates of the Average Treatment Effects on the Treated (ATT) for IHWAP using the machine learning approach as described in the main text. The WAP Treatment should be interpreted as percent energy savings attributable to the program. Engineering estimates use only two (one pre and one post) observations per home. The realization rate is calculated by dividing our estimates from column (2) by the engineering projections from column (1). All standard errors are clustered by home. Standard errors are bootstrapped (200 iterations) for machine learning. Significance at 1% is indicated by ***.

Table 3: Simulations of Measure-Specific Effects on the Wedge

	Baseline	Simulations					
		(1)	(2)	(3)	(4)	(5)	**(Total)**
Average Wedge (percentage points)	14.744	16.008	10.330	14.550	14.534	13.561	**8.744**
	(0.592)	(0.822)	(0.611)	(1.210)	(0.599)	(0.649)	(1.285)
Wedge Increase/Reduction Compared to Baseline		8.574%	-29.934%	-1.315%	-1.422%	-8.021%	**-40.692%**
		(4.209)	(1.909)	(7.760)	(0.465)	(1.548)	(8.388)
Observations	111,505	111,505	111,505	111,505	111,505	111,505	111,505
Zero marginal effect on the wedge from:							
Furnace Replacements		X					
Wall Insulation Spending			X				X
Furnace Spending				X			X
Attic Spending Above $2,400					X		X
Window Spending Above $1,400						X	X

Notes: This table presents results from simulations to assess how the average performance wedge would change if the marginal effect of spending on selected measures on the wedge were zero. We use the coefficient estimates from equation 2 for these simulations. The first column (Baseline) presents the average predicted wedge according to that model. The second column (simulation 1), for example, presents the average predicted wedge for a simulation assuming that the coefficients for furnace replacements (spending of over $1,800 on Furnaces) were zero. The third column (simulation 2) assumes that the marginal effect of any level of spending on Wall Insulation were zero, and so on. The "Wedge Increase/Reduction..." rows compare the simulated wedge from each column with the baseline average wedge of 14.744. Bootstrapped standard errors are presented in parentheses.

Table 4: Relationship Between Contemporaneous and Lagged Contractor Quality

Panel A: Coefficient Estimate on Lagged Contractor Quality		
	Specification	
Outcome: *Contractor Quality*	(1)	(2)
Lagged Quality	0.3352***	0.3449***
	(0.0633)	(0.0657)
Observations	88,249	88,249
R-squared	0.2100	0.3031
Controls:		
Housing Structure	Yes	Yes
Demographics	Yes	Yes
Weather	Yes	Yes
Interactions Between Measures	No	Yes

Panel B: Results from Oster Tests		
	Bias-Corrected Coeff.	δ
R-squared Max = .39	0.3384	5.0753
R-squared Max = .45	0.3399	3.7956
R-squared Max = .51	0.3418	3.0303

Notes: Panel A presents results from regression specification (3) from the main text, establishing a relationship between contemporaneous and lagged contractor quality. For brevity, we present only the coefficient associated with lagged quality. Bootstrapped standard errors are in parentheses. Significance at 1% is indicated by ***. Panel B presents results from Oster Tests to bound the potential effects of unobservable confounders on the estimated relationship (Oster, 2019). The tests compare coefficients and R-squares from specification (1) versus the fully saturated specification (2). The first column of Panel B presents results from bias-corrected coefficients, with varying levels of R-squared Max (i.e. how much of the relationship we expect to be able to explain) equal to 1.3, 1.5 and 1.7 times the R-squared from the saturated specification (2). We assume a coefficient of proportionality (δ) equal to one (i.e. observable and unobservable covariates are equally important in explaining the relationship). The bias-corrected coefficients are not significantly different from those presented in Panel A. The second column from Panel B is an alternative approach which produces bounds on the coefficient of proportionality necessary to drive our estimate to zero. Values of δ above 3 suggest that the unobservable confounders would need to be three times as important as the observables to nullify our estimates.

Table 5: Simulations for Workmanship Effects on the Wedge

	Baseline	"Best" Contractor Percentile			
		50th	75th	90th	95th
Panel A: Main Specification					
Avg. Pct. Point Wedge if **All Contractors** Become "Best"	15.357	15.406	12.871	10.452	**8.806**
	(0.621)	(0.638)	(0.734)	(0.977)	(1.205)
Wedge Reduction Compared to Baseline		0.315%	-16.190%	-31.939%	**-42.658%**
		(1.599)	(3.169)	(5.623)	(7.542)
Panel B: Interactions Between Wall Insulation Spending and Workmanship					
Avg. Pct. Point Wedge if **All Contractors** Become "Best"	15.357	15.097	12.566	10.137	**8.481**
	(0.621)	(0.661)	(0.762)	(1.004)	(1.225)
Wedge Reduction Compared to Baseline		-1.698%	-18.178%	-33.989%	**-44.777%**
		(1.924)	(3.462)	(5.854)	(7.703)
Observations	84,404	84,404	84,404	84,404	84,404

Notes: This table presents results from simulations for which we replace all contractors' marginal effects on the wedge with those for contractors identified as high-performers. We define high-performance contractors as those who are at the 95th, 90th, 75th, or 50th percentiles in terms marginal effects on the wedge. The simulations consist of applying those effects to all homes as if they had been served by the best contractors. We then calculate the resulting simulated average performance wedge. Panel A assumes there are no interactions between estimated contractor quality and spending on measures. For Panel B, we consider the interactive effects. Specifically, we replace the coefficient of the interactive effect between low (bottom 20%) performers and spending on Wall Insulation with zero, which leads to only a slight reduction of the wedge compared to Panel A. Bootstrapped standard errors are in parentheses. The "Wedge Reduction..." rows compare the resulting simulated wedge from each column with the "baseline" estimated wedge of 15.357.

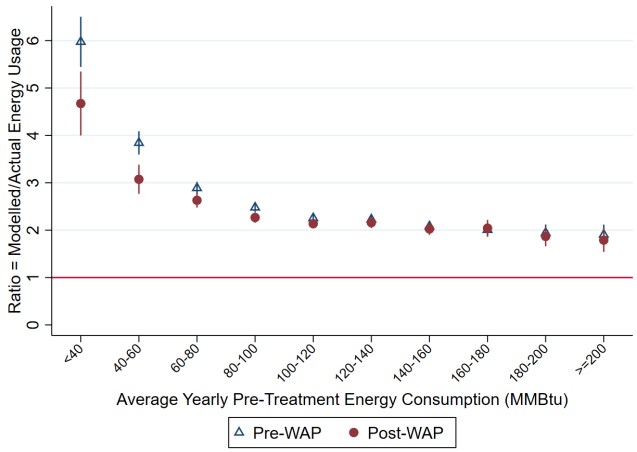
Table 6: Simulations for Impact of the Rebound Effect on the Wedge

	Baseline	Varying the balance point		
Balance Point (^{o}F)	61.8	61.6	**61.4**	61.2
Removed Rebound Effect (^{o}F)	0	0.2	**0.4**	0.6
Average Percentage Point Savings	-11.391	-11.874	**-12.352**	-12.824
	(0.543)	(0.542)	(0.540)	(0.539)
Savings Increase Compared to Baseline		4.246%	**8.442%**	12.585%
		(0.198)	(0.393)	(0.585)
Average Percentage Point Wedge	15.098	14.619	**14.140**	13.673
	(0.583)	(0.581)	(0.580)	(0.579)
Wedge Reduction Compared to Baseline		-3.177%	**-6.347%**	-9.443%
		(0.090)	(0.178)	(0.261)
Observations	128,670	128,655	128,644	128,631

Notes: This table presents results from simulations to assess how the average performance wedge changes by "eliminating" the rebound effect. We estimate post-treatment energy usage (according to equation 5) with balance points adjusted to reflect plausible changes in indoor air temperature due to the rebound effect (0.2, 0.4, and 0.6^{o}F). Lower indoor air temperature settings (lower rebound) directly map to lower balance points. Lower balance points indicate that the heating systems turn on at lower outdoor air temperatures, thus reducing energy consumption and the wedge. The "Wedge Reduction Compared to Baseline" compares the resulting simulated wedge from each column with the "Baseline" estimated wedge of 15.098. Bootstrapped standard errors are in parentheses.

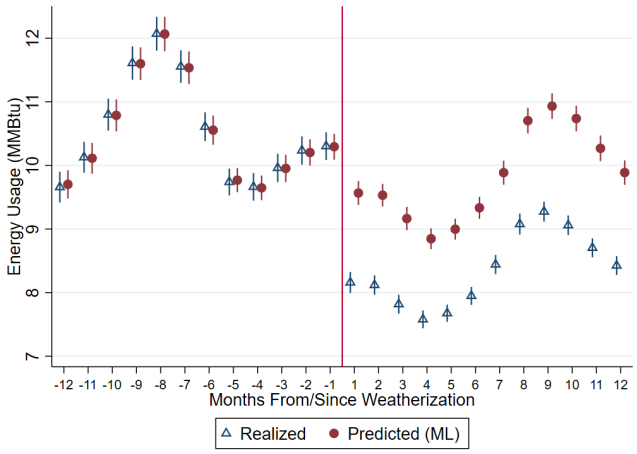
Figures

Figure 1: Ratio Between (WeatherWorks) Modeled and Actual Energy Usage



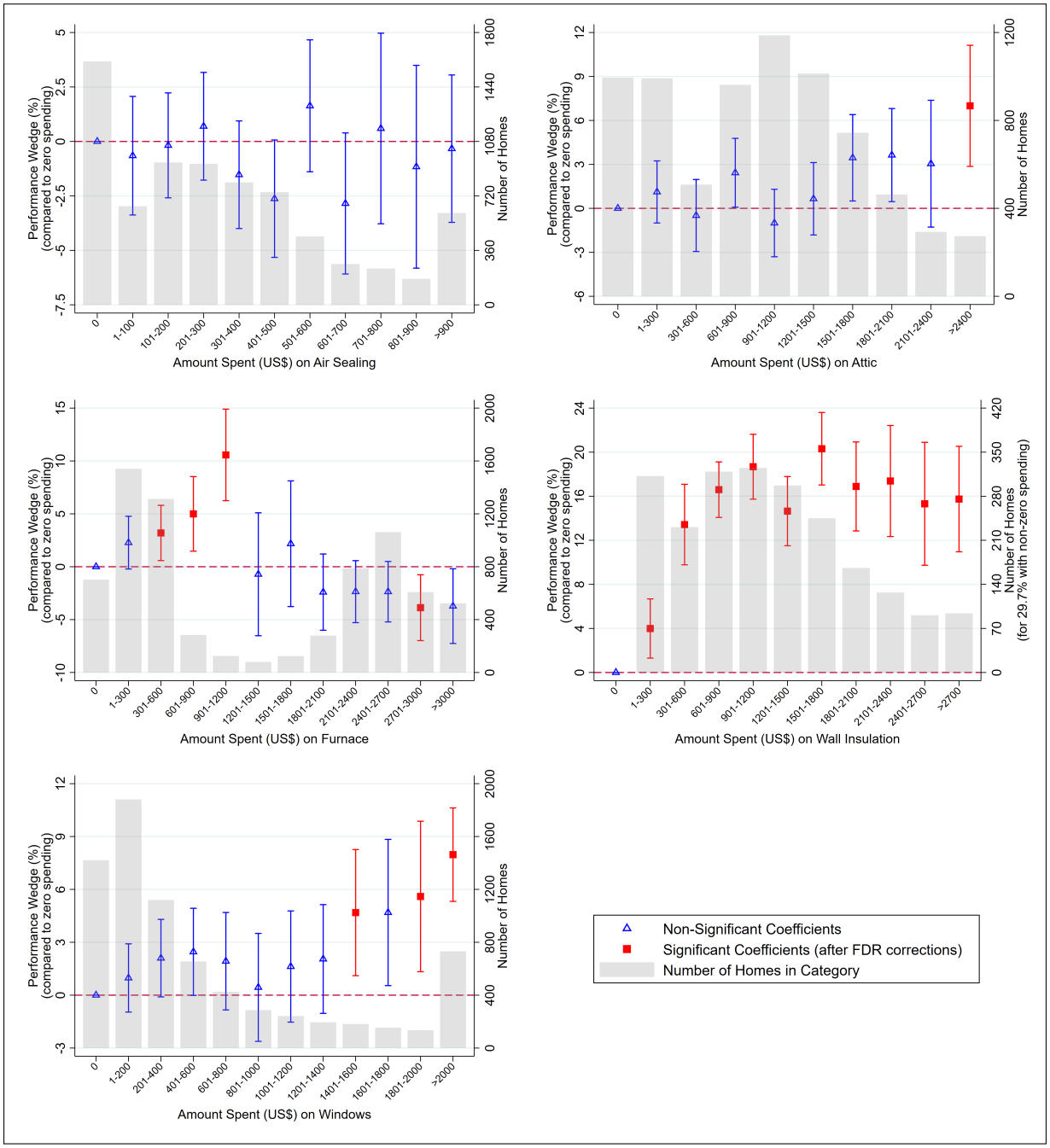
Notes: This figure presents average ratios between projected and realized energy usage of WAP-treated homes. The whiskers represent 95% confidence intervals. Averages are calculated for bins of homes' annual pre-treatment usage. For comparability with the yearly energy usage values provided by the engineering model, we limit the sample used in this figure to approximately 2,800 homes with at least a full year (Jan-Dec) of pre- and post-treatment billing data. As indicated by ratios above 1, we find that engineering models consistently overestimate homes' energy consumption. That is accentuated for smaller homes (with lower yearly consumption), and it holds both before and after treatment.

Figure 2: Realized versus Predicted Energy Usage by Timing of Treatment



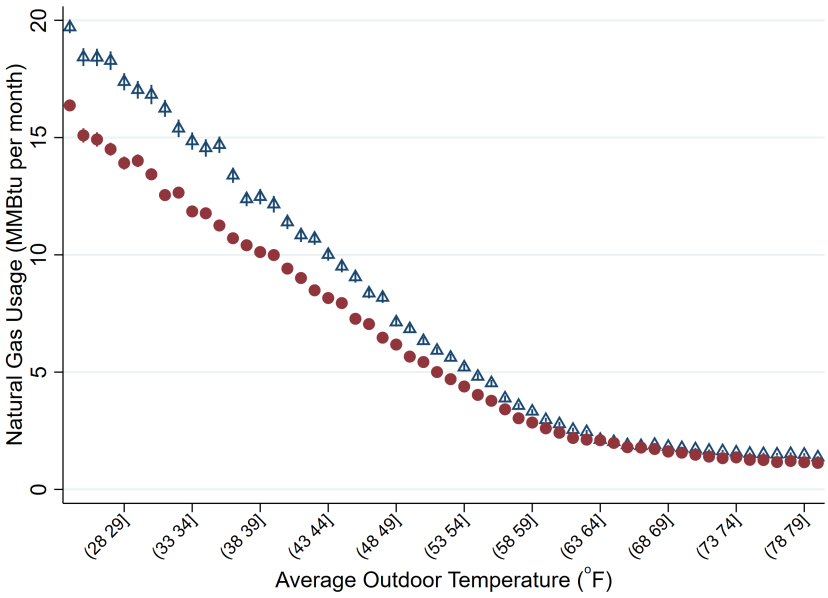
Notes: This is an event study graph comparing realized (ex post) versus predicted energy usage for WAP treated homes. The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. Predictions are based on a flexible machine learning model, as described in section 3.1. The model is trained with pre-treatment data only. We present cross-validated (out-of-sample) predictions for months before weatherization. The predictions after treatment represent counterfactuals (energy usage in case the homes had not been treated). The difference between the curves post-treatment represents the energy savings attributed to WAP.

Figure 3: Heterogeneity in the Performance Wedge by Spending on Retrofits

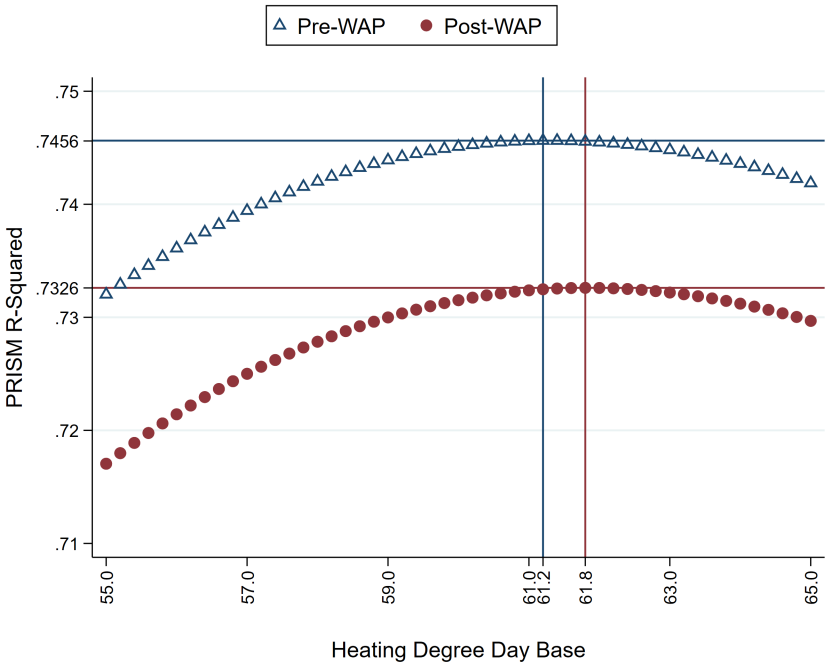


Notes: This figure presents estimates of how the performance wedge is affected by additional spending on the five major program measures. Coefficients are interpreted as percentage point increase/reduction in the wedge, relative to the omitted category (zero spending). The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. P-values have been corrected with the false discovery rate (FDR) procedure from Benjamini and Hochberg (1995), where red indicates significance after these corrections. We assume an overall uncorrected critical p-value of 0.05 for each group. Uncorrected p-values within groups are assumed to be nonnegatively correlated. Results are robust across FDR or family-wise error rate (FWER) correction procedures. The light grey bars represent the number of homes with spending in a given category, denoted on the right-hand vertical axis.

Figure 4: Identifying Change in Balance Points



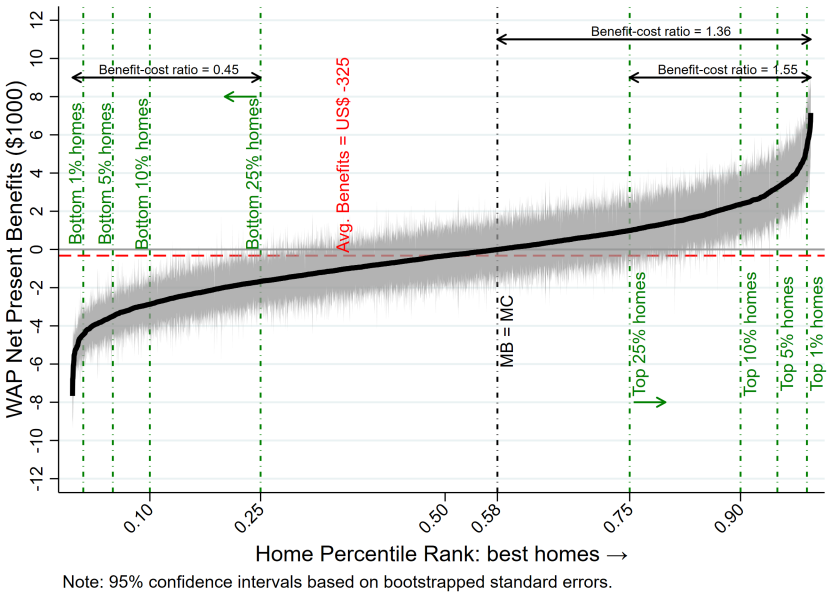
Panel A: Average Energy Usage by Temperature



Panel B: PRISM Optimal HDD Bases

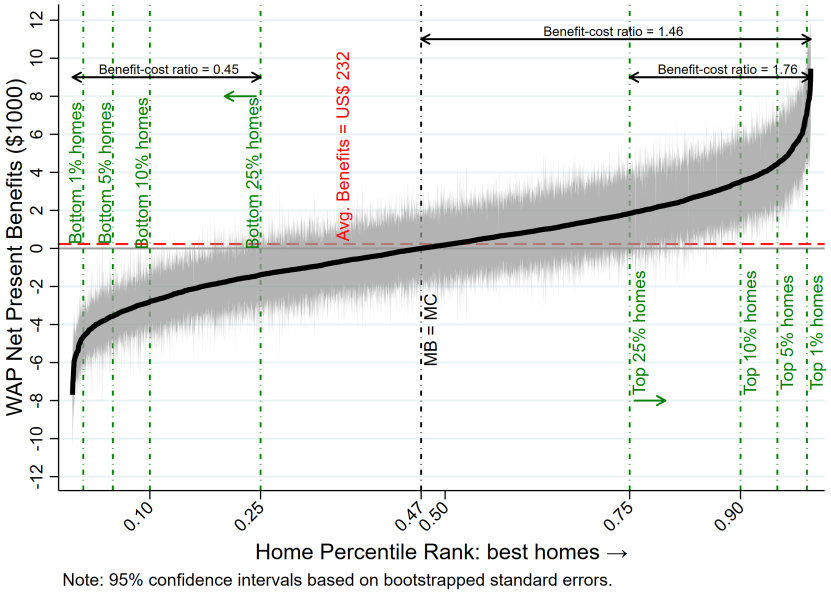
Notes: Panel A plots average natural gas consumption by outdoor temperature bins for homes served by WAP. We use the full sample for this analysis, plotting averages before and after treatment. Standard errors are clustered by home. Panel B plots results from PRISM analyses using the full sample of WAP homes. We iterate through many temperatures to identify the optimal HDD balance points for an average home, both before and after treatment. Balance points with highest R-squared are considered optimal.

Figure 5: Ranking of Homes by Net Present Benefits



Note: 95% confidence intervals based on bootstrapped standard errors.

a: Social Marginal Costs



Note: 95% confidence intervals based on bootstrapped standard errors.

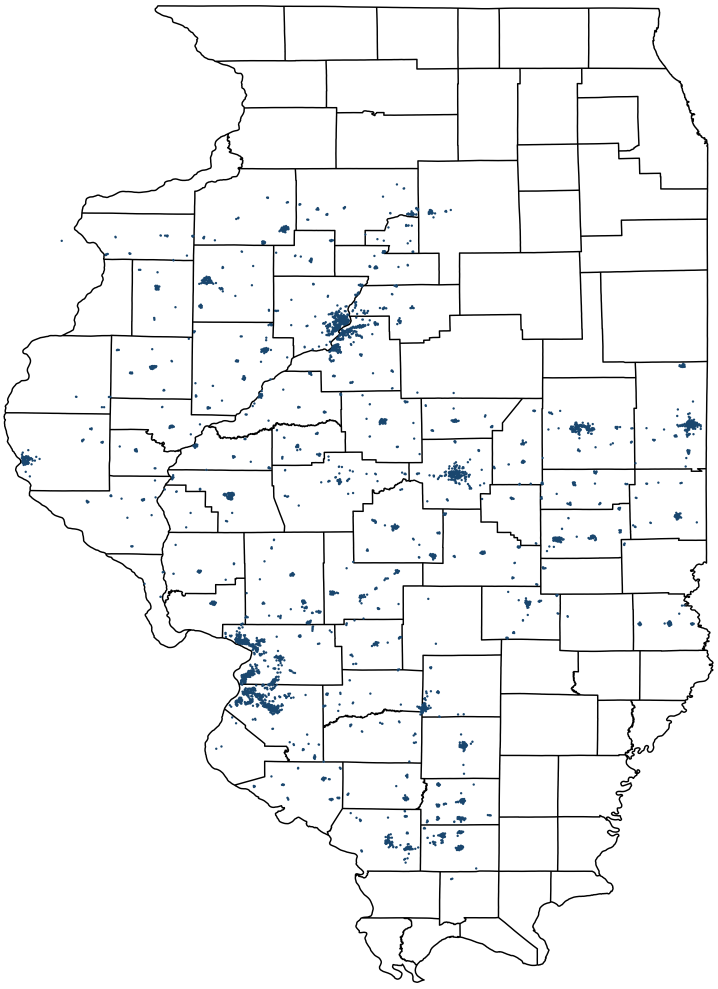
b: Retail Energy Prices

Notes: This figure ranks homes from lowest to highest WAP net benefits. The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. We a sample of homes with at least one full year of post-treatment data. The estimates of benefits assume a 3% discount rate, and upgrade lifespans of, on average, 30 years. For the bottom panel, energy savings are monetized based on retail energy prices. For the top panel, we adjust for social marginal costs. It is possible to note significant heterogeneity in program benefits, ranging from -$7,500 to +$7,000 according to the top panel. Even though average net benefits are close to zero, we identify many homes with significantly positive benefits.

Appendix – For Online Publication

A Additional Details on the Study Sample

Figure A.1: Geographical Distribution of WAP-Treated Homes in Sample



Notes: This Figure maps homes with available energy consumption data, which constitutes the sample used for the analyses in this paper. Each blue dot represents a home in our sample. The Northern and Southeastern parts of the state are not served by the utility that provided data for this project.

B Machine Learning Model Tuning and Diagnostics

We use the following sample restrictions to predict pre-treatment energy consumption: single-family homes; heating fuel is natural gas; for which both electricity and natural gas billing data is available. The model is trained only with data that would be available prior to weatherization: pre-treatment billing data, energy audit information, household demographics, and weather variation. Specifically, we include the following variables: energy usage in MMBtu (outcome), heating degree days (base 60, and 65), cooling degree days (base 75), min. outdoor temperature, max. outdoor temperature, precipitation, floor area (square feet), family size, number of windows, number of stories, number of bedrooms, vintage, county indicator, building shielding (measure of shielding provided by structures surrounding home), pre-treatment blower door test (CFM50), main heating system type, main heating system capacity (Btu), attic R-value, household income, indicators for householder's race, presence of disable occupant, presence of children, presence of elderly, home priority rank, audit date (month, year, and day), program year of audit, month of year, year of sample, and number of days in billing cycle. Our outcome (energy usage) varies by home and by month of sample (billing period). Weather also varies by month of sample, while information collected during WAP audit/applicaiton varies only by home.

We use the machine learning algorithm XGBoost, which is a computationally efficient implementation of gradient boosted trees, developed by (Chen and Guestrin, 2016). The concept of boosted trees involves iteratively combining weak predictive trees to form an ensemble. More weights are given to the trees with better predictive accuracy. By default, the algorithm uses mean squared errors (MSE) as a measure of accuracy. Each tree is constructed with a fraction of the provided sample and a different set of the variables described above. It is important to note that regression trees automatically consider variable interactions and non-linear functional forms (i.e. binning). As the tree "depth" increases, interactions become more complex. With more tree "branches," the model allows for more flexibility in how each variable is included.

To increase predictive accuracy of machine learning models, it is common practice

to "tune" the (hyper)parameters that control factors such as maximum tree depth. The following section describes the configurations that we considered for our model.

B.1 Hyperparameter Tuning

We perform cross-validation and hyperparameter tuning to identify which machine learning algorithm exhibits best out-of-sample prediction accuracy. We implement 5-fold cross-validation via the "SuperLearner" package in R Polley et al. (2018). We consider the following 5 types of predictions models: ridge regression; elastic net; lasso; random forest; XGBoost. Several hyperparmeter configurations are tested for each of the types of models.

Results from Table B.1 suggest that models with lower learning rate (shrinkage = 0.05) are generally more accurate in this setting, as measured by the cross-validated (out of sample) RMSE. Increasing the number of trees does not significantly affect performance. We therefore select a parsimonious model (number of trees = 1000). Our preferred specification is an ensemble of model IDs 1 and 3.

Table B.1: Hyperparameter Tuning - XGBoost

Model ID	N Trees	Max Tree Depth	Shrinkage	Min Obs per Node	In Sample RMSE	Cross-Validated RMSE
1	1000	20	0.05	30	0.745	2.614
2	2000	20	0.05	30	0.379	2.598
3	1000	30	0.05	30	0.459	2.641
4	2000	30	0.05	30	0.111	2.635
5	1000	20	0.50	30	0.001	3.002
6	2000	20	0.50	30	0.001	3.002
7	1000	30	0.50	30	0.005	3.061
8	2000	30	0.50	30	0.002	3.061

Table B.2: Hyperparameter Tuning - Random Forests

Model ID	N Trees	Max Nodes	Min Obs per Node	In Sample RMSE	Cross-Validated RMSE
1	1000	500	30	3.455	3.621
2	2000	500	30	3.455	3.621
3	1000	1000	30	3.253	3.538
4	2000	1000	30	3.255	3.538

Table B.3: Hyperparameter Tuning - Ridge/Elastic Net/Lasso

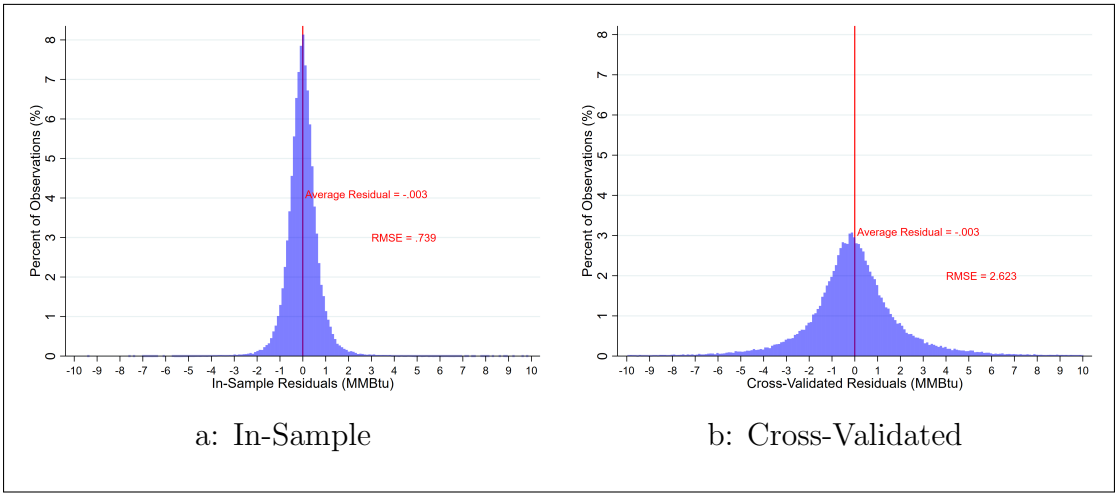
Model Type	Alpha	Max Variables	In Sample RMSE	Cross-Validated RMSE
Ridge	0.00	50	no convergence	no convergence
Ridge	0.00	75	no convergence	no convergence
Ridge	0.00	100	7.666	7.666
Ridge	0.00	150	7.666	7.666
Ridge	0.00	200	3.711	3.716
Elastic Net	0.25	50	3.917	3.911
Elastic Net	0.25	75	3.753	3.759
Elastic Net	0.25	100	3.707	3.712
Elastic Net	0.25	150	3.665	3.671
Elastic Net	0.25	200	3.664	3.670
Elastic Net	0.50	50	3.830	3.836
Elastic Net	0.50	75	3.739	3.742
Elastic Net	0.50	100	3.702	3.706
Elastic Net	0.50	150	3.663	3.672
Elastic Net	0.50	200	3.663	3.672
Elastic Net	0.75	50	3.807	3.804
Elastic Net	0.75	75	3.720	3.728
Elastic Net	0.75	100	3.696	3.699
Elastic Net	0.75	150	3.670	3.675
Elastic Net	0.75	200	3.670	3.676
Lasso	1.00	50	3.791	3.779
Lasso	1.00	75	3.715	3.722
Lasso	1.00	100	3.691	3.696
Lasso	1.00	150	3.673	3.677
Lasso	1.00	200	3.673	3.677

Notes: "no convergence" indicates that the algorithm did not arrive at a sufficiently precise lambda (or shrinkage) parameter.

B.2 Prediction Errors

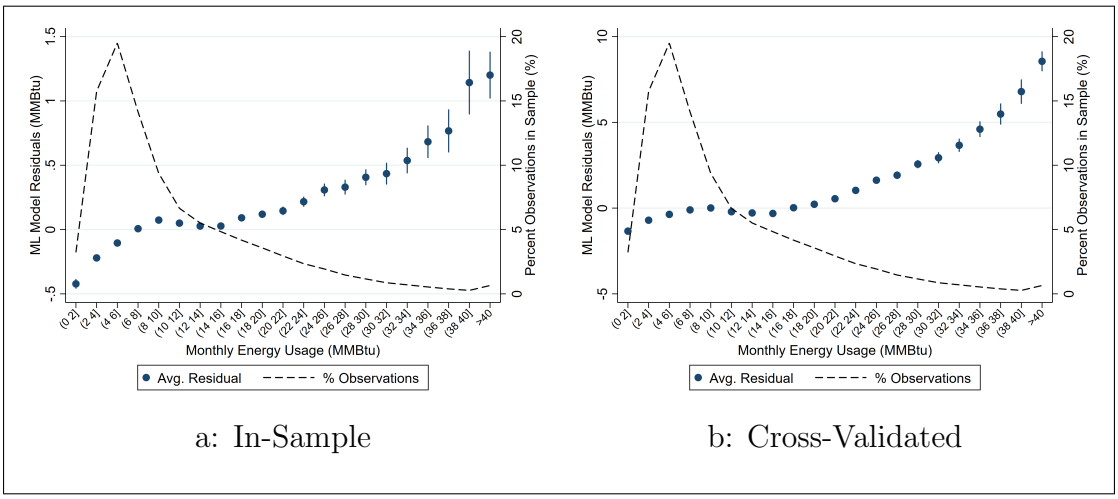
Figure B.1 presents the distributions of in-sample and cross-validated prediction errors (residuals) for the machine learning model. Both types of errors are approximately centered around zero, although cross-validated errors exhibit significantly fatter tails. In Figure B.2, we disaggregate the errors by bins of monthly energy consumption on the horizontal axis. The dashed lines represent the percent of months (on the right vertical axis) with a given level of observed energy consumption. The (5-fold) cross-validated errors serve as a measure for out-of-sample model performance. As expected, those are larger than in-sample errors. Nevertheless, significant errors occur only at the tails of the distribution (for months when energy usage was abnormally high or abnormally low). We can note slight overestimation of energy usage at the low end, and slight underestimation at the high end.

Figure B.1: Distribution of Pre-Treatment Residuals



a: In-Sample b: Cross-Validated

Figure B.2: Pre-Treatment Residuals by Actual Energy Consumption



a: In-Sample b: Cross-Validated

B.3 Correlation Between Prediction Errors and Covariates

In this section, we report the correlation between pre-treatment cross-validated errors and observable covariates. These graphs provide evidence on the relationship between errors in the machine learning model and observable characteristics of homes in our sample. We expect errors to be zero, on average, for fine scale bins of our controls. We note that estimates for some bins are very noisy and that some are statistically different from zero in one direction or the other. However, they are for sparse regions of the sample and are small in magnitude. The same is true for the graphs of prediction errors by program spending. These graphs suggest that our ML model errors are unlikely to drive the results reported in the paper.

Figure B.3: Cross-Validated Prediction Errors by Observable Covariates

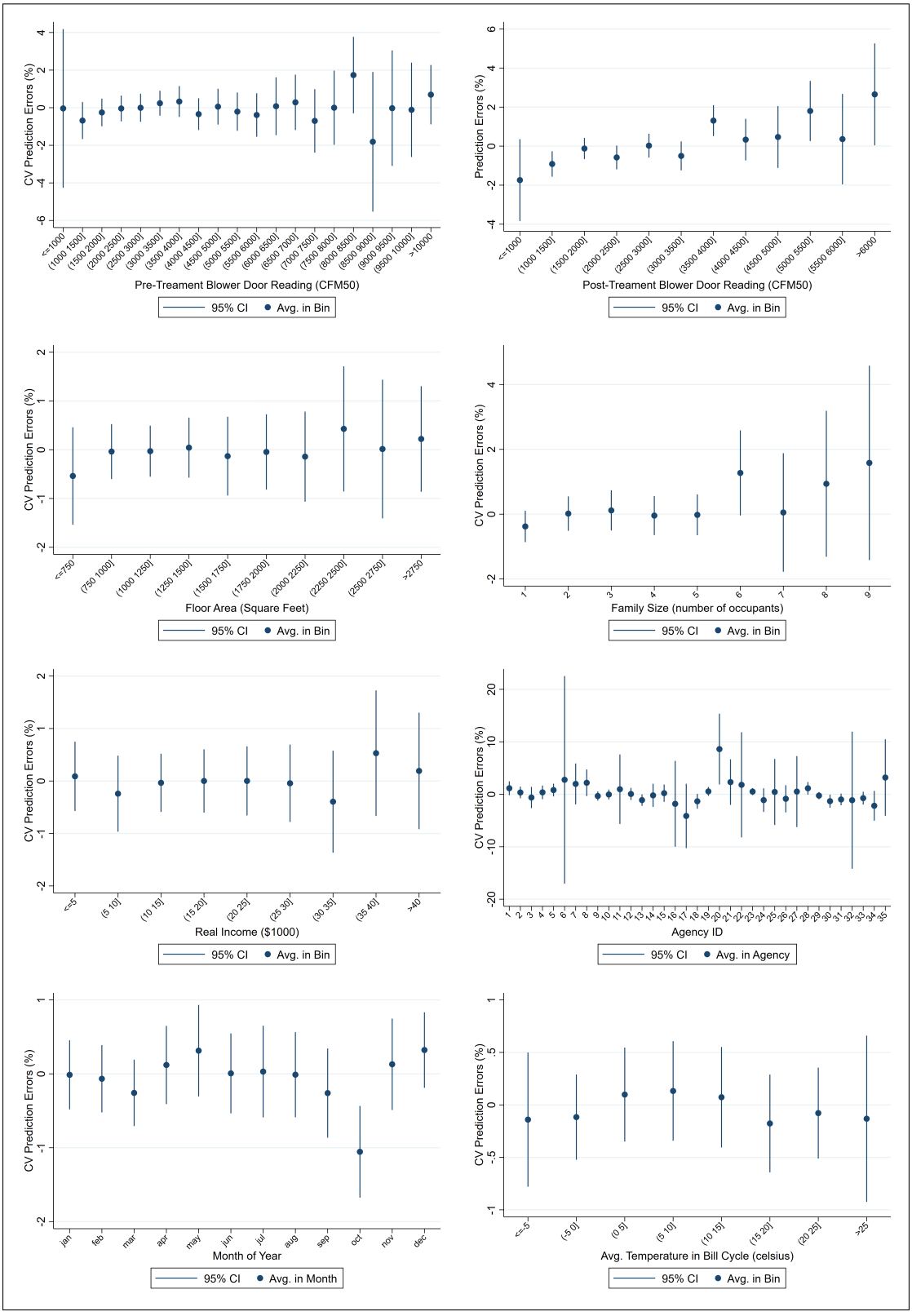


Figure B.4: Cross-Validated Prediction Errors by Program Spending

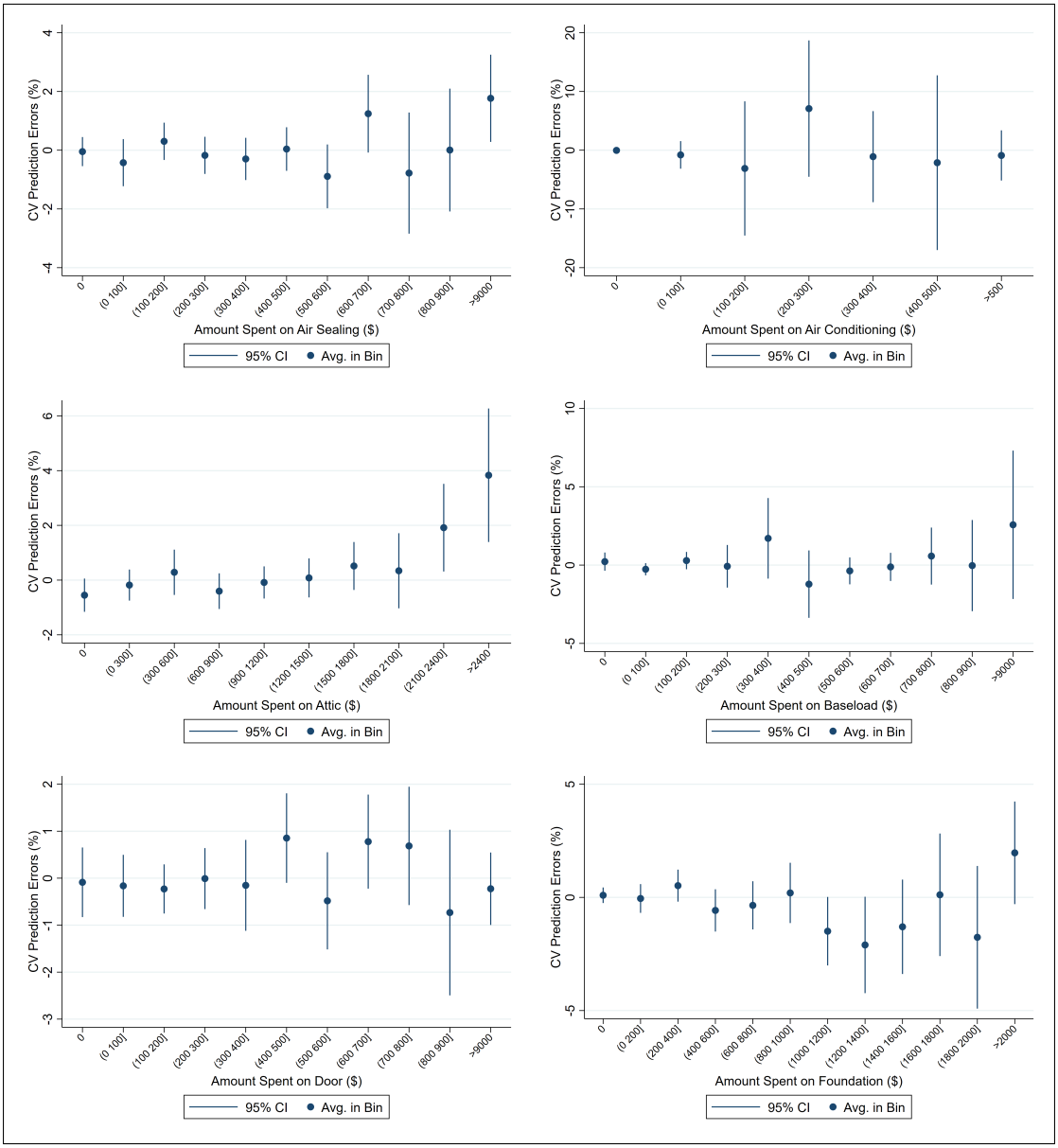
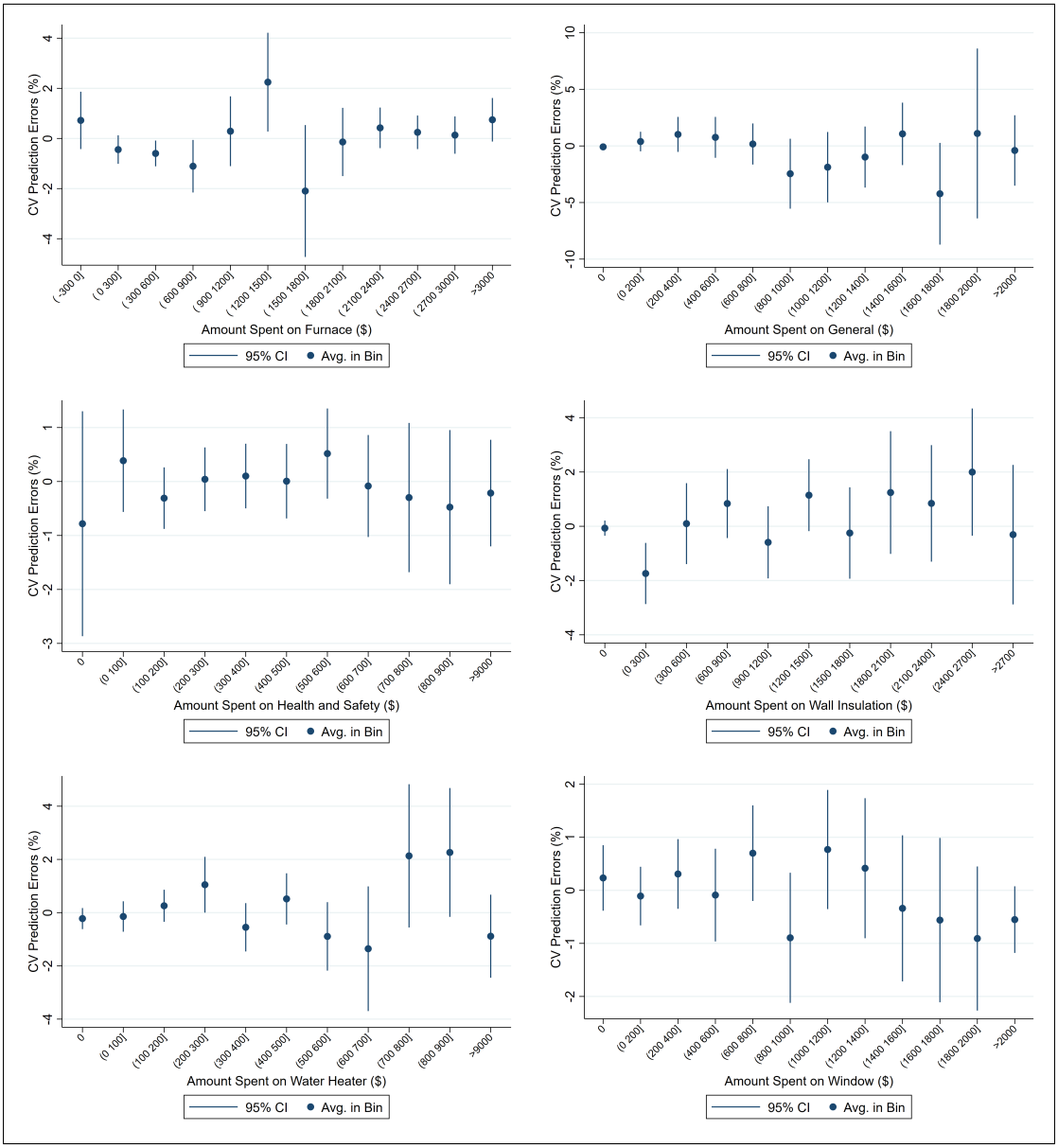


Figure B.4 (Continued): Cross-Validated Prediction Errors by Program Spending



C Obtaining Monthly Projections of Energy Savings from Weatherworks

This section describes the method that we use to obtain monthly projections from Weatherworks model that are comparable to our monthly observations from billing data. The Weatherworks engineering model provides 2 estimates of a home's energy usage: (1) pre-weatherization and (2) post-weatherization. These estimates are designed to represent a full year of energy demanded prior to weatherization and a full year post. We re-scale these estimates to obtain monthly observations by dividing by 12.[34] The average projected savings from the engineering model can then be obtained with a home fixed effects model:

$$ln(Y_{id}) = \beta^p \mathbf{1}[WAP]_{id} + \alpha_i + \varepsilon_{id} \tag{C.1}$$

where Y_{id} are engineering model calculations of monthly energy demand for home i in treatment status d. β^p measures the average projected energy savings for the sample of homes used in the regression. We restrict the sample to the same homes for which utility data was available. Since the regression specifies a log-linear form, we correct the estimated coefficients of interest so that they can be interpreted as percentage projected energy savings: $exp(\beta_p) - 1$.

[34] Rescaling the engineering projections does not change results that are in percentage terms. However, that transformation is necessary for analysis in levels (MMBtu) because our utility data is monthly.

D Comparing Estimates from Machine Learning and Two-Way Fixed Effects Models

We compare our machine learning estimates with variations two-way fixed effects regressions. Specifically, we consider the following specification:

$$ln(Y_{it}) = \beta^{TWFE}\mathbf{1}[WAP]_{it} + \alpha_i + \alpha_t + \varepsilon_{it} \tag{D.1}$$

where $ln(Y_{it})$ is the natural log of energy consumption from home i in billing cycle t; $\mathbf{1}[WAP]_{it}$ is a treatment indicator equal to one for time periods after a given home has been treated, zero otherwise; α_i are home fixed effects; and α_t are time fixed effects. We also consider variations of that specification with interactions of home by calendar month fixed effects, as well as month of sample by county fixed effects.

These results are reported in Table D.1. We find that the machine learning estimates are slightly higher than those from two-way fixed effects specifications.

Table D.1: WAP Average Treatment Effects on the Treated

Specification:	Engineering Projections	Machine Learning	Fixed Effects Models		
Outcome = log(Energy)	(1)	(2)	(3)	(4)	(5)
WAP Treatment	-0.2903***	-0.1483***	-0.1321***	-0.1295***	-0.1280***
	(0.0020)	(0.0037)	(0.0039)	(0.0038)	(0.0039)
Realization Rate		.5108			.441
Observations	22,394	142,327	277,182	239,135	238,167
Controls:					
Home FE	Yes	NA	Yes	No	No
Month of Sample FE	No	NA	Yes	Yes	No
Home by Calendar Month FE	No	NA	No	Yes	Yes
Month of Sample by County FE	No	NA	No	No	Yes
Heating and Cooling Degree Days	No	NA	Yes	Yes	Yes

Notes: This table presents Average Treatment Effects on the Treated (ATT) estimates for WAP. The coefficients on WAP Treatment should be interpreted as percent energy savings attributable to the program. No controls are used for the machine learning ATT, which is identified from the difference between post-treatment usage and predicted counterfactuals. Machine learning estimates use post-treatment monthly observations only (although predictive models are trained with pre-treatment data). Engineering estimates use only two (one pre and one post) observations per home. Realization rates are calculated by dividing estimates from columns (2) or (5) by the engineering projections from column (1). All standard errors are clustered by home. Standard errors are bootstrapped (200 iterations) for machine learning. Significance at 1% is indicated by ***.

Table D.2 reports estimates of WAP average treatment effects in levels rather than logs. We note that realization rates are significantly smaller in these specifications. This can be attributed to engineering model overestimation of energy usage both before and

after treatment, as shown in Figure 1. As discussed in Section 2.5, overestimation does not necessarily imply bias in percent projected energy savings. The general agreement among WAP stakeholders is that the engineering models aim to be accurate in projecting percentage energy reduction.

Table D.2: WAP Average Treatment Effects on the Treated - levels

Specification:	Engineering Projections	Machine Learning	Standard Econometrics		
	(1)	(2)	(3)	(4)	(5)
WAP Treatment	-5.1656***	-1.4529***	-1.5988***	-1.5380***	-1.5050***
	(0.0495)	(0.0413)	(0.0422)	(0.0346)	(0.0351)
Realization Rate		.2813			.2913
Observations	22,394	142,327	277,182	239,135	238,167
Controls:					
Home FE	Yes	NA	Yes	No	No
Month of Sample FE	No	NA	Yes	No	No
Home by Calendar Month FE	No	NA	No	Yes	No
Month of Sample by County FE	No	NA	No	No	Yes
Heating and Cooling Degree Days	No	NA	Yes	Yes	Yes

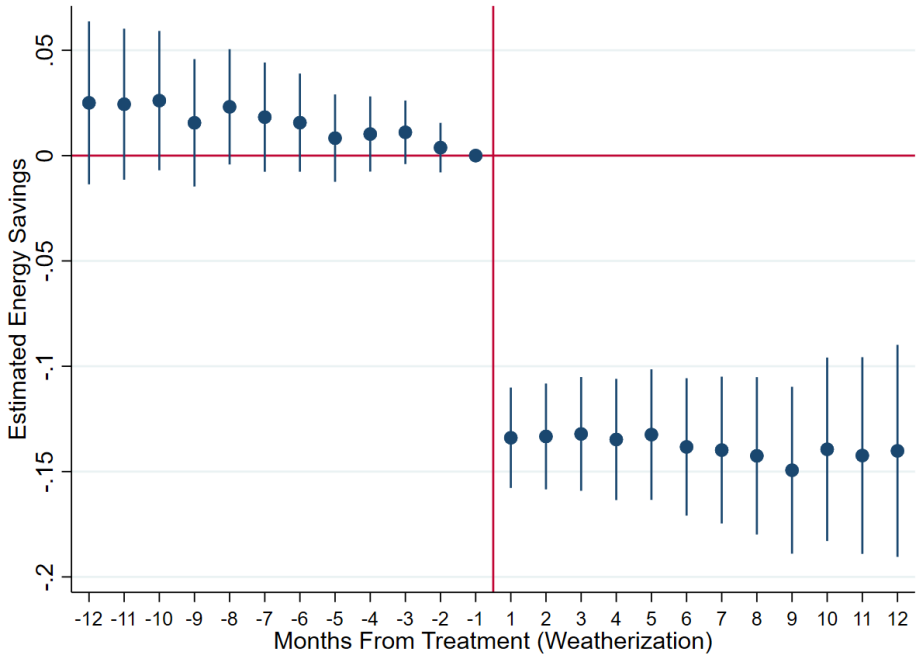
Note: This table presents ATT estimates of the effect of WAP on energy usage. The coefficients on WAP treatment should be interpreted as MMBtu energy savings attributable to WAP. No controls are used for the machine learning ATT, which is identified from predicted counterfactuals. Realization rates are calculated by dividing realized savings estimates with the engineering projected savings. All standard errors are clustered by home. For the machine learning estimates, standard errors are bootstrapped (200 iterations). Significance at 1% is indicated by ***.

We also assess the parallel trends assumption for the fixed effects models. We add lags and leads of timing of treatment to regression specification (D.1). With that, we estimate effects for 12 months before and 12 months after WAP upgrades, conditional on home and month of sample fixed effects. Recall that for our analyses we exclude "work in progress" months for which we believe upgrades are still being performed (months between audit and final inspection date). In our setting, the month immediately before a home's audit date is the omitted comparison group.[35] We normalize timing of treatment based on number of months before/elapsed since the construction phase.

Figure D.1 presents results from an event study regression using that approach. We cannot reject that the coefficients prior to treatment are equal to zero, such that parallel trends are likely to hold. For the months immediately after final inspection, there is a strong reduction in energy usage, with points estimates close to 13%. Effects do not seem to dissipate even a full year after treatment.

[35]Normally for event studies it is possible to identify a clear cutoff point after which treatment occurs. Given that WAP treatment may occur over many days, there is no clear cutoff in this context. Therefore we exclude monthly observations that are constituted of a mix of untreated and treated days.

Figure D.1: Event Study Results - Fixed Effect Models



Finally, Table D.3 presents results from specifications to recover the effects of WAP on natural gas and electricity usage separately. It can be noted that, given the focus of WAP, natural gas savings are significantly higher than electricity savings. Looking at the last three columns we find that, for the average home, approximately 83% of program savings can be attributed to reduced natural gas usage, while the remainder can be attributed to reduced electricity usage.

Table D.3: WAP Average Treatment Effects - Natural Gas Versus Electricity Savings

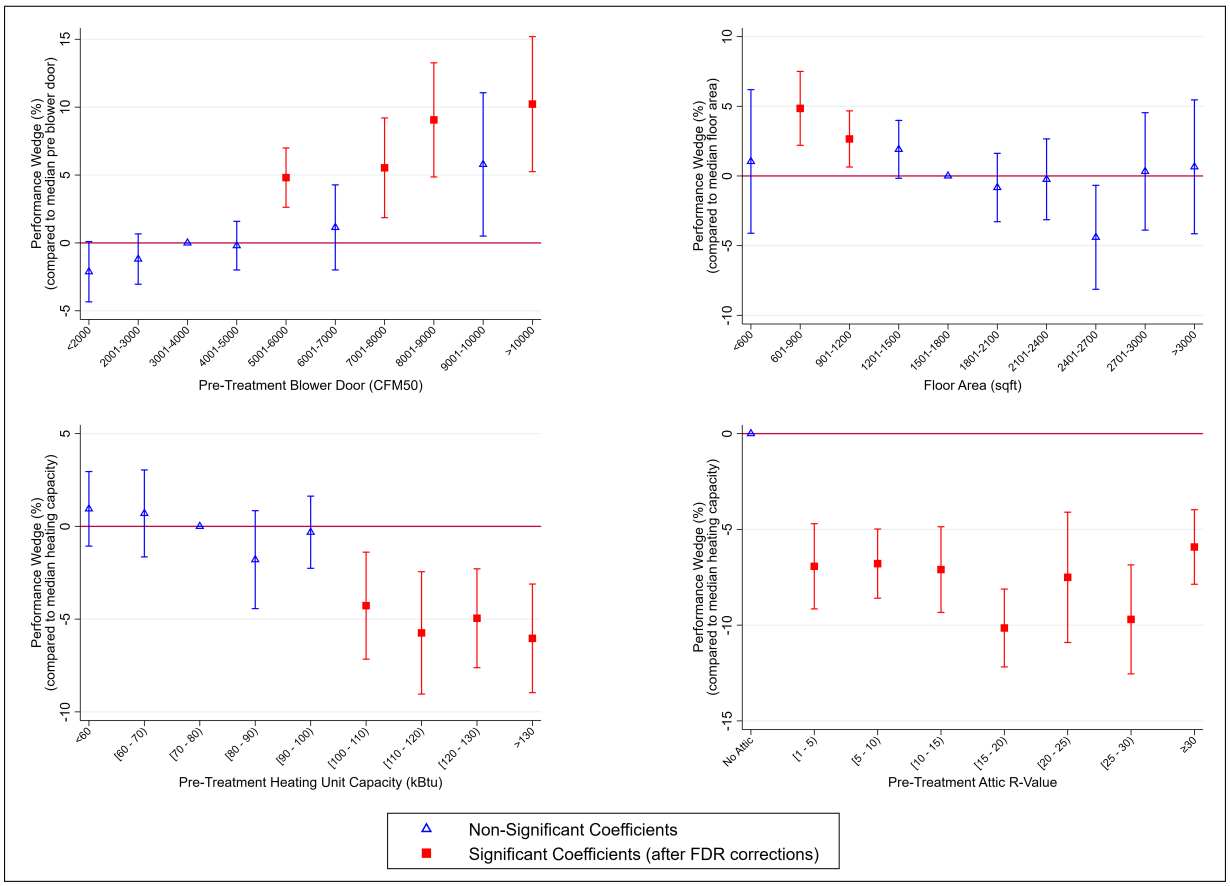
	Logs			Levels (MMBtu)		
Energy Outcome:	Total	Gas	Electricity	Total	Gas	Electricity
WAP Treatment	-0.1280***	-0.1472***	-0.0767***	-1.5050***	-1.2374***	-0.2676***
	(0.0039)	(0.0059)	(0.0052)	(0.0351)	(0.0352)	(0.0156)
Observations	238,167	238,167	238,059	238,167	238,167	238,167

Note: This table presents ATT estimates of the effect of WAP on total energy, natural gas, and electricity usage. These are results from fixed effects models that include home by calendar month FE, month of sample by county FE, plus weather controls (degree days). Standard errors are clustered by home. Significance at 1% is indicated by ***.

E Wedge Heterogeneity for Other Covariates

In the main text we interpret results for the five major retrofits related to energy savings and the performance wedge. The following suite of graphs presents heterogeneity results for other household or housing structure variables that were not discussed. These were all obtained from a same regression (2) that flexibly decomposes the performance wedge across many dimensions.

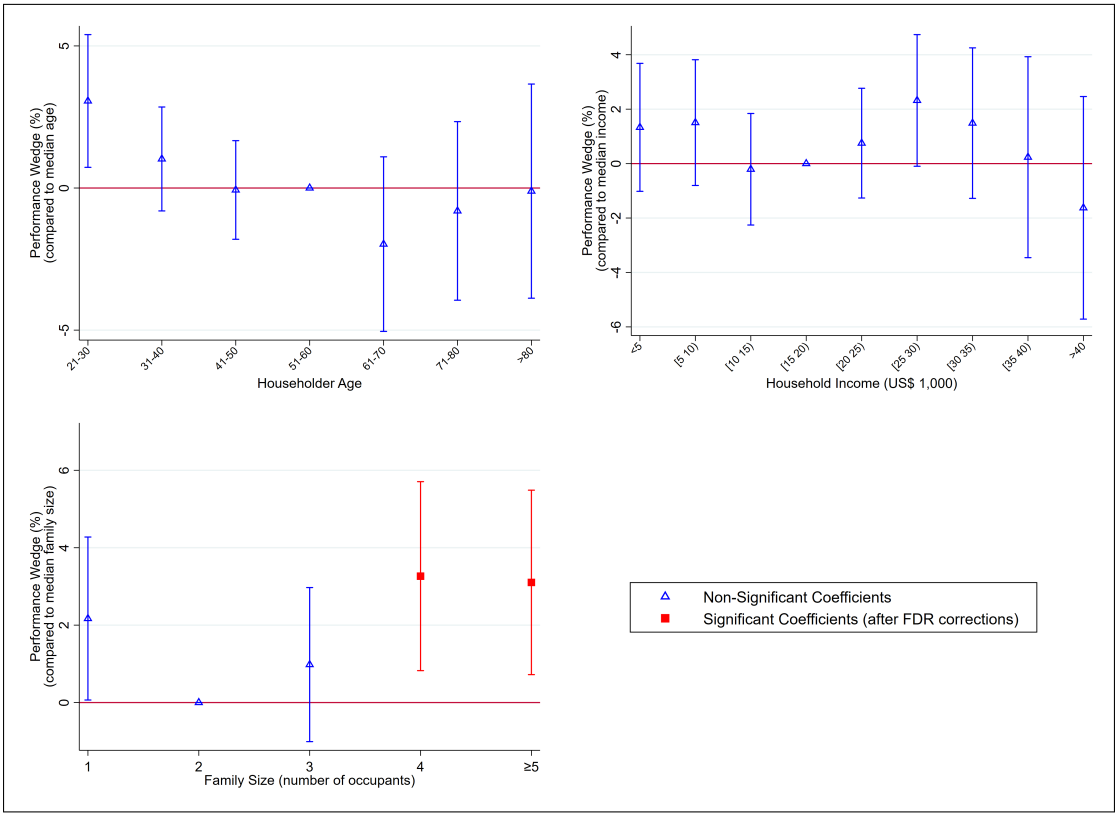
Figure E.1: Performance Wedge Heterogeneity by Other Covariates



Notes: This figure presents estimates of heterogeneity in the performance wedge by household or housing structure variables. Coefficients are interpreted as percentage point increase/reduction in the wedge, relative to the omitted category. The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. P-values have been corrected with the false discovery rate (FDR) procedure from Benjamini and Hochberg (1995), where red indicates significance after these corrections. We assume an overall uncorrected critical p-value of 0.05 for each group. Uncorrected p-values within groups are assumed to be nonnegatively correlated. Results are robust across FDR or family-wise error rate (FWER) correction procedures.

(Continues on next page)

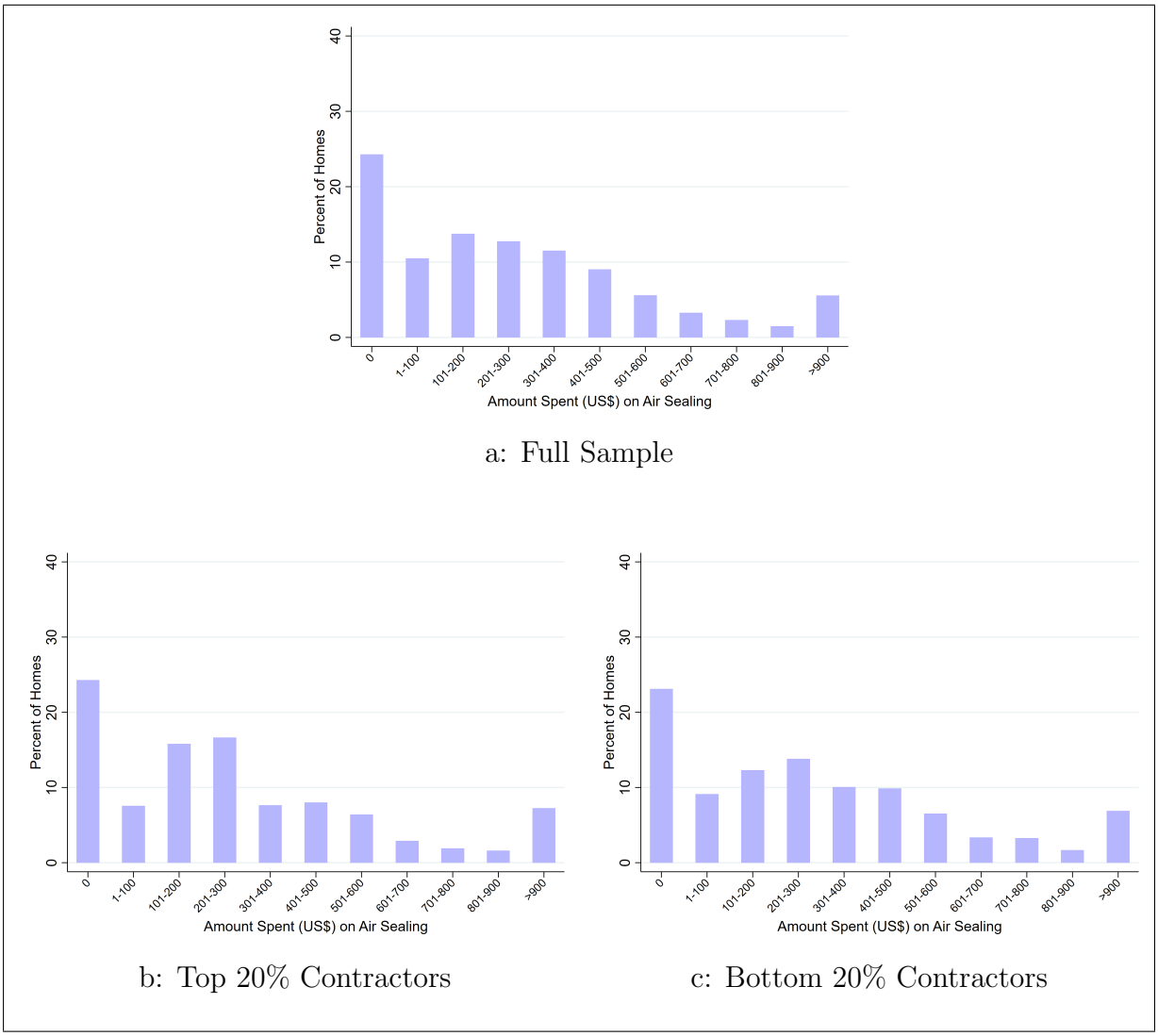
Figure E.1: Performance Wedge Heterogeneity by Other Covariates (continued)



Notes: This figure presents estimates of heterogeneity in the performance wedge by household or housing structure variables. Coefficients are interpreted as percentage point increase/reduction in the wedge, relative to the omitted category. The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. P-values have been corrected with the false discovery rate (FDR) procedure from Benjamini and Hochberg (1995), where red indicates significance after these corrections. We assume an overall uncorrected critical p-value of 0.05 for each group. Uncorrected p-values within groups are assumed to be nonnegatively correlated. Results are robust across FDR or family-wise error rate (FWER) correction procedures.

F Robustness of Contractor Quality Estimates and Interactions

Figure F.1: Histograms for Amount Spent on Air Sealing

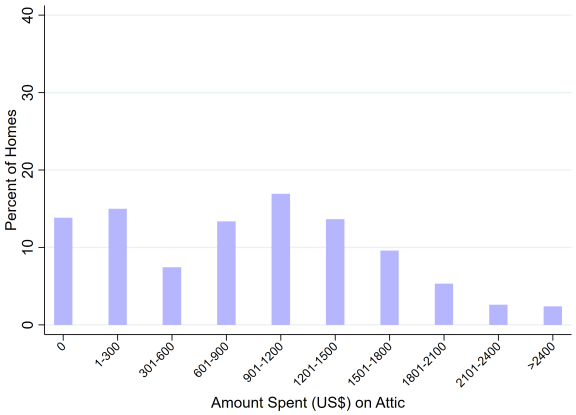


a: Full Sample

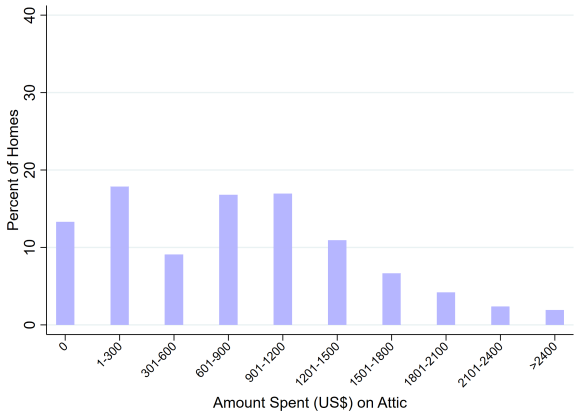
b: Top 20% Contractors

c: Bottom 20% Contractors

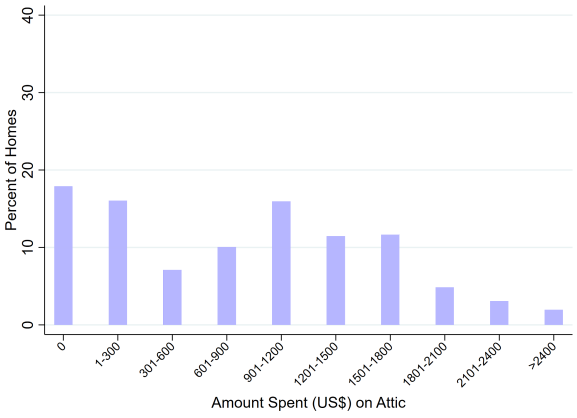
Figure F.2: Histograms for Amount Spent on Attic



a: Full Sample

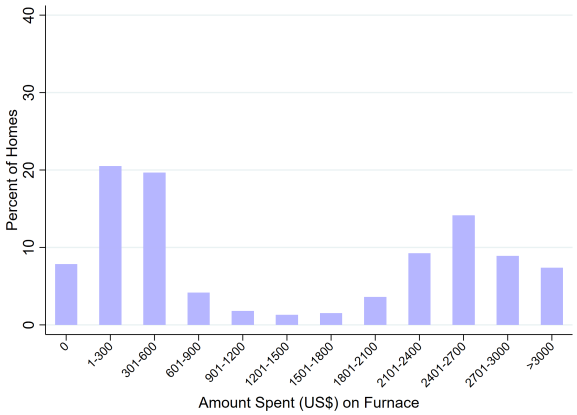


b: Top 20% Contractors

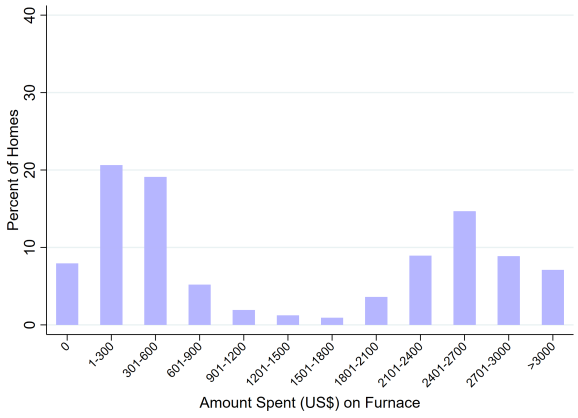


c: Bottom 20% Contractors

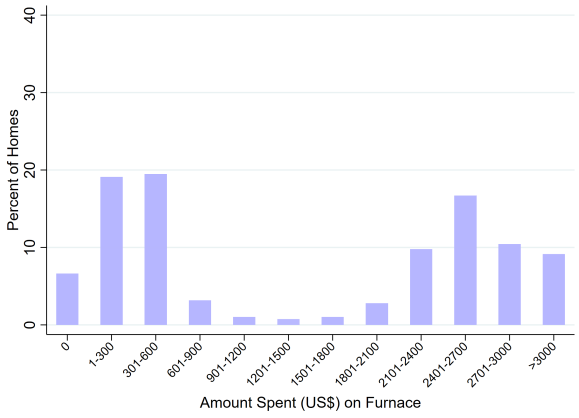
Figure F.3: Histograms for Amount Spent on Furnace



a: Full Sample

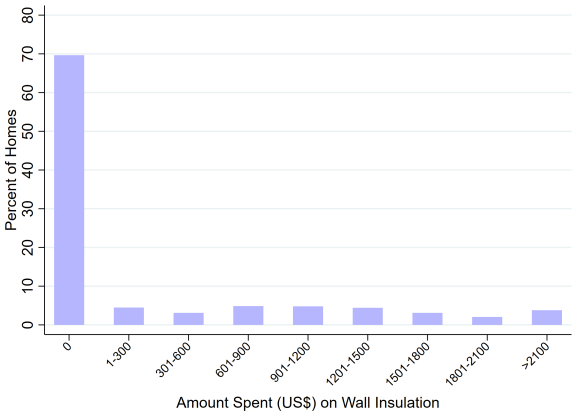


b: Top 20% Contractors

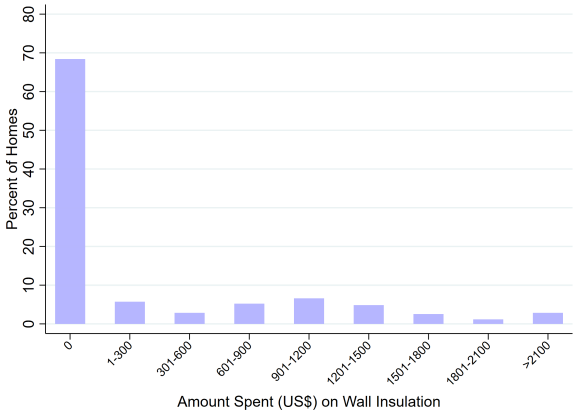


c: Bottom 20% Contractors

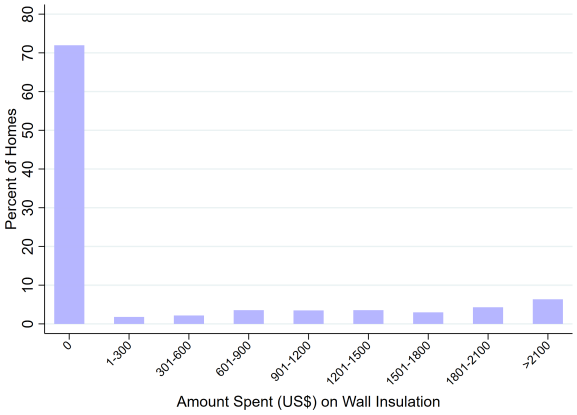
Figure F.4: Histograms for Amount Spent on Wall Insulation



a: Full Sample

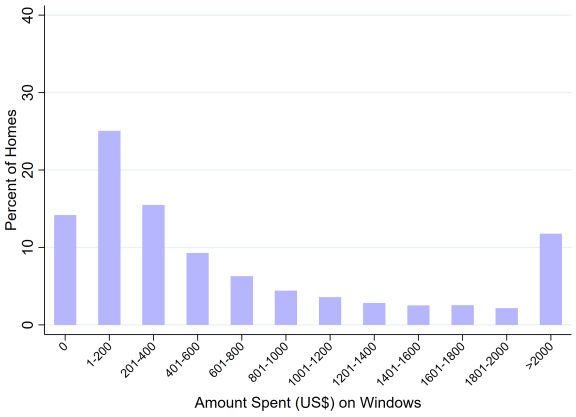


b: Top 20% Contractors

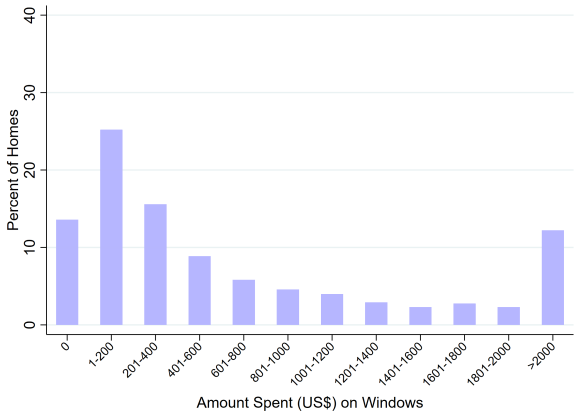


c: Bottom 20% Contractors

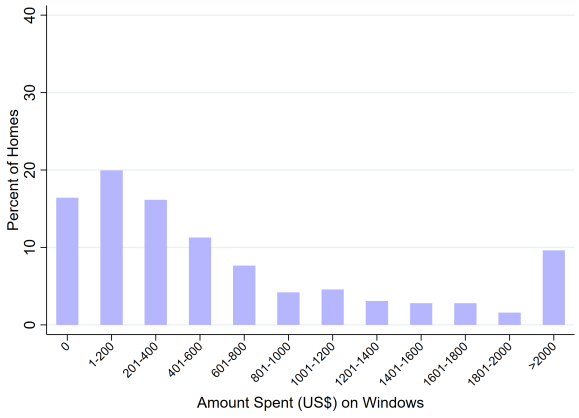
Figure F.5: Histograms for Amount Spent on Windows



a: Full Sample

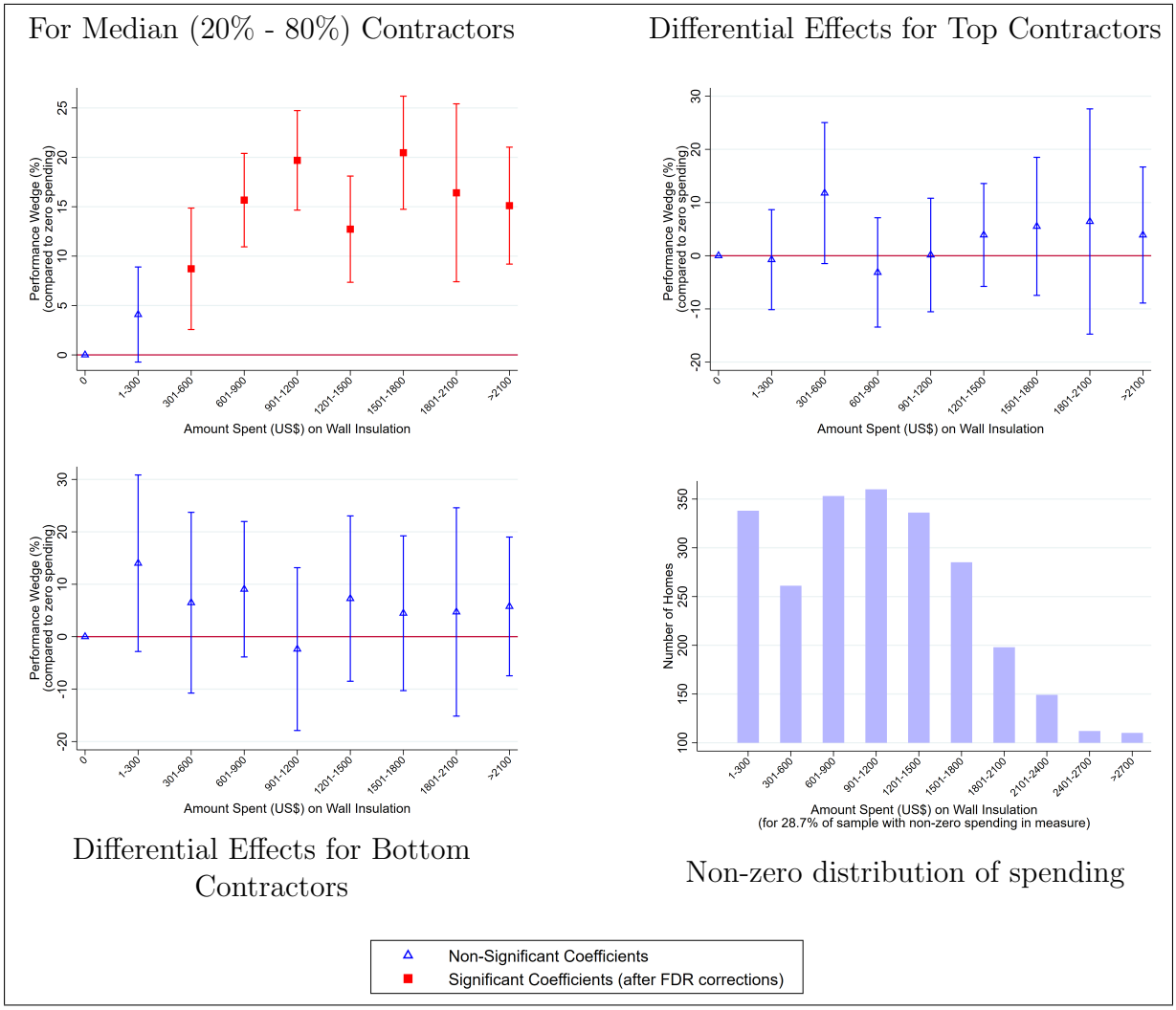


b: Top 20% Contractors



c: Bottom 20% Contractors

Figure F.6: Wedge by Wall Insulation, Interacted with Contractor Quality



Notes: This figure presents estimates of how the performance wedge is affected by additional spending on Wall Insulation. Coefficients are interpreted as percentage point increase/reduction in the wedge, relative to the omitted category (zero spending). The top right and bottom left panels represent additional interactive effects for the top and bottom performing contractors, respectively. The whiskers around the point estimates represent 95% confidence intervals, based on bootstrapped standard errors. P-values have been corrected with the false discovery rate (FDR) procedure from Benjamini and Hochberg (1995), where red indicates significance after these corrections. We assume an overall uncorrected critical p-value of 0.05 for each group. Uncorrected p-values within groups are assumed to be nonnegatively correlated. Results are robust across FDR or family-wise error rate (FWER) correction procedures. The bottom right panel presents the number of homes with spending in a given category.

G Further Details on PRISM and Behavioral Analyses

G.1 Engineering Structural Model for Residential Space Heating

In order to understand how household behavior might interact with the projected savings from WAP, we begin by describing the components of the physical relationships that drive space heating requirements for residential space. The heat interchange between a house and its surroundings can be written as follows (Johannesson et al., 1985):

$$\phi_h + \phi_i + \phi_s = \frac{A}{Be^F}(T_i - T_o) + S \tag{G.1}$$

where ϕ_h is the thermal output from the heating system (MMBtu), ϕ_i are internal heat gains from inhabitants, lighting, and other appliances (MMBtu), and ϕ_s are heat gains from absorption of solar radiation (MMBtu). The surface area of the house is indicated by A, while B is the thermal resistance of the wall, and e^F is the efficiency of the furnace. Taken together, the term $\frac{A}{Be^F}$ measures transmission and ventilation heat losses (MMBtu/^{o}F). The indoor and outdoor temperatures are T_i and T_o respectively (^{o}F), and S represents the rate of heat storage within the structure (MMBtu). Although all the terms vary with time, the equation may be applied to mean values over longer periods, such as a monthly billing cycle.

We test if changes in consumer preferences for indoor temperature T_i after weatherization can lead to significant differences in energy output projected by equation (G.1). The outdoor air temperature at which a heating system must be turned on to maintain a household's desired indoor temperature is known as a "balance point" (T_b). It is a function of indoor temperature (T_i), internal (ϕ_i) and solar (ϕ_s) gains, as well as transmission and ventilation heat losses (H), where $H = \frac{A}{Be^F}$.

$$T_b = T_i - (\phi_i + \phi_s)/H \tag{G.2}$$

The monthly output required from a heating system for a particular balance point, T_b, is linear in the outdoor air temperature as follows:

$$\phi_h = H(T_b - T_o) \tag{G.3}$$

We drop S from equation (G.1), assuming that net heat storage will be negligible over a heating season (Johannesson et al., 1985). An increase in a household's chosen indoor air temperature, T_i, increases energy consumption by raising the balance point, T_b, thus causing the heating system to turn on at higher outdoor air temperatures. The change in balance point following weatherization is as follows, where superscripts indicate pre and post weatherization and $\Delta T_b = T_b^{post} - T_b^{pre}$.

$$\Delta T_b = \Delta T_i - \left(\frac{\phi_i^{post} + \phi_s^{post}}{H^{post}} - \frac{\phi_i^{pre} + \phi_s^{pre}}{H^{pre}} \right) \tag{G.4}$$

In addition to the effects from households changing their preferred indoor air temperature, the balance point will be affected by weatherization through two structural channels. First, weatherization increases the thermal resistance of the structure so that $H^{post} > H^{pre}$, which serves to lower the balance point, such that the furnaces turn on later in the season, at colder temperatures. Second, more efficient lighting, appliances, and windows lower internal gains ($\phi_i + \phi_s$), and potentially counteract some of the effects of the change in H on the balance point. Lighting is upgraded in all homes as part of the weatherization process. In certain cases, new windows may be installed. Refrigerators are rarely replaced.

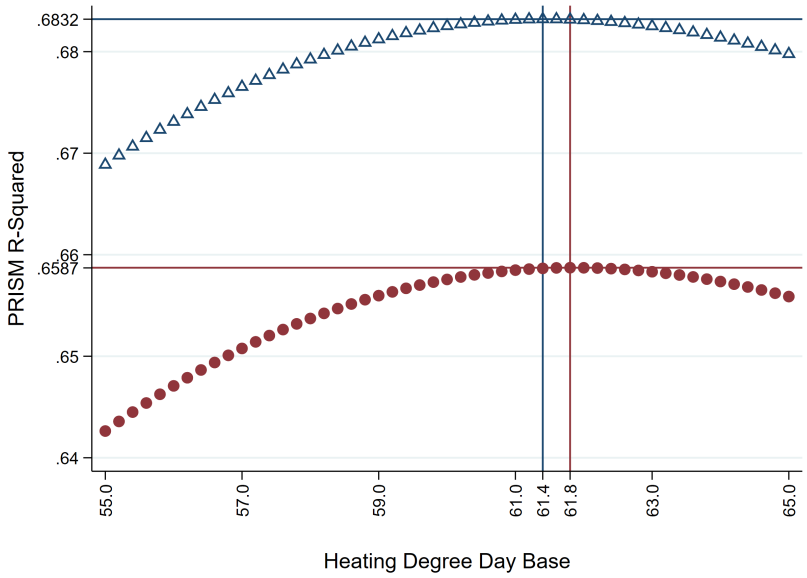
As described in the main text, we use the Princeton Scorekeeping Method (PRISM) to identify the balance point T_b for a given sample (Fels, 1986). We regress home-by-month energy usage on a constant plus heating degree days (HDD), iterating through several HDD bases s:

$$Y_{it} = \alpha + \beta \text{HDD}_{it}^s + \varepsilon_{it} \,, \tag{G.5}$$

In the main text, we present results for a specification that adds housing and demographic controls. Results without controls, presented in Figure G.1, suggest an increase

of 0.4^oF comparing pre- and post-weatherization samples.

Figure G.1: PRISM Optimal Heating Degree Day Bases - specification without controls



Notes: This figure plots results from PRISM analyses using the full sample of WAP homes. We iterate through many temperatures to identify the optimal HDD balance points for an average home, both before and after treatment. Balance points with highest R-squared are considered optimal. Results suggest an increase in balance points after treatment, such that heating systems turn on *earlier* in the season. That is evidence of behavioral effects, since most of WAP measures are expected to lower balance points (e.g. better heat retention should lead to heating systems turning on *later* in the season).

G.2 Savings Attributed to Heating Efficiency

The framework presented in the subsection above can be used to estimate how improvements to heating efficiency contribute to the overall program savings. For simplicity, here we focus on homes that use natural gas as their main heating fuel, as in the analyses from section 4.3. First, we estimate PRISM models, as in equation (5) from the main text, separately for pre- and post-treatment data. Then, we use the parameters from the pre-treatment PRISM model to predict the counterfactual (without weatherization) natural gas usage ($\hat{Y}_{it}^{PRISM}(0|D_{it}=1)$) during the post-treatment period. Further, we use the parameters from the post-treatment PRISM model to predict post-treatment usage with weatherization ($\hat{Y}_{it}^{PRISM}(1|D_{it}=1)$).

Lastly, we predict post-treatment energy usage with the pre-treatment model parameters, except for the slope β on the heating degree days, for which we use the post-treatment coefficient. We can denote those predictions as $\hat{Y}_{it}^{PRISM}(eff.|D_{it}=1)$, rep-

resenting the post-treatment natural gas usage if the program had only implemented improvements to heating efficiency. These elements allow us to estimate the savings from efficiency improvements as follows:

$$\hat{b}_{it}^{PRISM}(eff.) = \hat{Y}_{it}^{PRISM}(eff.|D_{it} = 1) - \hat{Y}_{it}^{PRISM}(0|D_{it} = 1) \ .$$

Those can then be compared to the total savings according to the PRISM model:

$$\hat{b}_{it}^{PRISM}(tot.) = \hat{Y}_{it}^{PRISM}(1|D_{it} = 1) - \hat{Y}_{it}^{PRISM}(0|D_{it} = 1) \ .$$

We find that the average savings from heating efficiency alone are about 1.1 MMBtu (15.76%) per month, while total savings are about 1.33 MMBtu (19.03%) per month according to the PRISM model. This suggests that heating efficiency improvements are the main channel through which the weatherization program operates. Almost 83% of the savings may be attributed to heating efficiency improvements, while the remaining 17% may be attributed to baseload or behavior changes.

G.3 Effects of Alternative Behavioral Factors on The Wedge

We investigate if behavioral factors, other than rebound effects on indoor temperature, may affect our estimates of the performance wedge. Specifically, we look at: (1) existence of non-working furnaces prior to WAP treatment, such that households were substituting to other heat sources (e.g. oven, electric space heating), or heat retention strategies (e.g. more layers of clothing, blankets); (2) occupants that were not home for significant periods of the year. For example, if the furnace was not working prior to the intervention, depending on how the household was compensating for the lack of central heat, installation of a new furnace could lead to substantial increases in overall energy consumption. If some occupants are typically away from home for significant periods of the winter, then realized savings from an improved envelope would be a smaller percentage than if they were home all year round. Non-working furnaces are directly identified by WAP energy auditors. To identify homes absent of occupants, we use a variation of the Princeton Scorekeeping Method (PRISM) described in the main text.

Rather than fitting the PRISM equation (5) for the whole sample, we fit it separately for each home. That allows us to identify homes for which energy consumption reliably follows variations in outdoor temperature. Strong deviations from those patterns are attributed either to measurement error, or behavioral discrepancies, such as a family not being home during winter (thus leading to low usage during those months). If the fit of equation (5) at a home's optimal balance point is not strong enough (R-squared is below 0.85), then it is deemed non-temperature responsive. In Figures G.2 and G.3 we report correlations between R-squares and optimal heating degree day bases according to this PRISM analysis. We assume that a given home is responsive to variations in outdoor temperature if the optimal base is between 43 and 73 degrees, with an R-squared above 0.85.

We then estimate our "wedge regression," equation (2), including those indicator variables for non-working furnace and non-responsiveness to outdoor temperatures. Table G.1 presents the results of this estimation. Neither of these factors appear to contribute in an economically or statistically significant way to the wedge between projected and realized savings.

Table G.1: Effects of Non-Working Furnace and Failing PRISM Restrictions

Outcome: *Percent Performance Gap*	
Non-Working Furnace	-0.8670
	(1.4576)
Failed PRISM	0.6477
	(0.9906)
Observations	60,855

Notes: This table presents how the performance wedge is different for homes that had a non-working furnace pre-treament and for homes that failed the PRISM sample restrictions (meaning that their energy consumption patterns are unresponsive to changes in outdoor temperature). These are coefficients obtained from the "wedge regression" described in the main text. That regression controls for other factors that can affect the wedge, such as housing structure, demographics, weather, and program spending. Note that none of the coefficients are significant, indicating that those set of homes do not have an average performance wedge that is different form the rest of the sample. Coefficients were obtained from a regression of the performance wedge on indicators for those two conditions, plus program spending variables, weather controls, demographics, and housing structure. Some homes were drop from this analysis because the home-specific PRISM procedure requires a full year of data both pre- and post-treatment. Standard errors in parentheses are bootstrapped.

Figure G.2: Correlation Between PRISM Optimal HDD Base and R-Squared - Pre-Treatment

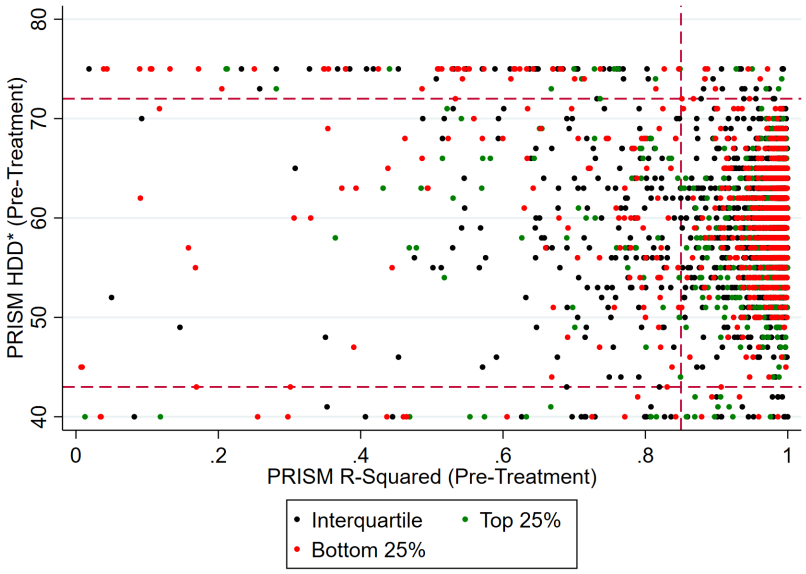
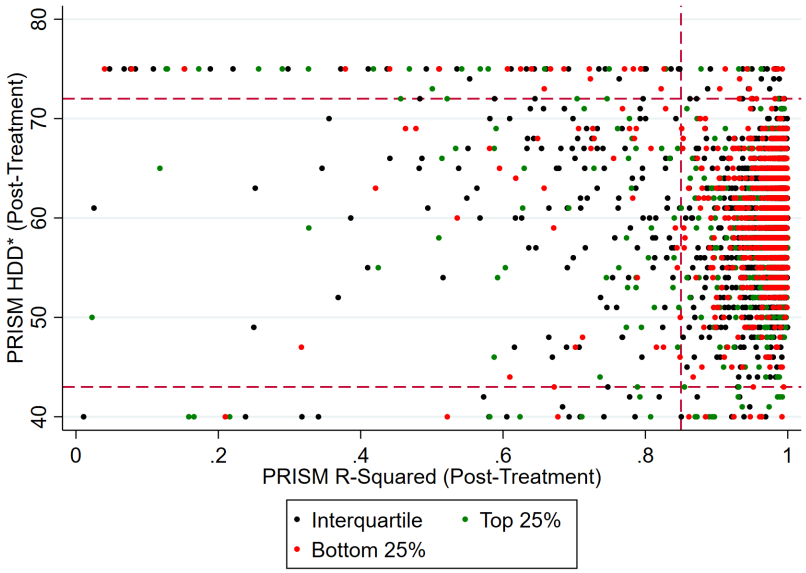


Figure G.3: Correlation Between PRISM Optimal HDD Base and R-Squared - Post-Treatment



H Further Details on Home-Specific Cost-Benefit Analysis

We use a two-step method presented by Souza (2019) to estimate the expected home-specific energy savings from weatherization. The first step consists of obtaining home-by-month energy savings $\hat{b}_{it}^{ml}$ with the ML method as described in section 3 of the main text. For the second step, we project those savings on available covariates as follows:

$$\hat{b}_{it}^{ml} = \alpha_0 + \sum_{k=1}^{K}\sum_{b=1}^{B_k} \beta_{kb} \mathbb{1}[Category = k]_{it} \cdot \mathbb{1}[Bin = b]_{it} + \sum_{g=1}^{G} \gamma_g X_{it}^{g} + \varepsilon_{it} \quad \forall\, t > t_i\,, \quad \text{(H.1)}$$

with notation as described in the main text. We use that model to obtain predictions of savings $\hat{\hat{b}}_{it}^{ml}$. Compared to equation (2) from the main text, equation (H.1) above differs in two ways: first, the outcome here are energy savings $\hat{b}_{it}^{ml}$; second, here we exclude contractor fixed effects.

As described in the main text, predictions from specification (H.1) capture heterogeneity in treatment effects based solely on the observable features of the house and household. Within the same context of this study, simulations from Souza (2019) illustrate how this two-step approach filters out unobservable home-specific idiosyncrasies. Therefore, the ranking it produces will reflect heterogeneity in savings based on characteristics that are observable to practioners ex ante.

Tables H.1 H.2 below present detailed results on our cost-benefit analyses. Table H.1 tests for sensitivity of net present benefits to varying assumptions, described in detail in the main text. Table H.2 presents both total program costs and total program benefits from weatherizing different subsamples of homes, depending on their cost-effectiveness ranking. Panel A monetizes benefits according to the social cost of carbon, while Panel B uses retail energy prices. Table H.2 uses baseline assumptions: discount rate of 3% and average upgrade lifespans of 30 years.

Table H.1: Heterogeneity in Cost-Benefit Estimates

Lifespan	Discount Rate	Percentile of Homes: All	Top 99%	Top 95%	Top 90%	Top 75%	Share with TB≥TC	Share with MB≥MC
Panel A: Evaluated at Social Marginal Costs of Energy								
		NPB (million $) with Baseline Assumptions						
30 years	**3%**	**-1.51**	**-1.38**	**-0.60**	**0.16**	**1.76**	**92%**	**42%**
		NPB (million $) with Alternative Discount Rates						
30 years	0%	9.58	9.71	10.41	11.01	11.85	100%	71%
30 years	6%	-7.52	-7.38	-6.54	-5.65	-3.50	34%	14%
		NPB (million $) with Alternative Lifespans						
10 years	3%	-12.76	-12.60	-11.65	-10.61	-7.88	4%	2%
20 years	3%	-5.80	-5.66	-4.85	-4.01	-2.07	50%	21%
40 years	3%	2.07	2.20	2.95	3.66	5.01	100%	56%
Panel B: Evaluated at Retail Energy Prices								
		NPB (million $) with Baseline Assumptions						
30 years	**3%**	**1.08**	**1.22**	**2.03**	**2.80**	**4.28**	**100%**	**53%**
		NPB (million $) with Alternative Discount Rates						
30 years	0%	13.66	13.81	14.57	15.20	15.94	100%	75%
30 years	6%	-5.73	-5.59	-4.73	-3.86	-1.83	55%	23%
		NPB (million $) with Alternative Lifespans						
10 years	3%	-11.68	-11.52	-10.58	-9.57	-6.95	7%	3%
20 years	3%	-3.80	-3.66	-2.82	-1.99	-0.18	74%	32%
40 years	3%	5.16	5.30	6.10	6.82	8.05	100%	63%
Number of Homes		4,649	4,626	4,440	4,208	3,510		

Notes: This table presents our calculations for net present benefits of WAP as a whole. Baseline assumptions include a 30-year lifespan of upgrades, and a 3% discount rate. We present results varying those assumptions, as well as the sample of homes considered. Panel A presents estimates based on social marginal benefits, while Panel B is based on retail energy prices. For example, WAP net benefits in Panel A are around -$ 1.51 million with baseline assumptions and for the full sample. However, considering only the top 75% homes, program benefits can be up to $1.76 million.

Table H.2: Heterogeneity in Costs and Benefits With Baseline Assumptions

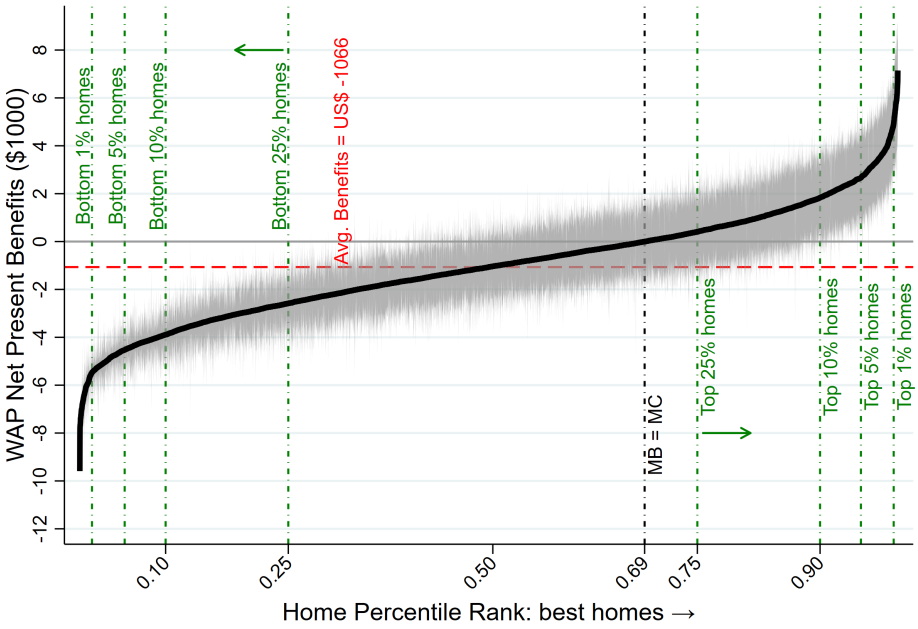
Panel A: Evaluated at Social Marginal Costs of Energy							
	Full Sample	Top 99% Homes	Top 95% Homes	Top 90% Homes	Top 75% Homes	TB≥TC	MB≥MC
Total Costs (million $)	21.52	21.37	20.36	19.17	15.60		
Total Benefits (million $)	20.01	19.99	19.76	19.32	17.36		
Net Benefits (million $)	-1.51	-1.38	-0.60	0.16	1.76	92%	42%
Number of Homes	4,649	4,626	4,440	4,208	3,510		
Panel B: Evaluated at Retail Energy Prices							
	Full Sample	Top 99% Homes	Top 95% Homes	Top 90% Homes	Top 75% Homes	TB≥TC	MB≥MC
Total Costs (million $)	21.52	21.38	20.40	19.25	15.86		
Total Benefits (million $)	22.60	22.61	22.44	22.05	20.13		
Net Benefits (million $)	1.08	1.22	2.03	2.80	4.28	100%	53%
Number of Homes	4,649	4,626	4,440	4,208	3,510		

Notes: Assuming a discount rate of 3% and average upgrade lifespans of 30 years.

H.1 Home-Specific Net Benefits with Alternative Sampling Restrictions

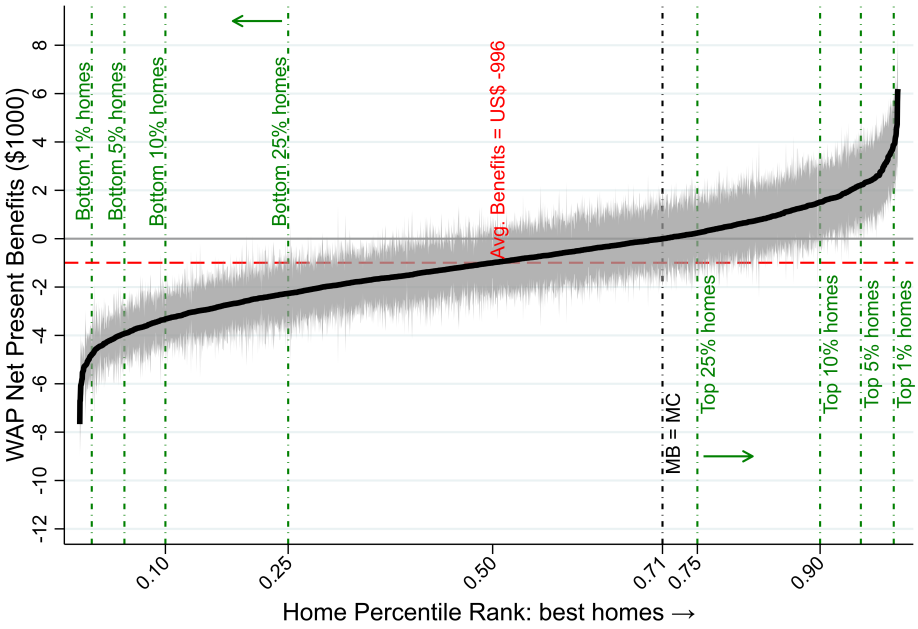
Figures H.1 and H.2 below plot, respectively, a ranking of home-specific net present benefits for our full study sample and for a PRISM-compliant sample of homes (i.e. PRISM R-squared above 0.85 and optimal HDD bases between 43 and 73 degrees). Note that the PRISM restrictions do not necessarily require homes with a full year of post-treatment data. For both figures, average program benefits are substantially lower than those presented in Figure 5 from the main text. That is due to the inclusion of homes with incomplete data (without a full year of post-treatment data) in Figures H.1 and H.2, which leads to underestimates of savings.

Figure H.1: Ranking of Homes by Net Present Benefits - not requiring a full year of pre- and post-data



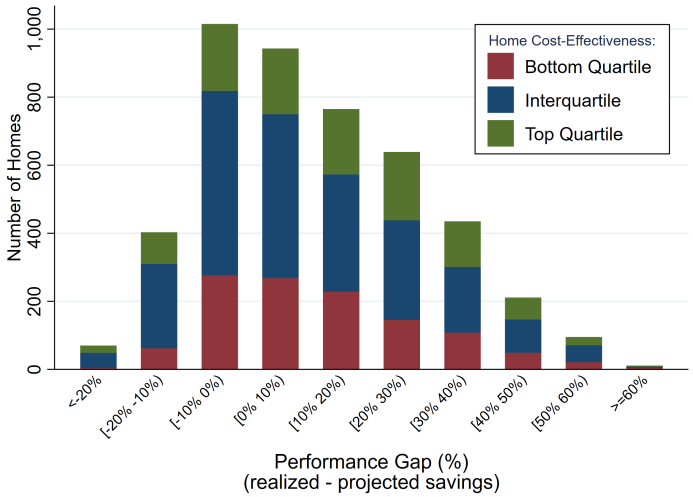
Note: 95% confidence intervals based on bootstrapped standard errors.

Figure H.2: Ranking of Homes by Net Present Benefits - PRISM sample restrictions



Note: 95% confidence intervals based on bootstrapped standard errors.

Figure H.3: Distribution of the Energy Savings Wedge, by Cost-Effectiveness Ranking



Notes: This graph presents the distribution of the performance wedge, correlating it with our calculations of WAP homes' cost-effectiveness. As expected, the wedge is smaller (larger) for the most (least) cost-effective homes.

H.2 Correlates With Cost-Effectiveness

Here we dig deeper into the factors that are associated with home cost-effectiveness. We compare the expenditures of homes in the top and bottom quartiles to those in the interquartile range. Results are reported in Table H.3 Panel A. On average, we find that per home expenditures in the bottom quartile of cost-effectiveness are $568 higher than those in the interquartile range, while expenditures were $465 lower in the top quartile of cost-effectiveness. This difference suggests that cost-effectiveness is likely affected by diminishing returns on measure spending.

When we examine particular types of retrofits, we find that the least cost-effective homes are characterized by higher spending on windows. This does not imply that window replacements completely fail to save energy. Rather, spending may have been at levels characterized by diminishing marginal returns. Conversely, the bottom homes spend less on attic insulation and foundation, which are, according to our assumptions, measures with inherently high value due to their long lifespans (50 years). Differences in wall insulation expenditures, however, are not statistically significant across groups, which is consistent with returns for that measure diminishing faster than what the engineering

models project.

In H.3 Panel B we show that, as expected, realized savings for homes increase as we move from the least to the most cost-effective homes. Projected savings also increase in that direction, nevertheless the performance wedge remains substantial for the top performing homes. Panel B also provides suggestive evidence of a correlation between cost-effectiveness and the ratio of modeled to actual post-treatment *consumption*. Interestingly, that correlation does not seem to exist when looking pre-treatment consumption.

Finally, Table H.4 presents average home-specific net present benefits by program year. Here we note that the IHWAP implemented a number of changes to improve quality after Program Year 2013; results from Table H.4 suggest that the IHWAP's efforts yielded substantial benefits. The average social NPB for homes through program year 2012 was -$808 whereas the average social NPB for program years 2013 and later was $623, a $1,431 improvement.

Table H.3: Expenditures of Least and Most Cost-Effective Homes

	Least Cost-Effective (Bottom 25%) Homes (1)	Interquartile (25%-75%) Homes (2)	Most Cost-Effective (Top 25%) Homes (3)	Difference in Means Diff. (1)-(2)	Diff. (3)-(2)
Panel A: Nonzero Amount Spent per Home or Upgrade (US$)					
Total, Excluding Health and Safety	5165.234	4597.372	4132.073	567.861***	-465.300***
				(44.369)	(56.380)
Air Conditioning	535.845	512.766	277.924	23.079	-234.842
				(143.641)	(120.130)
Air Sealing	396.134	394.457	343.181	1.676	-51.277***
				(14.636)	(12.555)
Baseload	242.530	229.266	204.795	13.263	-24.472*
				(9.801)	(9.686)
Door	515.926	337.582	249.868	178.344***	-87.714***
				(14.526)	(9.708)
Foundation	566.697	641.370	639.340	-74.673**	-2.030
				(27.038)	(24.039)
Furnace Repair	532.094	420.124	359.303	111.970***	-60.821***
				(17.583)	(15.391)
Furnace Replacement	2659.821	2620.282	2582.130	39.539	-38.152
				(25.254)	(24.244)
General	709.830	376.918	267.526	332.912***	-109.392*
				(59.985)	(52.807)
Health and Safety	468.150	527.390	605.216	-59.240***	77.826***
				(11.948)	(14.898)
Window	1602.355	649.296	329.825	953.059***	-319.471***
				(36.830)	(18.351)
Water Heater	280.913	264.390	246.154	16.523	-18.237
				(12.683)	(13.851)
Insulation Measures					
Attic Insulation	791.401	1081.339	1216.315	-289.938***	134.977***
				(27.084)	(22.437)
Wall Insulation	1207.796	1131.467	1115.909	76.329	-15.558
				(67.003)	(44.423)
Panel B: Energy Usage and Savings					
Ratio Modeled/Actual Usage Pre	2.225	2.350	2.293	-0.126*	-0.058
				(0.056)	(0.075)
Ratio Modeled/Actual Usage Post	1.972	1.860	1.545	0.111**	-0.315***
				(0.043)	(0.043)
Realized Savings (%)	-10.778	-15.937	-18.939	5.158***	-3.002***
				(0.436)	(0.382)
Projected Savings (%)	-22.108	-25.361	-32.311	3.253***	-6.950***
				(0.636)	(0.700)

Notes: This table compares average nonzero expenditures of least, most, and interquartile cost-effective homes, according to the home-specific cost-benefit analyses (adjusting for social marginal benefits). The fourth column presents differences in means between least cost-effective and interquartile homes. The fifth column presents differences in means between most cost-effective and interquartile homes. Bootstrapped standard errors are presented in parentheses. Significance at 1%, 5% and 10% are indicated by ***, ** and *, respectively.

Table H.4: Average Net Present Benefits by Program Years

Panel A: Evaluated at Social Marginal Costs of Energy			
Program Years	Average NPB (US$)	Std. Dev.	Number of Homes
PY 2009	-432.96	1909.92	497
PY 2010	-1019.02	1819.57	1015
PY 2011	-1143.87	1748.79	990
PY 2012	-176.37	1903.50	570
PY 2013	722.17	2110.03	489
PY 2014	738.36	1805.59	438
PY 2015	618.92	1816.10	554
PY 2016	-381.92	1850.94	96
PYs 2009-2012	-808.09	1867.21	3072
PYs 2013-2016	623.18	1927.35	1577

Panel B: Evaluated at Retail Energy Prices			
Program Years	Average NPB (US$)	Std. Dev.	Number of Homes
PY 2009	90.05	2280.92	497
PY 2010	-564.90	2202.13	1015
PY 2011	-745.66	2147.72	990
PY 2012	322.46	2290.48	570
PY 2013	1510.67	2472.68	489
PY 2014	1496.84	2179.20	438
PY 2015	1375.84	2152.61	554
PY 2016	238.60	2211.37	96
PYs 2009-2012	-352.55	2253.89	3072
PYs 2013-2016	1382.03	2284.55	1577

Notes: This table presents average home-specific net present benefits by program year. Those were obtained by first estimating home-specific net benefits, as in section 5, and then taking simple averages of those net benefits based on which homes were served in each program year.

H.3 Costs per CO2

Table H.5 below presents authors' calculations of the weatherization program's implied average costs per CO2 abatement. Net present benefits calculated in Section 5 are compared to projections of CO2 abated over the lifespan of retrofits. We assume a 3% discount rate and a social cost of carbon of $40 per ton.

Table H.5: Average Costs per CO2 Abated

	Costs per ton of CO2 (US$ /tCO2)		
Lifespan Assumption	All Homes	Top 25% Homes	Bottom 25% Homes
10 years	194.05	76.82	358.62
20 years	43.30	-18.94	178.16
30 years	7.65	-39.40	104.27
40 years	-7.85	-45.67	71.71