Experimental Research Designs – Energy Applications

Meredith Fowlie

ARE 261

Lecture 2

September 2021

RCTs in EEE

- Under certain conditions, a randomized control trial provides the most valid estimate of an intervention's causal impact.
- Assigning treatment at random ensures that there is no systematic variation in unobserved factors across treated and untreated groups.
- Real-world complications can pose challenges for experimental research design and implementation.
- Increasingly detailed/high frequency data can also open up the possibility for richer experimental analysis.

Lecture Game Plan

- Complication 1: Hawthorne effects
  - Gosnell, List, and Metcalfe (2019)
- Complication 2: Imperfect Compliance
  - Fowlie, Greenstone, Wolfram (2018)
- Heterogeneous treatment effects
  - Knittel and Stolper (2021)

Hawthorne effects

- Some influential experiments were conducted at the Western Electric Company Hawthorne Plant near Chicago in 1924.
- Associated studies were designed to shed light (ha!) on how shop-floor lighting affected workers' productivity.
- Researchers concluded that workers' knowledge that they were being experimented upon seemed to directly alter workers' behavior.

Hawthorne effects and energy consumption

- Why might expect Hawthorne effects to be an issue in the context of household energy consumption?

Hawthorne effects and energy consumption

- Why might expect Hawthorne effects to be an issue in the context of household energy consumption?
- People do not pay attention to electricity/gas consumption in general. If participating in a study increases salience, this could impact consumption!
- Two studies underscore the importance of assessing the impact of being in a study of energy consumption behavior, separate from making inferences about the impact of experimental manipulations.

Schwartz et al. 2013

- Participants were randomly selected from residential customers of a mid-Atlantic electricity utility.
- Households in the treatment group received their notifications that they had been selected to be in a one month long study about electricity use in their home.
- No action was required on the part of 'treated' households.
- The control group received nothing.
- Intervention reduced monthly use by 2.7%(!!!)!
- Treatment effect vanished when the intervention ended.

Gosnell, List, and Metcalfe (2019)

- Researchers observe more than 110,000 binary behavioral outcomes across 40,000 unique flights over a 27-month period for the entire population of captains within Virgin Atlantic Airways(!)
- Pilots randomly assigned to:
  1. Arm's length performance monitoring (i.e., control group),
  2. Informational performance feedback
  3. Target setting
  4. Prosocial incentives
- Monitored outcomes: flight-level measures of fuel-related efficiency across three distinct phases—pre-flight, in-flight, and post-flight.
- Estimated fuel savings: \$0.98-\$1.96 million using an SCC range of of \$40-\$80.

Gosnell, List, and Metcalfe (2019)

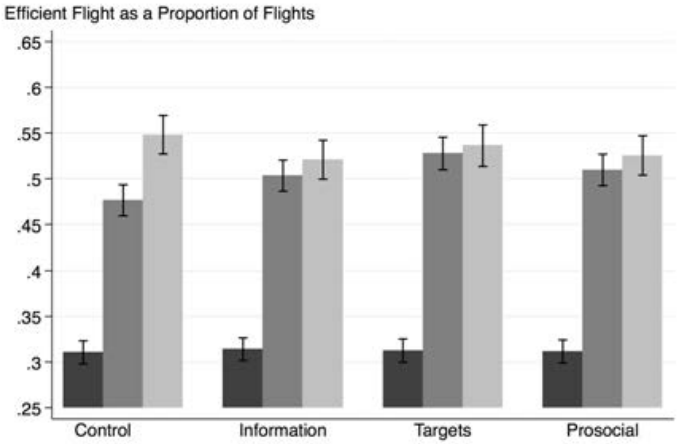
Table 1
Treatment Group Design

	Monitoring	Information	Targets	Prosocial
Control	✓			
Treatment Group 1	✓	✓		
Treatment Group 2	✓	✓	✓	
Treatment Group 3	✓	✓	✓	✓

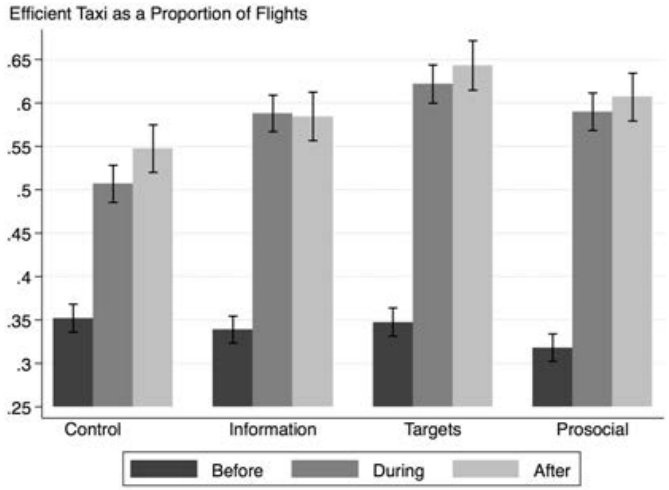
There is no 'pure' control - hard to implement if pilots are aware the experiment is underway.

How big is the Hawthorne effect?

(b) Efficient Flight



(c) Efficient Taxi



Notes: The y-axis for each of the above graphs represents the proportion of flights for which a fuel-related behavior has been implemented before

Take away?

- Hawthorne effects are a potential source of bias in randomized experiments.
- In general, when you are experimenting with an intervention that is designed to change behavior, you need to be able to disentangle the effect of the intervention from the effects of the experiment.
- Keep this in mind as you design your experiments! Build in checks and balances if you have the statistical power to do so.
- Also - some compelling questions about Hawthorne effects and habit formation. Can the way in which an intervention is introduced affect long run impacts?

Lecture Game Plan

- Complication 1: Hawthorne effects
  - Gosnell, List, and Metcalfe (2019)
- Complication 2: Imperfect Compliance
  - Fowlie, Greenstone, Wolfram (2018)
- Emerging opportunity: Rich data!
  - Knittel and Stolper (2020)

RCTs work really well.....

When the unit of study stays where you put/assign it.

source:www.education.com/reference/controlled-experiments/

Economic agents are not avocado pits!

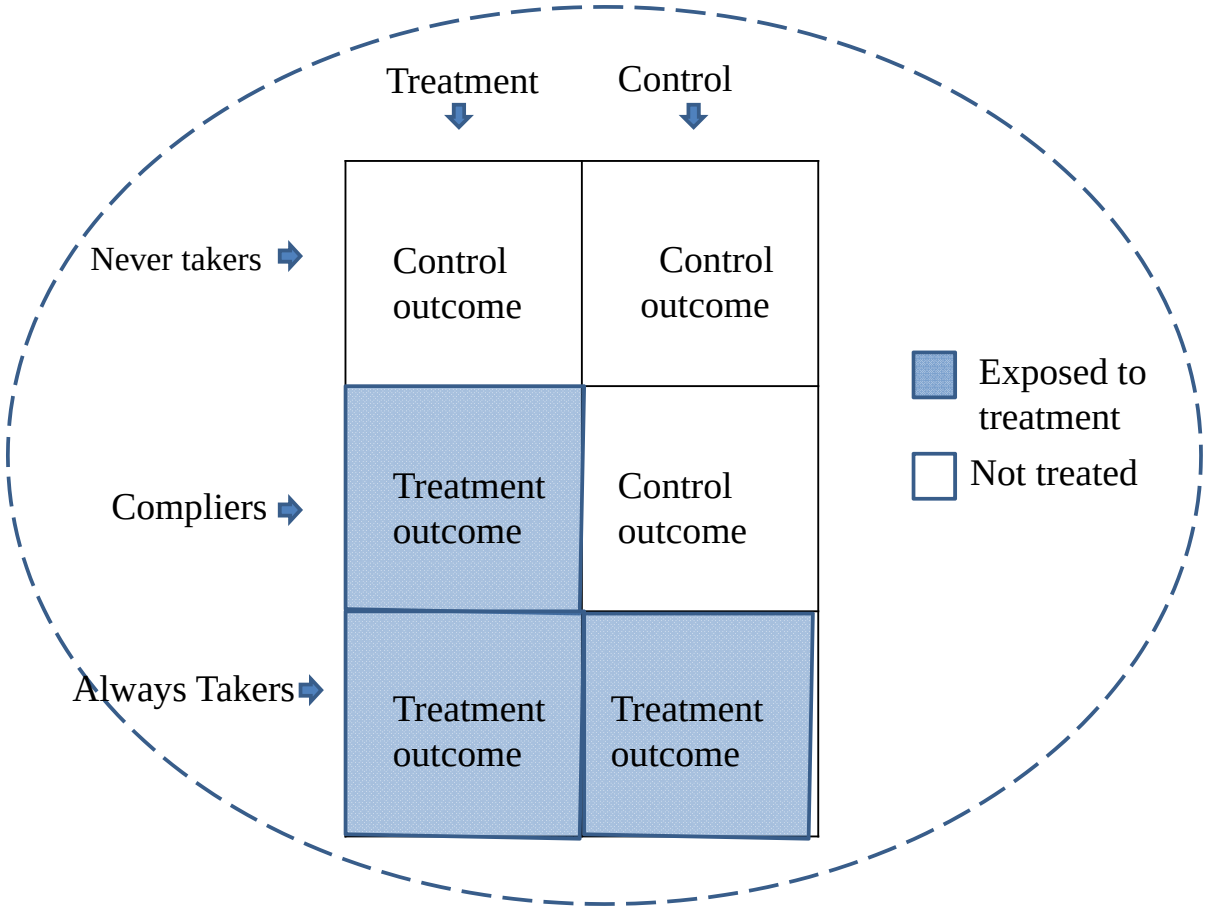
- Economic researchers rarely have complete control over treatment assignment... individuals or firms may not stay where you want to put them!
- "Non-compliance" with treatment assignment can undo the random assignment to an intervention in a very non-random way.
- This can undermine the credibility of (and introduce bias into) causal impact estimates.
- What's the problem?

Imperfect compliance with treatment assignment

Under imperfect compliance, we make a conceptual distinction between (four) types of units (e.g. households, firms, etc.).

1. **Always takers** take up the intervention regardless of their treatment assignment.
2. **Never takers** do not take up the intervention regardless of their treatment assignment.
3. **Compliers** comply with their treatment assignment.
4. **Defiers** do the opposite of what you assign them to do!

Imperfect compliance with treatment assignment



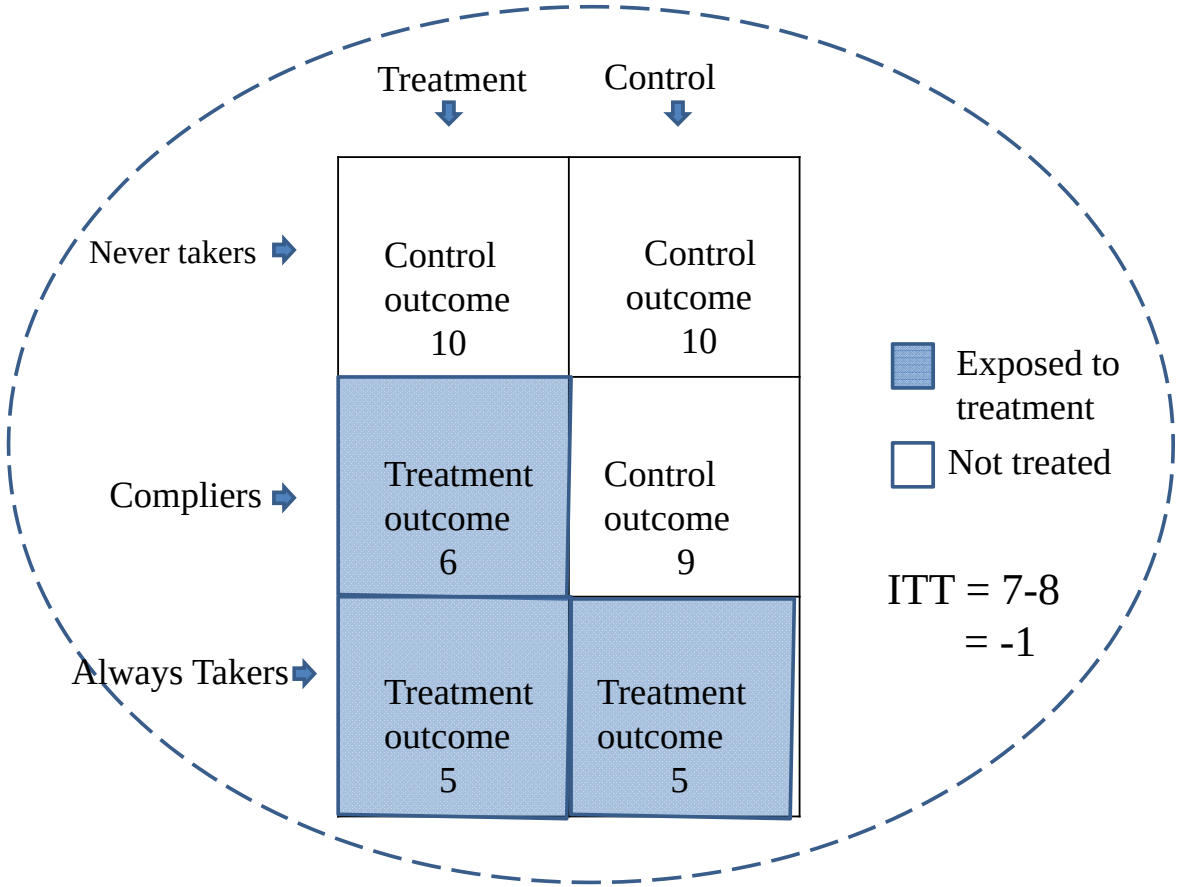
The composition of treated units and untreated units is no longer the same (in expectation).

When mandatory assignment fails?

1. Be satisfied with estimating the effect of the *intent* to treat.
2. Redefine your population of interest to include only those whose treatment status you *can* manipulate.
   - Recruit-and-deny
   - Recruit-and-delay
3. Randomly manipulate the *probability* of treatment.

The intent to treat (ITT)

We can obtain an unbiased estimate of the effect of treatment *assignment* (versus the treatment itself).



When mandatory assignment fails?

When random assignment of a treatment/intervention to individuals in a population is impossible, impractical, or ineffective, you have some options...

1. Be satisfied with estimating the effect of the *intent* to treat.
2. Redefine your population of interest to include only those whose treatment status you *can* manipulate.
   - Recruit-and-deny
   - Recruit-and-delay
3. Randomly manipulate the *probability* of treatment.

"Oversubscription approaches"

Good news:

- Non-compliance with treatment assignment less likely to be a problem (because you've pre-screened the sample).

Limitations?

"Oversubscription approaches"

Good news:

- Non-compliance with treatment assignment less likely to be a problem (because you've pre-screened the sample).

Limitations?

- A comparison of average outcomes across treated and control groups yields an unbiased estimate of what?
  - The causal effect of the treatment in the sub-set of the population that sought out the treatment.
- Construct validity?

When mandatory assignment fails?

When random assignment of a treatment/intervention to individuals in a population is impossible, impractical, or ineffective, you have some options...

1. Be satisfied with estimating the effect of the *intent* to treat.
2. Redefine your population of interest to include only those whose treatment status you *can* manipulate.
   - Recruit-and-deny
   - Recruit-and-delay
3. Randomly manipulate the *probability* of treatment.

When mandatory assignment fails

One sided noncompliance:

- The treatment is only made available through the experiment (the control group cannot access treatment).
- You might be tempted to compare the average outcome among those accepting the treatment with the average outcome in the control group.

When mandatory assignment fails

One sided noncompliance:

- The treatment is only made available through the experiment (the control group cannot access treatment).
- You might be tempted to compare the average outcome among those accepting the treatment with the average outcome in the control group.
- Self-selection out of the treatment opens up the possibility that confounding factors that determine selection also play a role in determining the outcomes.

Two-sided non-compliance

- Units assigned to your control group are able to select into the treatment if they so choose.
- Added identification complication: you may have units who will seek out treatment if assigned to the control group, but who will select out of treatment if assigned to the treatment group (defiers)

Randomized encouragement designs

- If mandatory random assignment infeasible/impractical, there may be a way to randomly manipulate the probability of treatment.
- By randomly assigning an encouragement into treatment, we can randomly manipulate the probability of receiving treatment *among those units whose treatment status is affected by our encouragement!*.
- This is an instrumental variables approach to identifying local average treatment effects.

Randomized encouragement design

Let Z_i denote encouragement status.

- **"Never takers"**: $D_i = 0|Z_i = 1,\ D_i = 0|Z_i = 0$.
- **"Always takers"**: $D_i = 1|Z_i = 1, D_i = 1|Z_i = 0$.
- **"Compliers"**: $D_i = 1|Z_i = 1,\ D_i = 0|Z_i = 0$.
- **"Defiers"**: $D_i = 0|Z_i = 1,\ D_i = 1|Z_i = 0$.

In order for the RED to work, the encouragement instrument must have a weakly positive effect on participation/treatment probabilities (i.e. no defiers!).

Sample composition

Let π^T represent the population proportions of compliance types, where $T = A$(always taker), N (never taker), C, (complier), D (defier).

Monotonicity implies that $\pi^D = 0$.

Under an RED, the units are distributed across treatment cells (up to a random error) as follows:

	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	$(\pi^{NT} + \pi^C)(1 - P)N$	$\pi^{NT} PN$
$D_i = 1$	$\pi^{AT}(1 - P)N$	$(\pi^{AT} + \pi^C)PN$

What allows us to estimate these these π proportions directly from sample statistics?

Sample composition

Let π^T represent the population proportions of compliance types, where $T = A$(always taker), N (never taker), C, (complier), D (defier).

Monotonicity implies that $\pi^D = 0$.

Under an RED, the units are distributed across treatment cells (up to a random error) as follows:

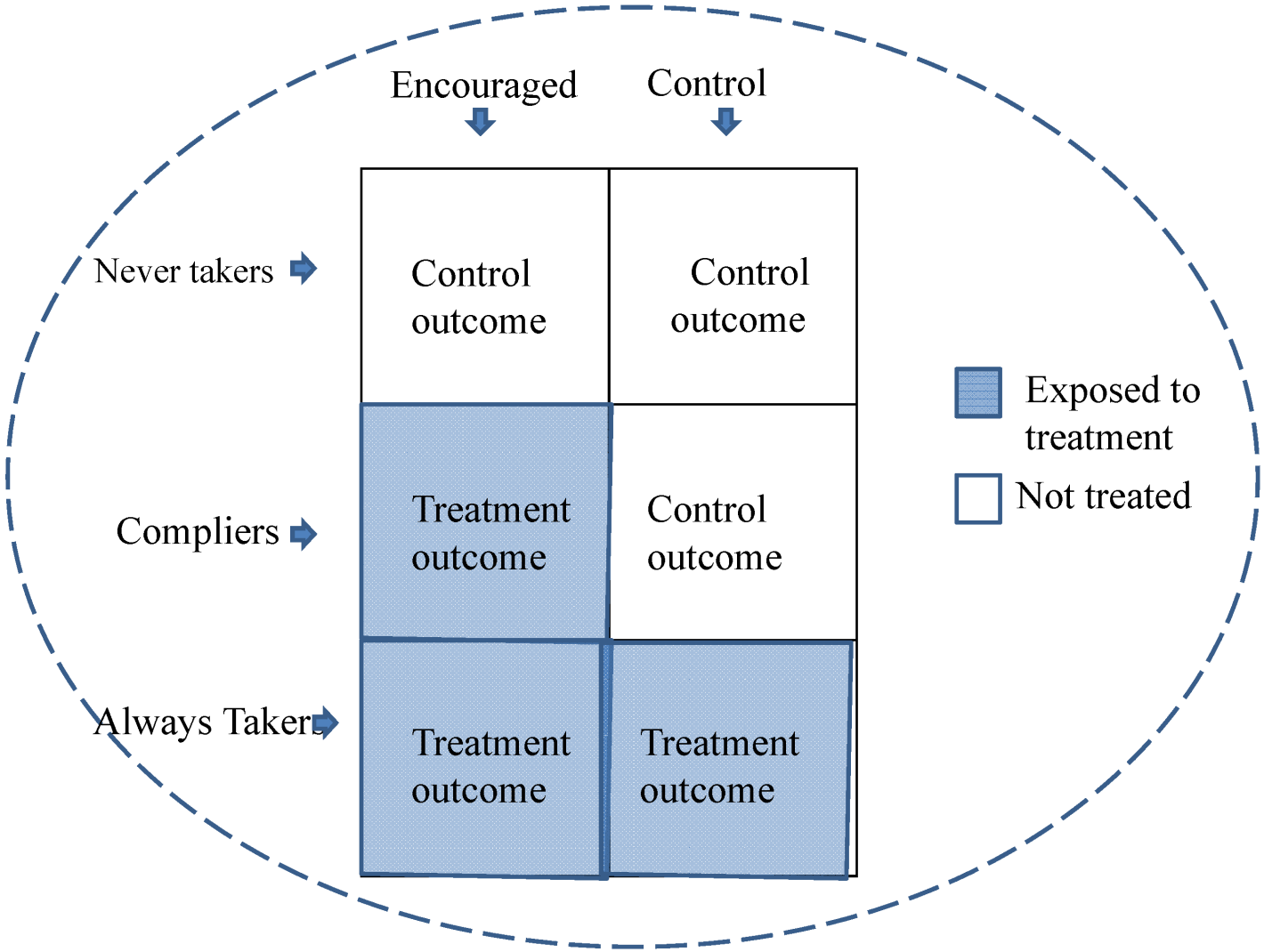
	$Z_i = 0$	$Z_i = 1$
$D_i = 0$	$(\pi^{NT} + \pi^C)(1 - P)N$	$\pi^{NT} PN$
$D_i = 1$	$\pi^{AT}(1 - P)N$	$(\pi^{AT} + \pi^C)PN$

What allows us to estimate these these π proportions directly from sample statistics?

Monotonicity, and randomization of the encouragement instrument assignment.

Randomized encouragement design (assume no defiers!)

The proportion of never takers/always takers/compliers is the same in expectation across the encouraged and control groups.



Randomized encouragement design (assume no defiers!)

More formally: :

$$
\begin{aligned}
E[Y_i|Z_i &= 1] = \pi^{NT} E[Y(0)|N] + \pi^{C} E[Y(1)|C] + \pi^{AT} E[Y(1)|A] \\
E[Y_i|Z_i &= 0] = \pi^{NT} E[Y(0)|N] + \pi^{C} E[Y(0)|C] + \pi^{AT} E[Y(1)|A]
\end{aligned}
$$

Taking the difference in these average outcomes:

$$
E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] = \pi^{C}(E[Y(1)|C] - E[Y(0)|C])
$$

RED Mechanics in 2 Stages

If we difference average outcomes across encouraged and control groups, we get the following:

$$\underbrace{\pi^{CM}}_{\text{Complier proportion}} * \underbrace{(Y_{CM}(1) - Y_{CM}(0))}_{\text{treatment effect among compliers}}$$

This should look familiar....

RED Mechanics in 2 Stages

If we difference average outcomes across encouraged and control groups, we get the following:

$$\underbrace{\pi^{CM}}_{\text{Complier proportion}} * \underbrace{(Y_{CM}(1) - Y_{CM}(0))}_{\text{treatment effect among compliers}}$$

This should look familiar....

This is the average effect of the intent to treat (ITT) *specific to the encouragement* Z.

RED Mechanics

First stage: Compare outcomes across encouraged and control groups

$$ITT = \underbrace{\pi^{CM}}_{\text{compliance rate}} * \underbrace{(Y_{CM}(1) - Y_{CM}(0))}_{\text{treatment effect among compliers}}$$

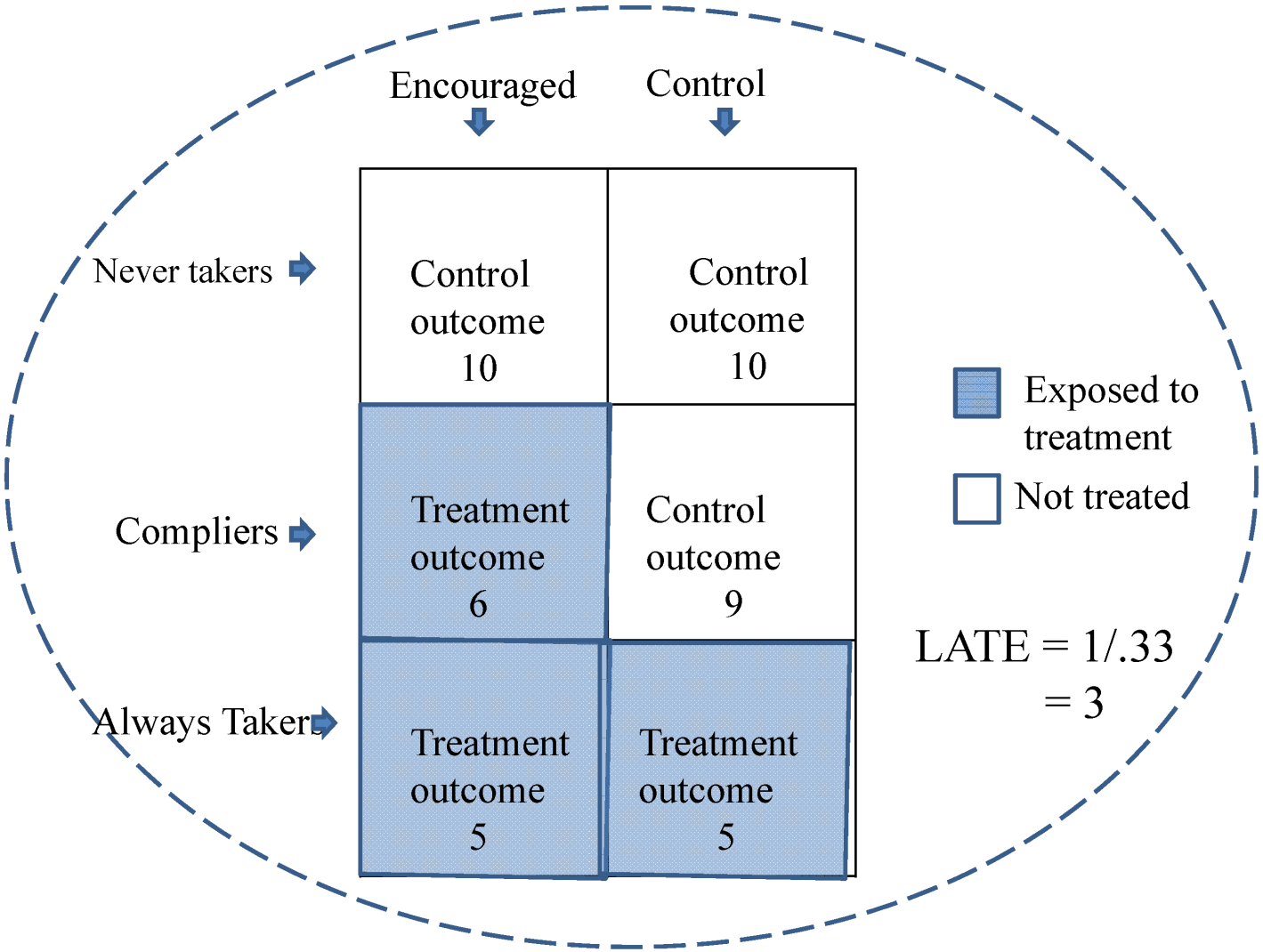
Second stage: Divide through by the proportion of compliers in the population to estimate a "local average treatment effect":

$$LATE = Y_{CM}(1) - Y_{CM}(0)$$

How exactly do we estimate π^{CM}?

Randomized encouragement design (assume no defiers!)

The proportion of never takers/always takers/compliers is the same in expectation across the encouraged and control groups.



Identification assumptions?

Identification assumptions?

- Ignorable assignment of the encouragement Z:

$$(Y_i(1), Y_i(0), D_{iZ=1}, D_{iZ=0}) \perp Z_i$$

- Stable unit treatment values
- An important exclusion restriction:

$$Y_i(D, Z = 0) = Y_i(D, Z = 1)$$

- Monotonicity

$$D_{iZ=1} \geq D_{iZ=0} \; \forall \; i.$$

The exclusion restriction:

Under the exclusion restriction, a comparison of average outcomes across encouraged and control groups yields:

$$\begin{aligned} &\pi^{AT}(Y(1)|Z=1,AT)+\pi^{C}(Y(1)|Z=1,C) \\ &\qquad +\pi^{N}(Y(0))Z=1,N) \\ -(&\pi^{AT}(Y(1)|Z=0,AT)+\pi^{C}(Y(0)|Z=0,C) \\ &\qquad +\pi^{N}(Y(0)|Z=0,N) \end{aligned}$$

If exclusion restriction holds, this difference reduces to: $\pi^{C}(Y(1)) - \pi^{C}(Y(0))$.

If it fails... we are confounding the LATE with the effects of the encouragement!

Monotonicity

Suppose that we violate the monotonicity assumption:

$$\begin{aligned} E[Y_i|Z_i = 1] &= \pi^N E[Y(0)|N] + \pi^C E[Y(1)|C] \\ &\quad + \pi^A E[Y(1)|A] + \pi^D E[Y(0)|D] \\ E[Y_i|Z_i = 0] &= \pi^N E[Y(0)|N] + \pi^C E[Y(0)C] \\ &\quad + \pi^A E[Y(1)|A] + \pi^D E[Y(1)|D] \end{aligned}$$

This implies:

$$\begin{aligned} E[Y_i|Z_i = 1] - E[Y_i|Z_i = 0] &= \pi^C (E[Y(1)|C] \\ &\quad - [Y(1)|C]) + \pi^D (E[Y(0)|D] - [Y(1)|D]). \end{aligned}$$

Yuck! The LATE is no longer identified!

RED meets WAP

- Empirical context
- Causal question to be answered
- What is being randomly assigned (and how)?
- Research design
- Key results
- Concerns and considerations (e.g. spillovers, non-compliance, etc.)

Weatherization (Fowlie et al. 2018)

Primary question: What is the causal impact of weatherization assistance on energy consumption/expenditures among participating households in Michigan?

Ancillary questions:

- How do experimental estimates of efficiency impacts compare to ex ante engineering estimates?
- Do we find evidence of "rebound" in the demand for home heating?
- Are these energy efficiency investments cost effective?

Conceptual framework

Returns on investment in energy efficiency are realized through two main channels:

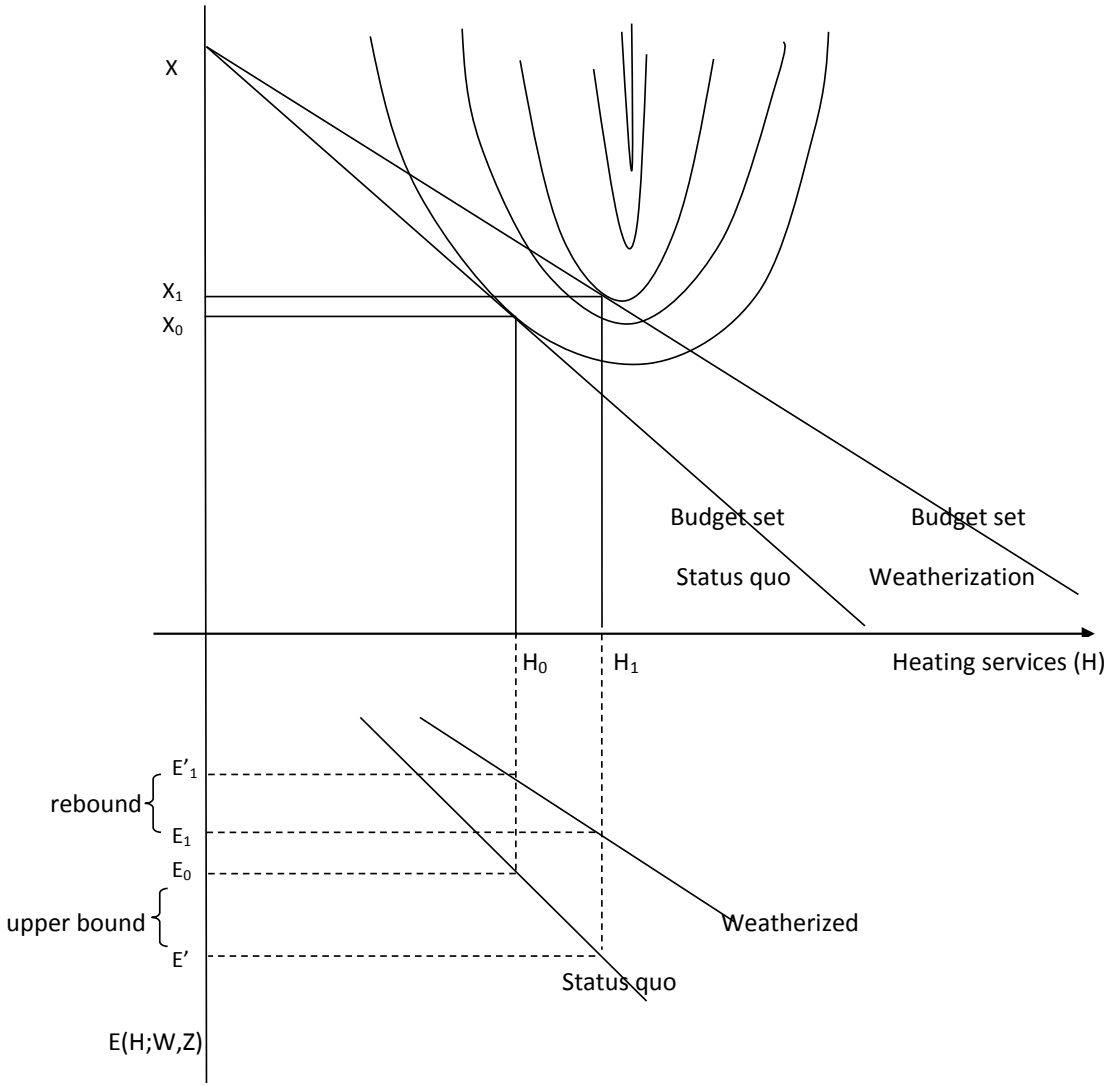
1. Reductions in energy expenditures.
2. Potentially, an increase in consumption of energy service in response to reduced cost.

A simple model informed by economic theory and basic principles of heat transfer in buildings provides a framework for interpreting our experimental estimates.

1. Reductions in energy expenditures measured directly.
2. Increase in consumption of energy services valued using a revealed preference approach.

Returns on investment manifest in two ways:

Figure 1 : Utility maximization of a representative low-income consumer



Federal weatherization assistance program (WAP)

- The largest residential energy efficiency program in the country,
- Weatherization retrofits (including insulation, furnace replacement, infiltration reduction) are provided for free to eligible households.
- An estimated 7 million U.S households have participated in the program.

source: http://www.princegeorgescountymd.gov/

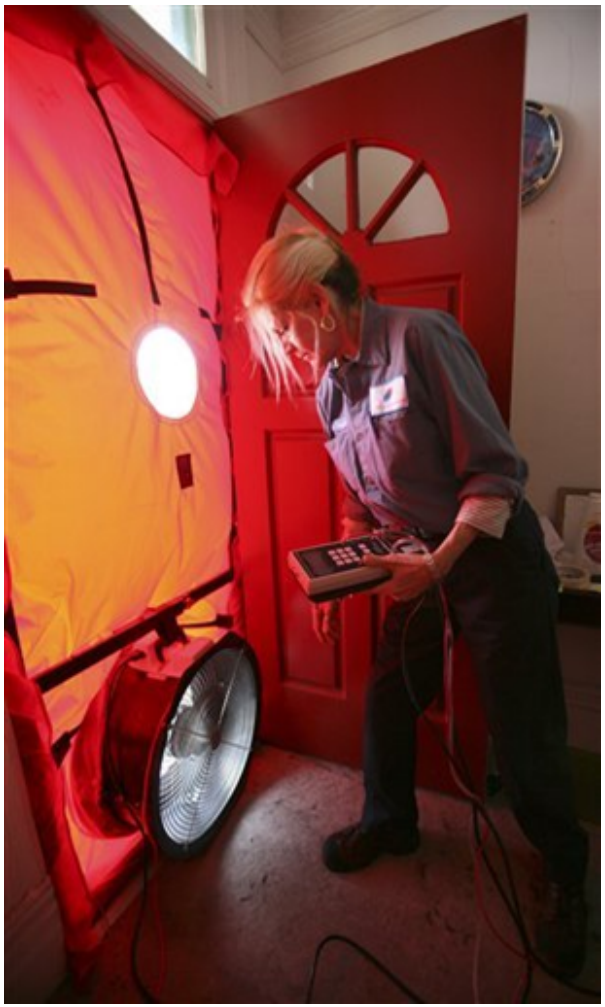
Empirical Context

- Energy efficiency audits and retrofits are provided for free to eligible households.
- The maximum average expenditure per household raised to $6,500 under ARRA.
- It is estimated that the average household receiving weatherization assistance will reduce heating costs by 20-25 percent in our field location (Michigan).

source: www.waptac.org

Federal weatherization assistance program

- Participating households receive a free energy audit/survey of building envelope characteristics.
- Audit data used to calibrate a detailed building simulation model (NEAT).
- In order for a measure to be implemented, the "cumulative" savings to investment ratio must be 1 or greater.



source: www.waptac.org

Empirical challenge

- To estimate the causal impacts of weatherization we need a credible estimate of counterfactual household energy consumption.
- We take two complementary strategies:
  1. **Randomized encouragement design**: Instrument for program participation using a randomly assigned encouragement intervention.
  2. **Quasi-experimental design**: Use data from all applicant households across five Michigan counties.

Our randomized encouragement design

Start with a sample of 34,161 presumptively eligible households

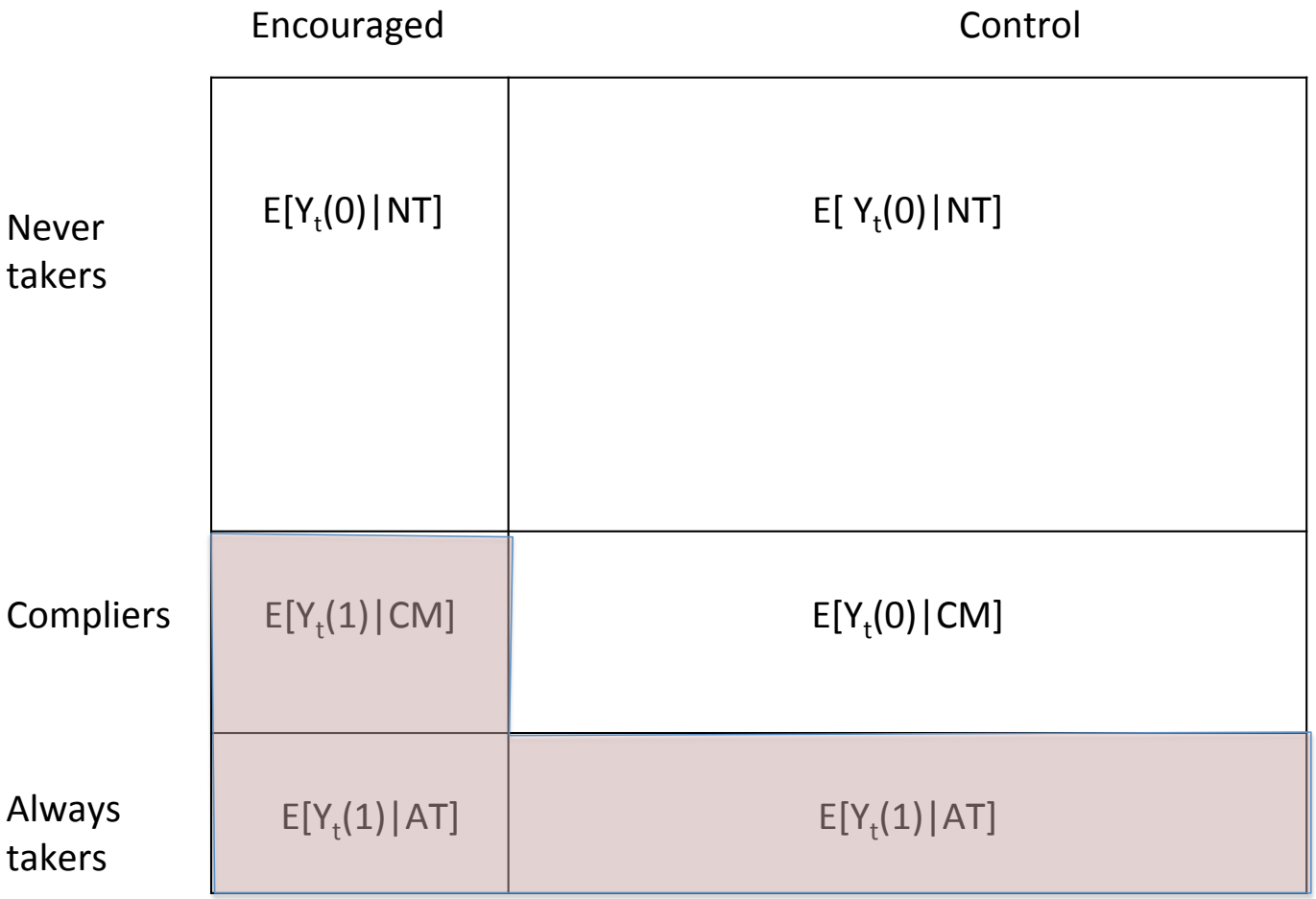
Our randomized encouragement design

Randomly assign 25 percent to an "encouragement treatment. N= 8,648	Simply observe the control group N=25,513

Our randomized encouragement design

	Encouraged N=8,648	Control N=25,513
Never takers	$E[Y_t(0)\|NT]$	$E[Y_t(0)\|NT]$
Compliers	$E[Y_t(1)\|CM]$	$E[Y_t(0)\|CM]$
Always takers	$E[Y_t(1)\|AT]$	$E[Y_t(1)\|AT]$

Our randomized encouragement design



Our encouragement intervention

- We issued an RFP to identify a firm specializing in grassroots mobilization and outreach.
- Fieldworks LLC has extensive experience with designing communication strategies and managing neighborhood canvassing operations.
- Michigan-based Fieldworks staff helped us develop a persuasive recruit-and-assist strategy, cut turf, hire local people from the community, support staff on the ground, and manage field operations.

Data

- Monthly energy consumption (natural gas and electricity) at all households: 2008-2014.
- Data tracking encouragement efforts.
- WAP application data on all households.
- Detailed efficiency audit data and work logs from all weatherized households.

Encouraging households to get weatherized

Panel A: Encouragement effort	
Encouraged group (households)	8,648
Initial home visits	6,694
Robo-calls	23,500
Personal calls	9,171
Follow up appointments	2,720
Average cost/hh	$55.00

Source: Fowlie, Greenstone, and Wolfram (2017)

Returns on encouragement effort

	Application	Efficiency audit	Weatherization complete
Base rate	0.02**	0.01**	0.01**
	(< 0.01)	(< 0.01)	(< 0.01)
Encouragement	0.13**	0.05**	0.05**
	(< 0.01)	(< 0.01)	(< 0.01)
Households	28,889	28,889	28,889

Notes: The unit of observation is a household.
** Significant at the 1 percent level

Source: Fowlie, Greenstone, and Wolfram (2017)

Energy efficiency is a tough sell...even when it's free!

In this context, we can safely rule out some of the standard explanations for the energy efficiency gap:

- Capital constraints
- Information costs/lack of information.
- Landlord-tenant split-incentive problems.

Results suggest that hard-to-measure "process" costs of pursuing energy efficiency improvements are large/prohibitive.

Treatment effect estimates - Total energy

	Total energy (log(MMBtu))		
	OLS-FE	IV-FE all	IV-FE gas only
WAP	-0.10**	-0.20*	-0.21**
	(0.01)	(0.08)	(0.08)
Hh-month FE	Y	Y	Y
Month-of-sample FE	Y	Y	Y
Households	27,990	27,229	26,054
Observations	1,662,781	1,653,583	1,528,526

** Significant at the 1 percent level
* Significant at the 5 percent level

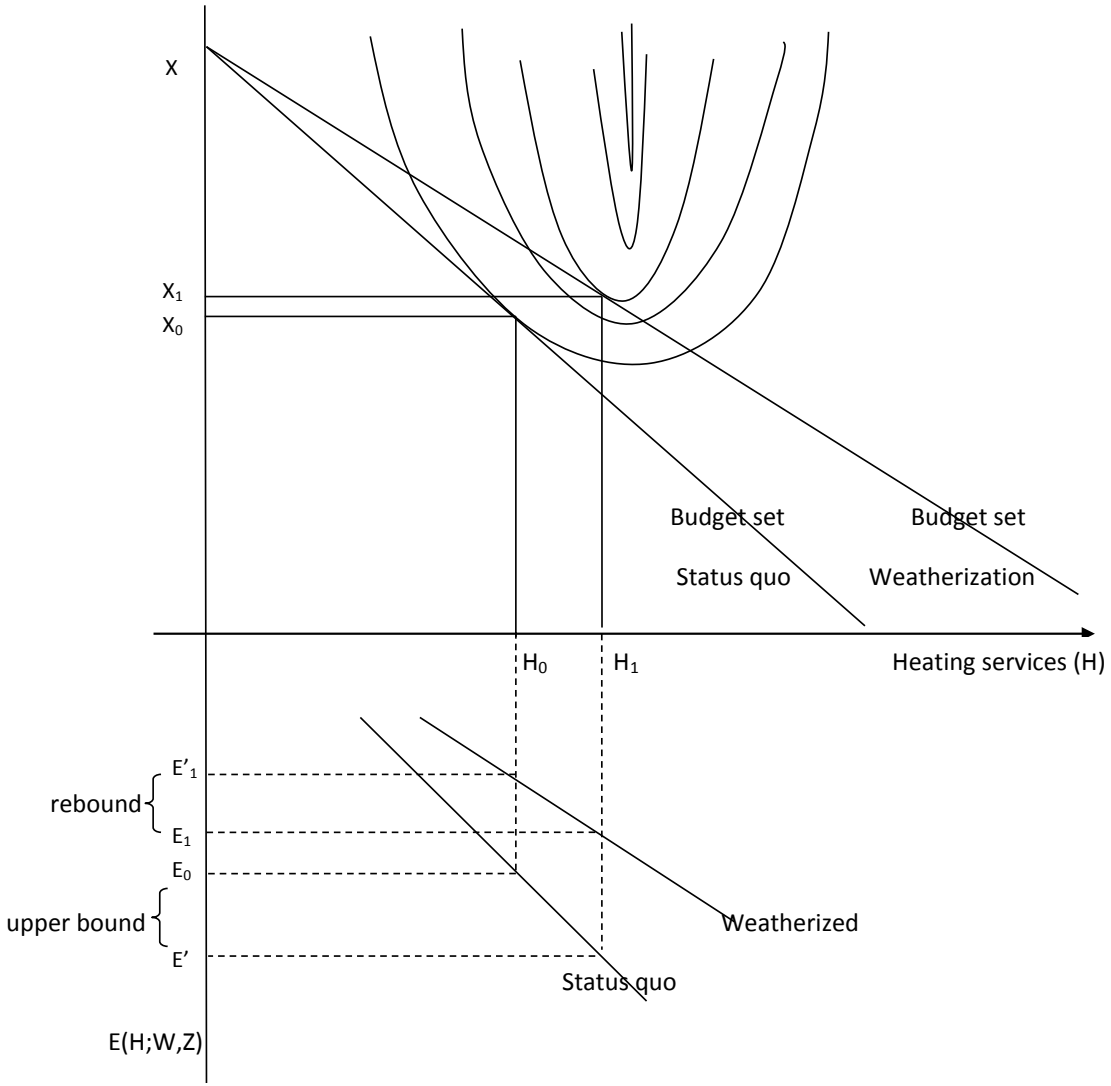
Source: Fowlie, Greenstone, and Wolfram (2017 WP)

We document a 'projection gap'

- We estimate economically significant energy savings: a 20 percent reduction in annual energy consumption on average.
- Our preferred experimental estimates amount to less than a third of projected savings.
- Prior work has attributed this projection gap to demand rebound (see, for example, Davis et al. 2014).
- Our theoretical framework guides estimation and valuation of this energy-demand rebound.

Returns on investment manifest in two ways:

Figure 1 : Utility maximization of a representative low-income consumer



Can rebound rationalize the projection gap?

- We estimate that the efficiency-induced increase in demand for heating explains less than 5 percent of the measurement gap.
- Indoor temperatures would need to increase by more than 26 degrees F to explain the gap between projected and realized energy savings!
- Upper bound on net welfare gain from 0.6 degree increase: approximately \$0.29 per winter month.
- Calibration error and overly optimistic assumptions about technology performance likely explain the measurement gap.

Returns on EE Investment

Time horizon	Ex ante (NEAT) projections (1)	Empirical estimates (2)
Panel A: Private internal rate of return		
10 years	7.0%	-10.5%
16 years	11.8%	-2.2%
20 years	12.8%	0.3%

Returns on EE Investment

Time horizon	Ex ante (NEAT) projections (1)	Empirical estimates (2)
Panel C: Social internal rate of return		
10 years	-1.0%	-20.0%
16 years	5.4%	-9.5%
20 years	7.0%	-6.1%

** Implies an abatement cost in excess of \$200 per ton CO_2.**

Conclusions

- First stage result comes as a surprise to some: energy efficiency is a tough sell, even when it is 'free'.
- Although realized energy savings are economically significant, they fall short of ex ante estimates of energy savings.
- Using a revealed preference-based approach, estimate welfare gains from rebound to be small.
- Our findings underscore the importance of field-testing of projected returns on energy efficiency investments.

Strengths and limitations of the RED

Advantages:

- Generates random variation in treatment status when mandatory assignment to treatment is out of the question.
- Supports the estimation of highly credible estimates of causal impacts without meddling with the program design/implementation.

Limitations?

Strengths and limitations of the RED

Advantages:

- Generates random variation in treatment status when mandatory assignment to treatment is out of the question.
- Supports the estimation of highly credible estimates of causal impacts without meddling with the program design/implementation.

Limitations?

- The LATE may differ from the ATE if treatment effects vary systematically in the population.
- Relies on some additional assumptions to achieve identification.
- Statistical power reduced vis a vis an RCT.

Concerns? Considerations?

- Spillovers?
- Valid exclusion restriction?
- Hawthorne effects?
- Other?

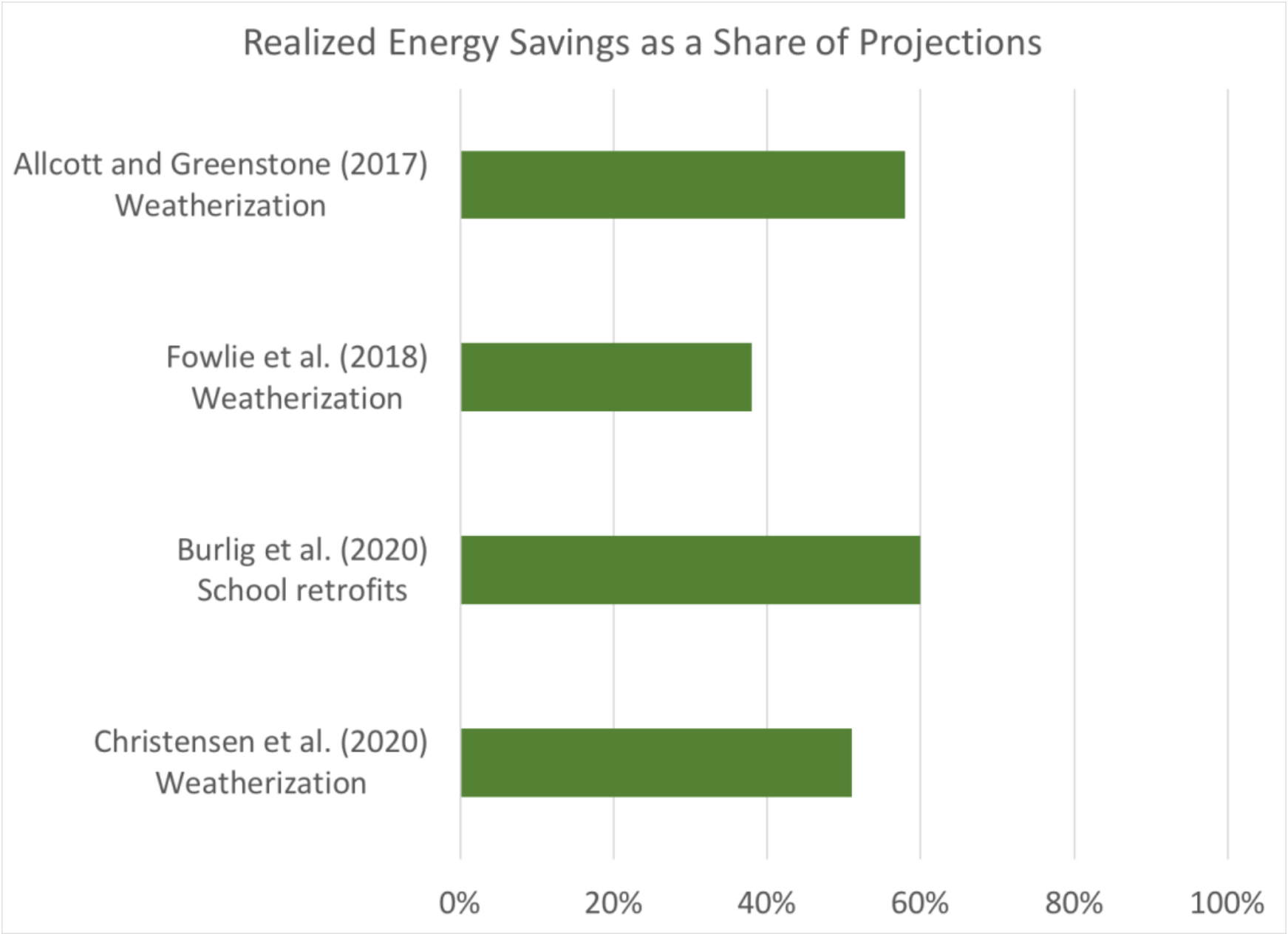
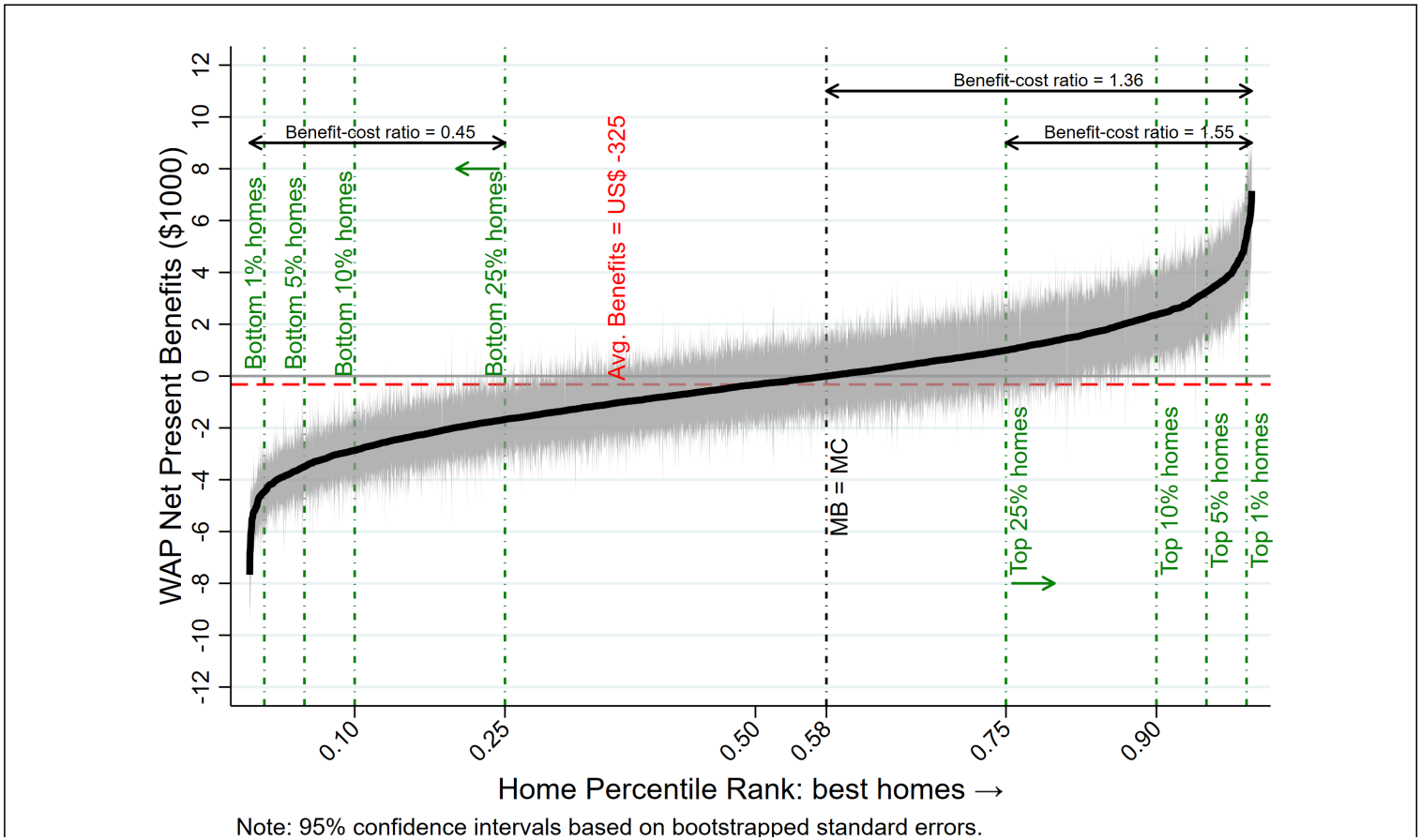

Realized Energy Savings as a Share of Projections
Allcott and Greenstone (2017)
Weatherization
Fowlie et al. (2018)
Weatherization
Burlig et al. (2020)
School retrofits
Christensen et al. (2020)
Weatherization
0%
20%
40%
60%
80%
100%

Figure 5: Ranking of Homes by Net Present Benefits



Decomposing the Wedge Between Projected and Realized Returns in Energy Efficiency Programs (Christensen et al. 2021)

Lecture Game Plan

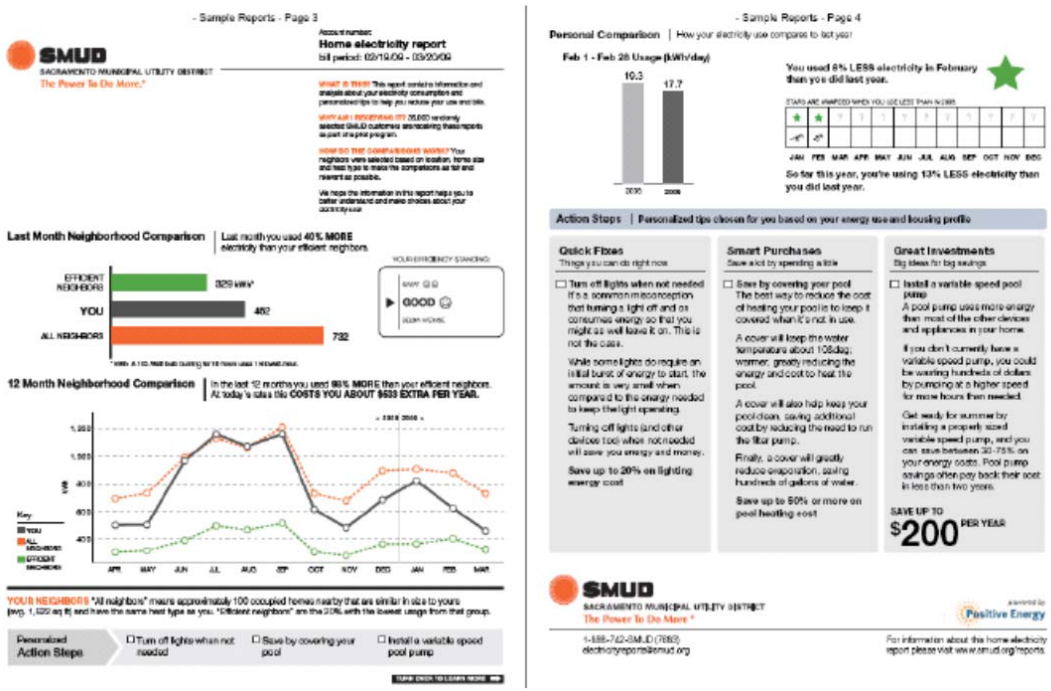
- Complication 1: Hawthorne effects
  - Gosnell, List, and Metcalfe (2019)
- Complication 2: Imperfect Compliance
  - Fowlie, Greenstone, Wolfram (2018)
- Treatment effect heterogeneity
  - Knittel and Stolper (2021)

Heterogeneous treatment effects?

- In many applications it will be interesting to look beyond the average in order to understand how the causal effects vary in the population.
- Standard approach: include interaction terms between covariates and the treatment indicator in you regression equation.
- Problem?: To see which variables predict heterogeneous treatment effects, we have to include many interaction terms.
- It's hard to know along which dimensions to look for interesting treatment effect heterogeneity ... cue machine learning methods.

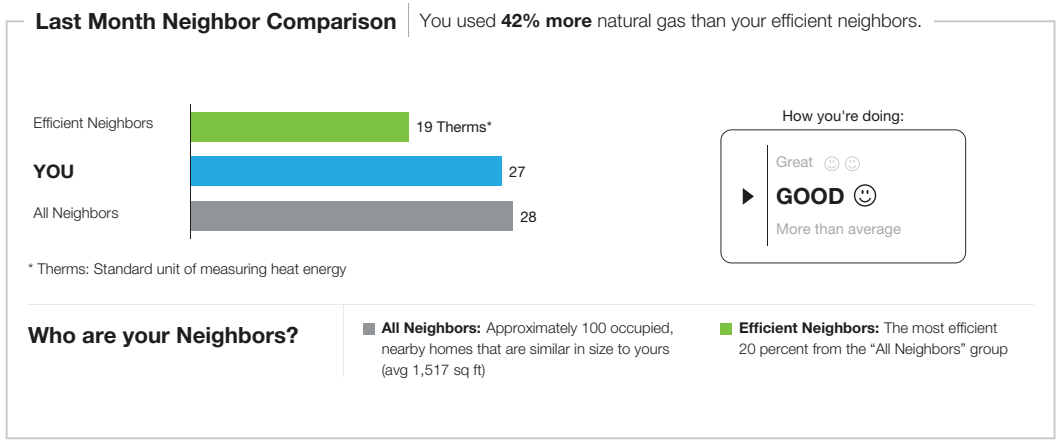
Application: A ubiquitous nudge

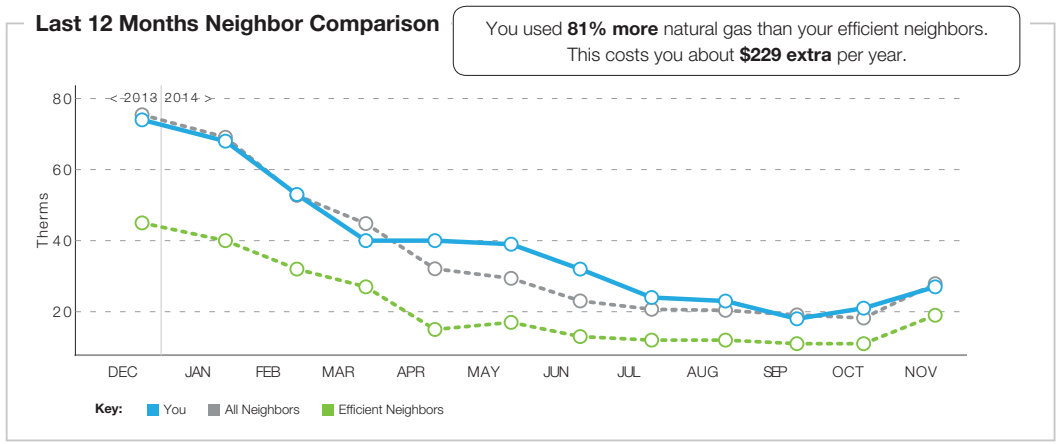
The Opower Home Energy Report



HERs are the most frequently studied nudges in EEE.

Empirical context: Home Energy Reports





Empirical context: Home Energy Reports

Personalized tips | For a complete list of energy saving investments and smart purchases, visit **utilityco.com/rebates**.

Quick Fix
Something you can do right now

☐ **Open your shades on winter days**

Taking advantage of winter's direct sunlight can make a dent in your heating costs. Open blinds and other window treatments during the day to capture free heat and light.

South-facing windows have the most potential for heat gain, and the sun is most intense from 9 a.m. to 3 p.m.

When you let the sun in, remember to lower the thermostat by a few degrees. These two steps combined are what save money and energy.

SAVE UP TO
$10 PER YEAR

Smart Purchase
An affordable way to save more

☐ **Program your thermostat**

A programmable thermostat can automatically adjust your heat or air conditioning when you're away, then return to your preferred temperature when you're home to enjoy it.

If you don't already have a programmable thermostat, look for one at your local home improvement store. For comfort and convenience, be sure to program your thermostat with energy-efficient settings.

If you need help installing or programming your thermostat, consult your manual or call the manufacturer for assistance.

SAVE UP TO
$65 PER YEAR

Smart Purchase
An affordable way to save more

☐ **Weatherstrip windows and doors**

Windows and doors can be responsible for up to 25% of heat loss in winter for a typical home.

If you're comfortable doing the task yourself, you can weatherize your home in just a few hours. Seal windows for about $1 each with rope caulk, or install more permanent weatherstripping for $8-$10 per window. Also, install sweeps at the bottom of exterior doors.

A professional can help you with this work if you prefer.

SAVE UP TO
$10 PER YEAR

Home Energy Reports: Implementation

- Electric/gas utilities contract with Opower (or other provider) to send HERs to residential consumers
- HERs have both informational and persuasive components
- Typical program: Send HERs repeatedly over several years
- Each HER has similar structure but can deliver different information:
  - Households often change rank relative to neighbors
  - Different conservation tips

Knittel and Stolper explore heterogeneity in these impacts

- Machine-learning (ML) methods can offer a disciplined way to search non-parametrically for heterogeneity in data rich contexts.
- ML methods are increasingly applied to causal inference problems.
- These authors use a 'causal forest' algorithm to investigate heterogeneity in responses to home energy reports.
- They also want to assess the potential efficiency gains associated with targeting treatments to maximize impacts.

Empirical Context

- Opower has run 26 waves (denoted w) of home energy report experiments in Eversource territory (2011-2017).
- This experimentation generates 902,581 households and 49,491,297 hh-month observations.
- Experian data used to measure household income, demographics, home value etc.
- Use household-monthly panel data on electricity consumption to estimate ATEs :

$$\text{kwh}_{it} = \alpha_1 + \alpha_2 T_{it} + X_i\eta + \theta_i + \omega_t + e_{iwt}$$

A Random Forest?

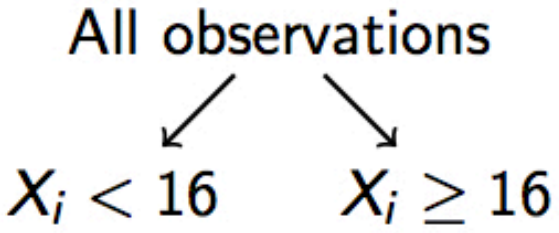
As a point of departure, think about predicting subgroups of X that have different average outcomes?

- At each node of our tree, we split the observations such that the resulting groups are as different from each other as possible, but the members of each resulting subgroup are as similar to each other as possible (minimize MSE).
- For example, look for the binary split that gives us the maximum improvement in MSE (often requiring that the improvement increment must clear a minimum threshold).
- In a random forest, we end up with 'trees' that are trained on different subsets of data ('bagging') and use different covariates to make splitting decisions (to avoid state dependence).
- Average prediction across trees to make predictions.

Growing a Tree to Predict Outcomes

$$\text{MSE}_0 = \frac{1}{n}\sum(Y_i - \bar{Y})^2$$

$$\text{MSE}_1 = \frac{1}{n}\sum(Y_i - \bar{Y}_{j:x_j \in \ell(x_i|\Pi_1)})^2$$



Partition $\Pi \in \mathbb{P}$ $\longrightarrow$ $\left\{\ell_1 = \{x_i : x_i < 16\},\ \ell_2 = \{x_i : x_i \geq 16\}\right\}$

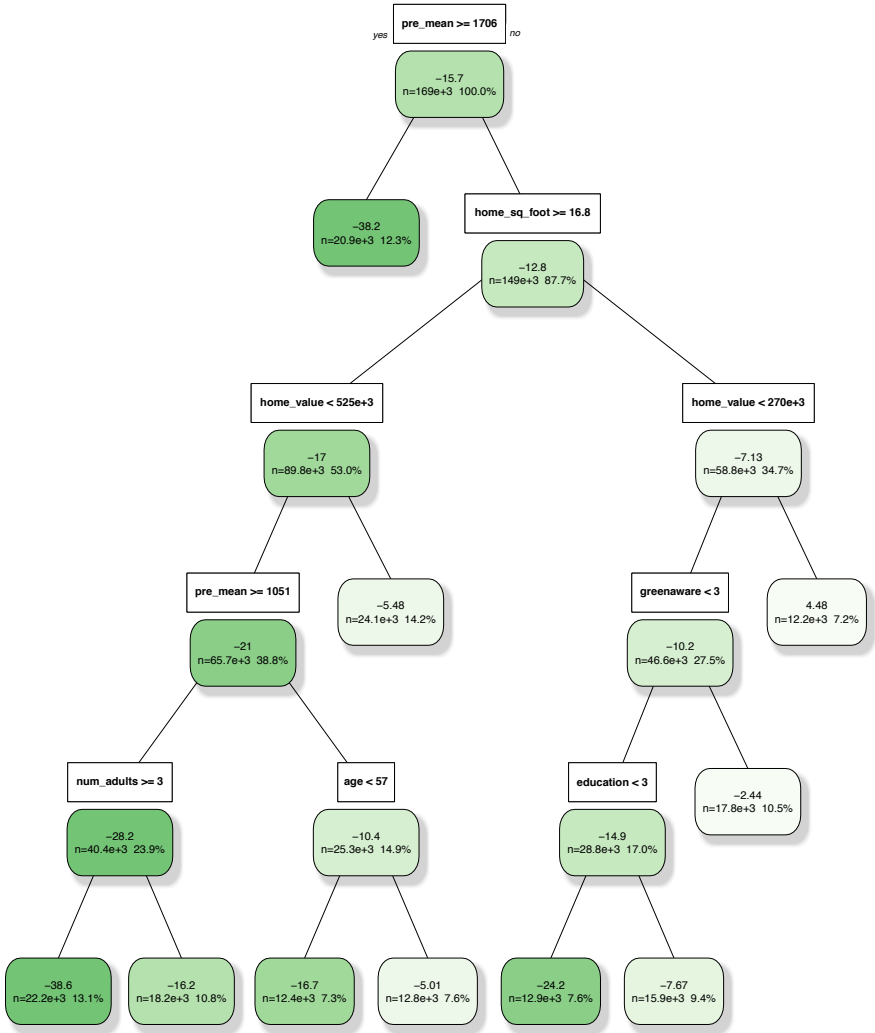
Leaves

Prediction rule for new x:

$$\hat{\mu}(x) = \bar{Y}_{j:x_j \in \ell(x_i|\Pi)}$$

Random versus causal forest? Splitting criterion

- In Random Forests, the split at each tree node is often performed by minimizing the mean squared error of the outcome variable Y.
- In Causal Forests, we cannot observe treatment effects for individuals!
- The prediction of a treatment effect is given by the difference in the average outcomes Y between the treated and the untreated observations in a leaf
- Splitting criteria searches for a partitioning that maximizes differences in treatment effects across nodes 'subject to penalties for within-node variance in ATEs and treatment-control imbalance' (??)).



Notes: The tree is constructed from the Connecticut "base" wave beginning in April 2014. The dependent variable is the difference between average monthly electricity usage in program year 2 and the year prior to program start. Reported numbers in each box are leaf-specific ATE (in kWh), the number (n) of households falling into this leaf, and the corresponding proportion (in %) of total households used.

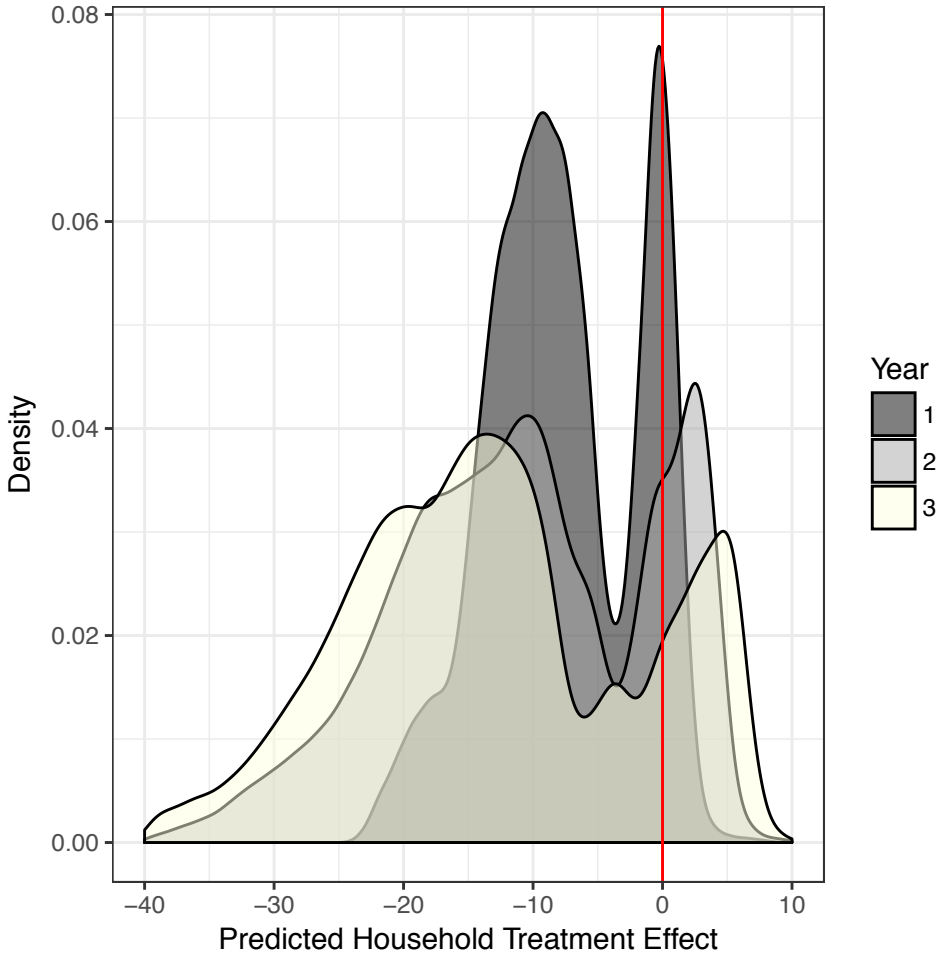
Honest trees?

- Random Forests assess accuracy of the predictions using an error measure such as the mean squared error.
- But we cannot ground-truth causal forests in this way as we as we do not observe individual treatment effects.
- Use a training sample to grow our causal tree. Then project each unit in an estimating subsample onto a terminal leaf.
- Estimate the treatment effects within each leaf by taking the difference between the mean of the treatment and the mean of the control cases in the estimating sub-sample.

Weighting function?

- After growing many (10,000) trees, each household associated with many predicted treatment effects (5,000 in expectation)
- Instead of averaging over the estimates, Generalized Random Forests uses a weighted set of nearby training examples.
- The 'adaptive neighborhood estimator' assigns weights to neighbors based on the number of times you land in the same leaf.
- Use the weighted average of neighbor treatment effects.

Figure 7: Distribution of Predicted Treatment Effects



Notes: Each plotted distribution is a kernel density of household treatment effects in a specific year (1, 2, or 3) of HER programming. Treatment effect predictions come from our causal forest (Section 2.2).

Some conclusions

- Paper demonstrates how ML methods can shed light on questions about treatment effect heterogeneity and targeting.
- Pre-treatment consumption and home value are the strongest predictors of individual responses, but other characteristics have predictive power as well.
- Authors suggest that machine learning could be used to improve the effectiveness of interventions via targeting – but don't incorporate parameter uncertainty into this simulation exercise.