SMART GRID INVESTMENT GRANT
CONSUMER BEHAVIOR STUDY ANALYSIS

Go for the Silver?

"Gold Standard" RCT Versus Quasi-Experimental Methods

Patrick Baylis[†], Peter Cappers[*], Ling Jin[*], C. Anna Spurlock[*], Annika Todd[*]

[*] Lawrence Berkeley National Lab
[†] University of California, Berkeley

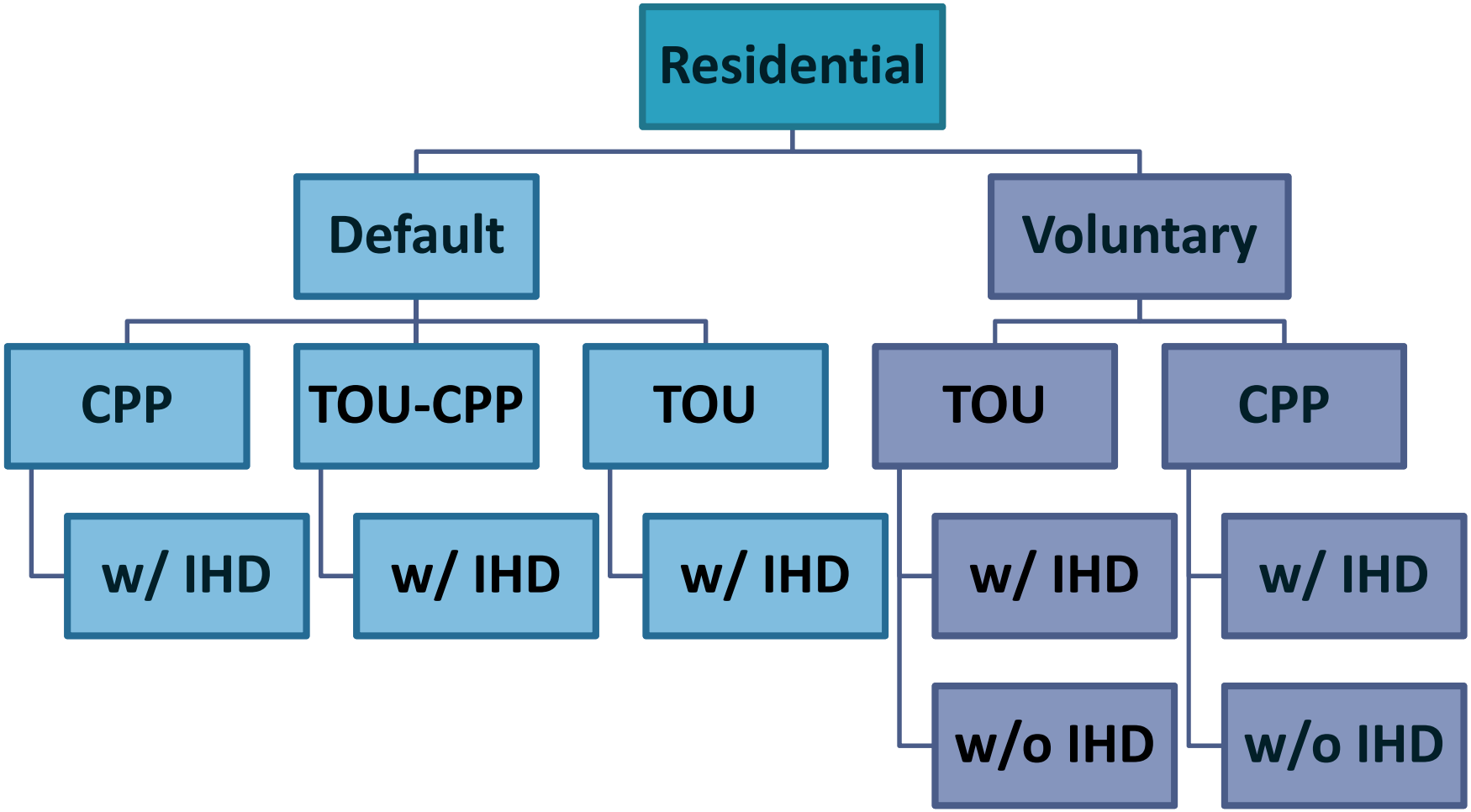
Motivation

- **Randomized controlled trials (RCTs) are widely viewed as the "gold standard" of evaluation.**
  - Widely used in applied research fields such as health and public policy
  - Requires forethought and planning in program implementation, and may not always be possible
- **Evaluations of DR and energy pricing programs have largely been conducted through non-RCT ("quasi-experimental") methods.**
  - Can be applied after a program has taken effect

Smart Grid Investment Grant – Consumer Behavior Study: Sacramento Municipal Utilities District (SMUD)

- **Types of Time-based pricing tested:**
  - CPP
  - TOU
  - CPP-TOU
- **Recruitment strategies used**
  - Voluntary (Opt-in)
  - Default (Opt-out)
- **Technology tested**
  - IHD
- **Treatment period**
  - June 1st – September 31st, 2012
  - June 1st – September 31st, 2013
- **We evaluate seven of the treatment arms from SMUD's pilot**
- **We focused on electricity use during Critical Peak Pricing (CPP) event days**

SMUD Experimental Design



Data from SMUD

- **Large Sample**
  - Because of the very large sample size (particularly of the control group) in SMUD's polite, we are able to estimate treatment effects using a variety of evaluation methods

Group	Customers
CPP opt-in	9,190
CPP opt-in no IHD	1,214
TOU opt-in	12,735
TOU opt-in no IHD	7,632
CPP opt-out	846
TOU opt-out	2,407
CPP-TOU opt-out	727
Control	45,839

Data from SMUD

- **Hourly electricity consumption**
  - In kWh
  - Pre-treatment (June 1st, 2011 – May 31st, 2012)
  - During Treatment (June 1st, 2012 – September 31st, 2013)
  - For every household
    - Randomized into treatment
      - Treated
      - Not Treated (opted out or didn't opt in)
    - Randomized into control
- **Hourly weather data**
  - Dry and wet bulb temp
  - humidity

Overview

Compared RCT to non-RCT evaluation methods for:

- **Overall program impact evaluation**
  - Estimating peak period energy savings on average
- **Baseline methods used to generate household-specific savings on event days individually**
  - RCT can't be used to generate household-specific savings
  - RCT can be compared to aggregate savings estimates when baseline used

Evaluation & Baseline "Gold Standard" Method

- **Randomized Controlled Trial (RCT)**
  - Randomized Encouragement Design (RED)
    - Households need to be assigned randomly between treatment and control groups, but "compliance" can be imperfect (i.e., people can chose not to opt in or can opt out).
  - Uses Difference-in-differences instrumental variables (IV) regression.

Non-RCT Evaluation Methods

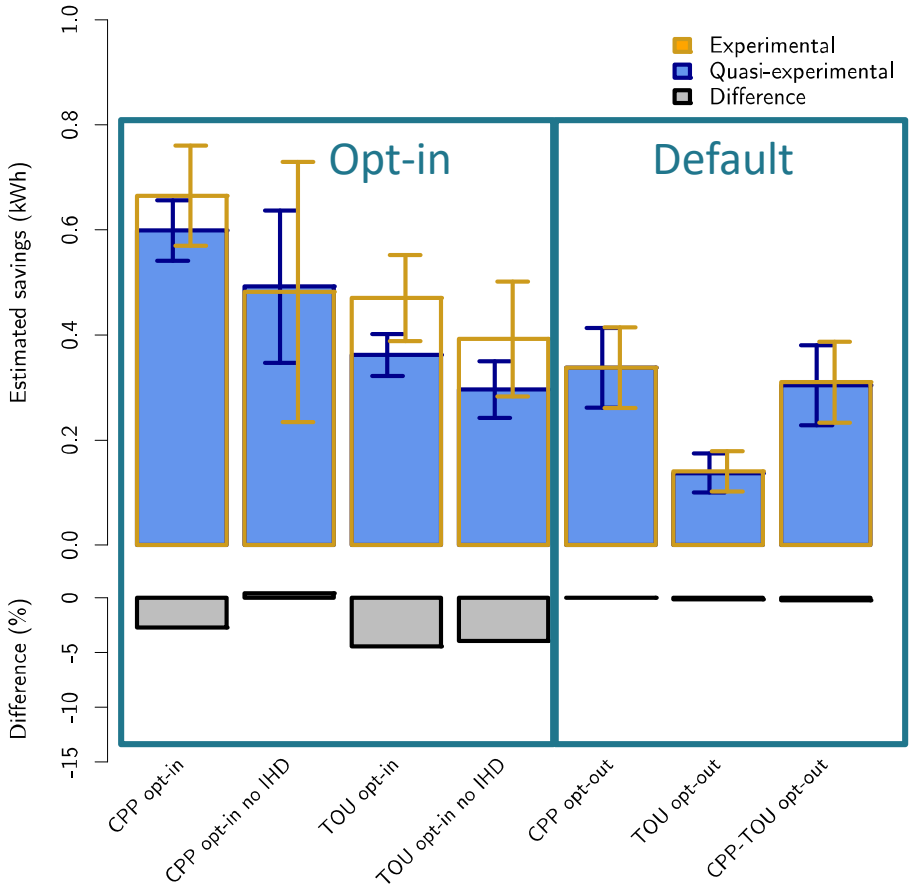
- **Difference-in-differences (DID)**
  - Compare difference between pre-treatment and treatment usage of those who actually end up in treatment to a randomized control group "held out" (i.e., not offered treatment)
- **Propensity Score Matching (P-score)**
  - Construct estimates of each customer's enrollment likelihood (propensity score) based on their pre-treatment usage
  - estimate a regression using treated and control households that accounts for this likelihood by using weights based on this propensity score.
- **Regression Discontinuity (RD)**
  - Not discussed in this presentation.

Non-RCT Baseline Methods

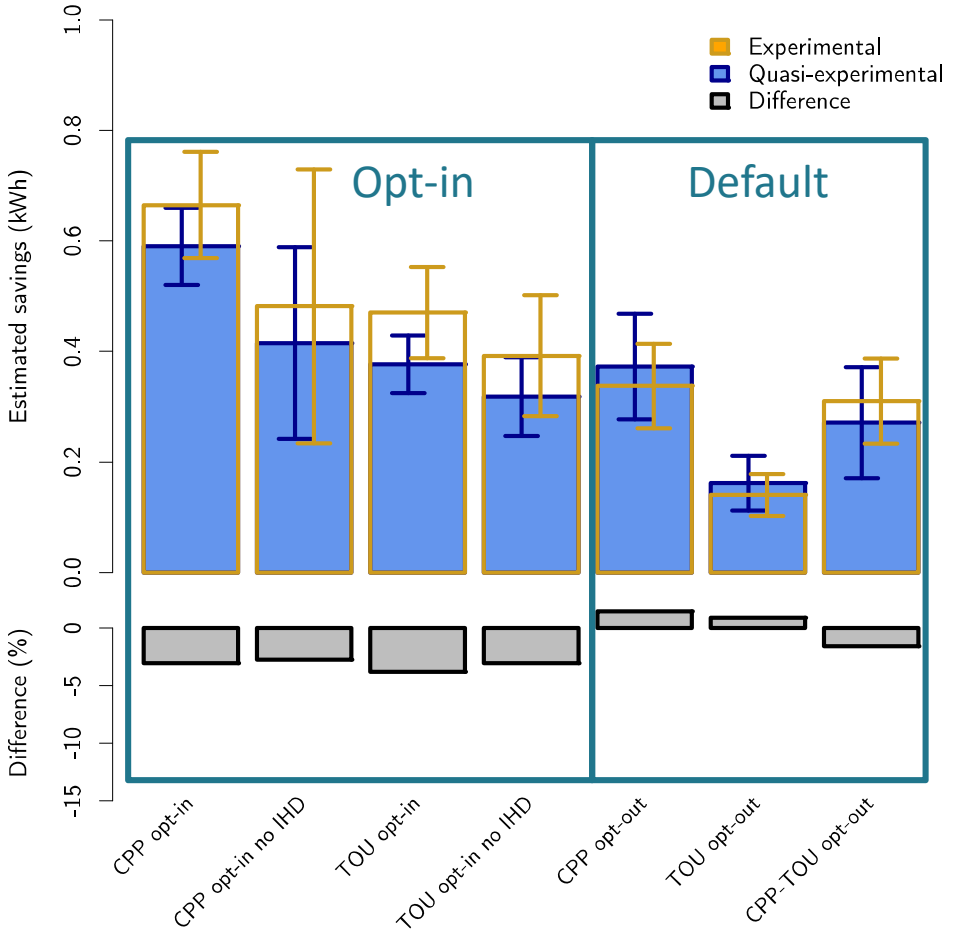
- **Four-in-five day baseline**
  - Without Adjustment
    - Average consumption from four highest consumption days in the the last 5 non-event business days
  - With Additive Adjustment
    - Same as above, but includes an adjustment to control for weather and underlying usage pattern differences across days based on baseline and event-day off-peak usage
    - Found by KEMA (2011) to be the preferred method based on several metrics of accuracy, consistency, etc. among a variety tested
- **Individual customer regression baseline**
  - Uses non-event hours across the treatment period for each customer, controlling for temperature, as a baseline for event-day peak period consumption for that customer

Evaluation Results: DID & P-Score

DID



Propensity Score



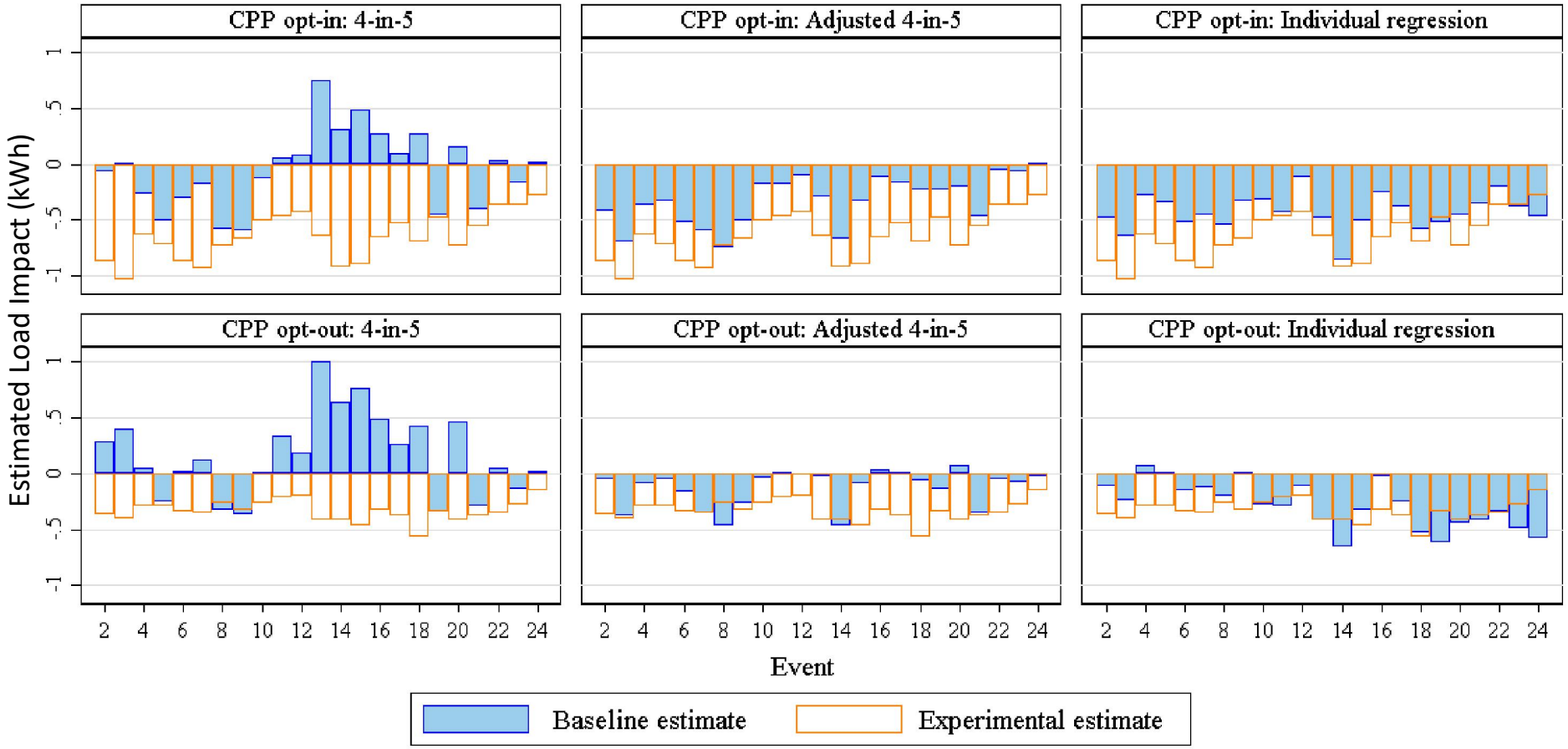
Evaluation Results

- **DID & P-Score tend to underestimate treatment effect, particularly for opt-in treatments**
  - Bias as much as 5 percentage points (i.e., estimated effect would have been 15% of average peak period consumption when effect was actually 20%)
  - Average Absolute Difference: DID (1.7%), P-score (2.4%)

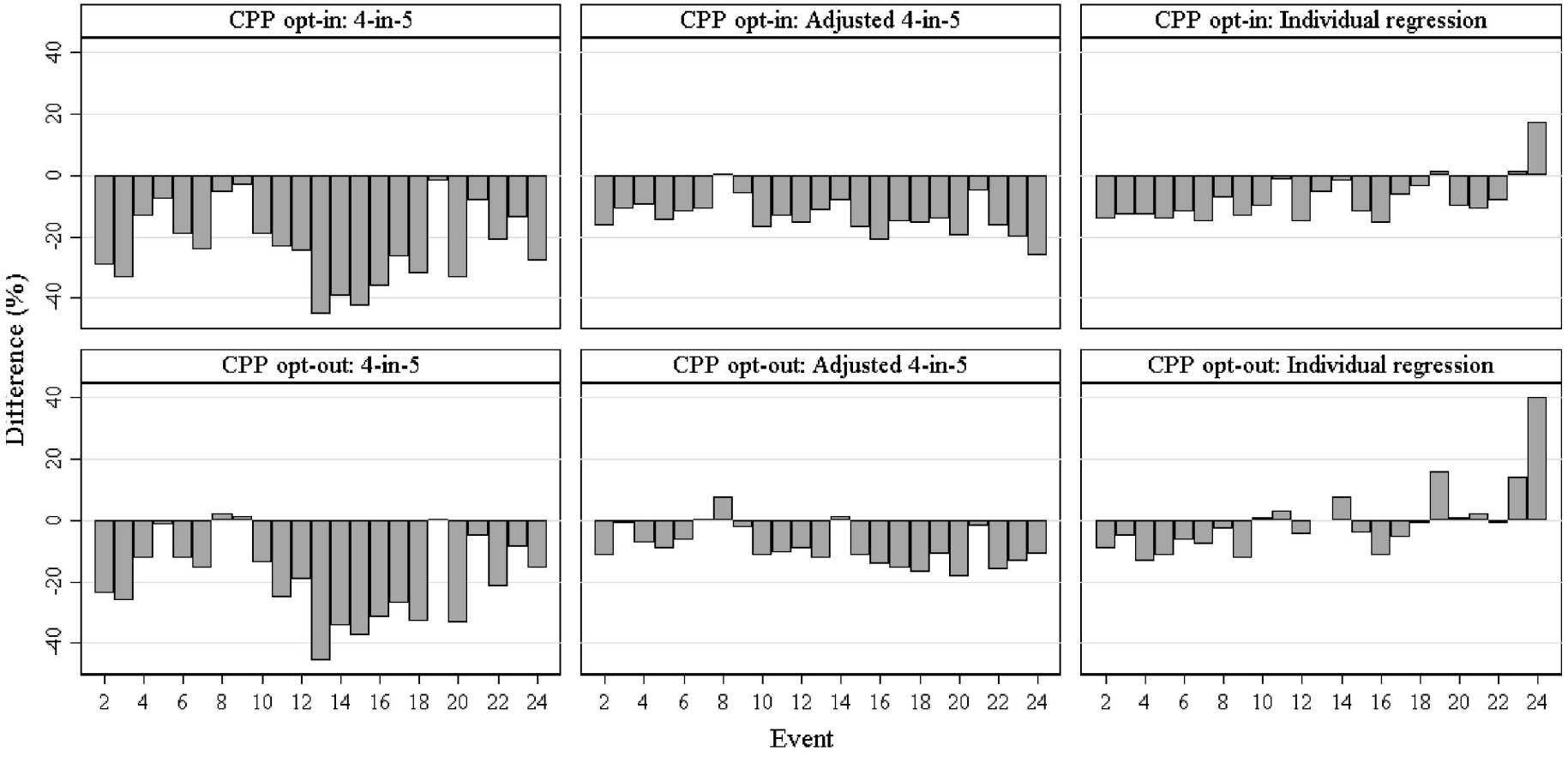
Evaluation Results

- **Biases are more pronounced in opt-in vs opt-out designs**
  - Highly suggestive of the role of selection effect in generating the bias
  - Opt-in achieved at most 20% enrollment, default was 90% enrollment
  - This means default treatment group resembles control group to a much greater extent than the opt-in treatment group.

Baseline Results: Estimate Comparison



Baseline Results: Difference



Baseline Results

- **Savings based on all non-RCT baseline methods underestimated treatment effect**
  - With differences as much as 20 percentage points in some cases (meaning a true effect of 25% of peak consumption would have been estimated at only 5%)
  - Baseline with additive adjustment did better than without adjustment
  - Individual customer regression did slightly better on average
- **Results strongly suggest the role of spillover effects in causing bias**
  - Customers change behavior on non-event hours and non-event days as a result of the rate because of habitual behavior change, reprograming equipment (e.g., thermostats), or new equipment investments.
  - These changes are caused by the rate and cause consumption in the "baseline" hours to be lower.
  - Therefore a comparison of event consumption to baseline consumption underestimates the true effect of the rate.

Conclusions

- **Evaluation Methods**
  - We find evidence of systematic bias when evaluation is conducted using commonly used evaluation methods (DID and P-score). DID relative to selected-out group does worse even.
  - The bias, given this program, was within 5% of average peak period consumption.
  - Opt-in programs exhibit bias while default programs do not to the same extent.
  - Likely cause of bias is selection into treatment
- **Baselines**
  - We find evidence of systematic bias when using commonly employed baseline methods
  - Bias was frequently between 10-20% of average peak period consumption when 4-in-5 method with additive adjustment was used.
  - Likely cause of bias is spillover effects
- **Further work**
  - We are expanding the comparison of evaluation methods to RCT to other SGIG-CBS utilities.
  - We are more carefully documenting and exploring the role of spillover effects in the bias shown with baseline methods and providing thoughts regarding baseline methods that are less likely to suffer from this type of bias.

SMART GRID INVESTMENT GRANT

CONSUMER BEHAVIOR STUDY ANALYSIS

Contact:

Anna Spurlock
caspurlock@lbl.gov

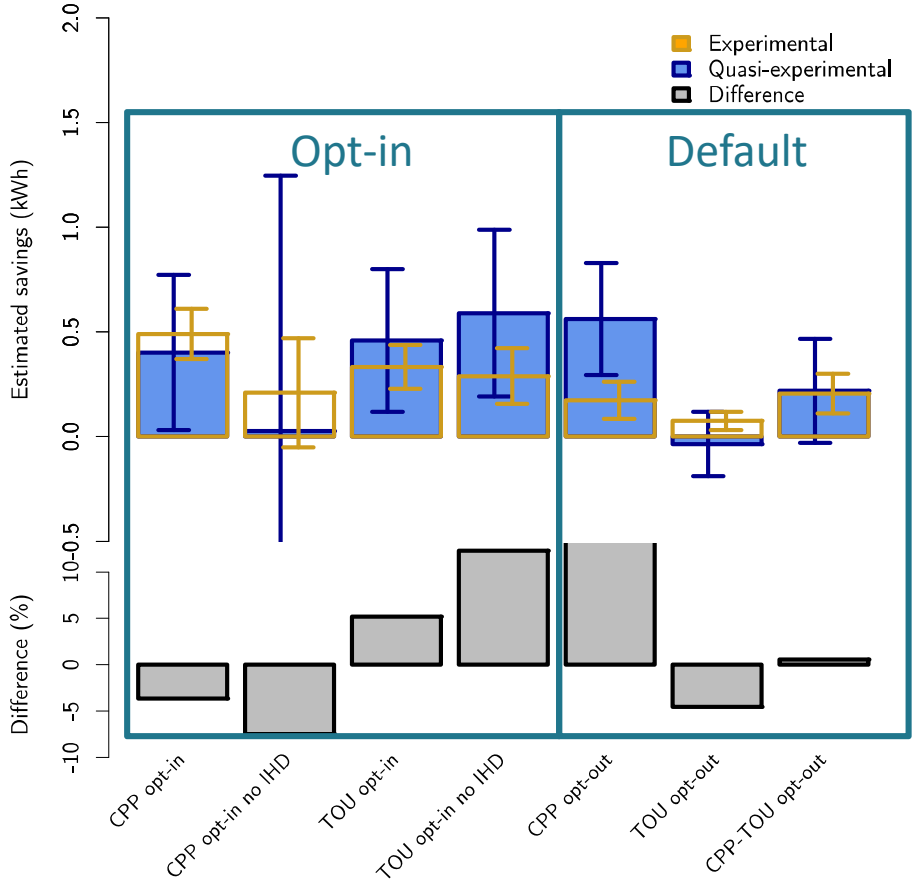
Peter Cappers
pacappers@lbl.gov

Annika Todd
atodd@lbl.gov

Thank you!

Evaluation Results: RD

RD (30th Percentile Cut-off)



RD (50th Percentile Cut-off)