machine learning conference

Measuring Long-Term Effects using Dialed Up Experiments

Chris Khawand
@khawandc
Sr. Economist | CMA

Miikka Rokkanen
@miikka
Economist | CMA

Matt Gentzkow
@mgentzko
Contractor | Amazon Search

Background and Motivation

- Long-term effects of long-term exposure are a fundamental question for launch decisions
- Typically experiments at Amazon run for only a short period of time
- We develop a new method for measuring long-term effects of long-term exposure using dialed up short-term experiments
- The method allows for arbitrary treatment effect dynamics in exposure time by restricting treatment effects to be stable in calendar time
- Our end goal is to build a tool that lets experiment owners monitor treatment effects after dial-up and potentially overturn the launch decision

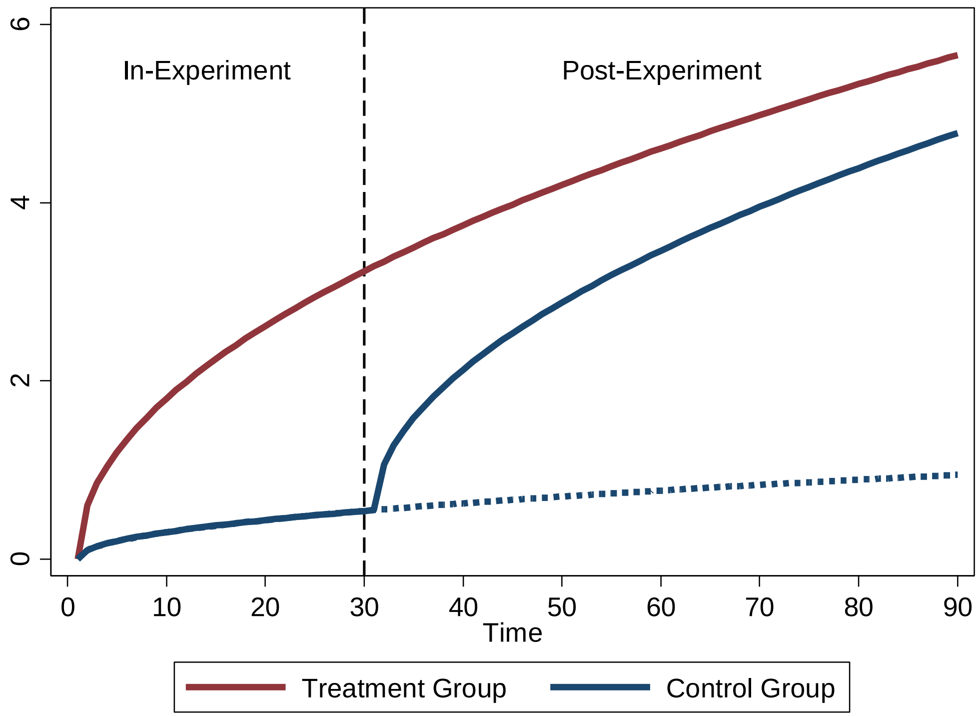
Identification and Estimation (Simplified Example)

- Experiment runs during days 1-30:
  - 50% of customers assigned to $T1$
  - 50% of customers assigned to C

- Experiment dialed up to 100% on day 31

- Customers trigger their assigned experience immediately on the day of assignment
  - $T1$ triggers treatment experience on day 1
  - C triggers treatment experience on day 31

- Variation in trigger timing handled in the paper

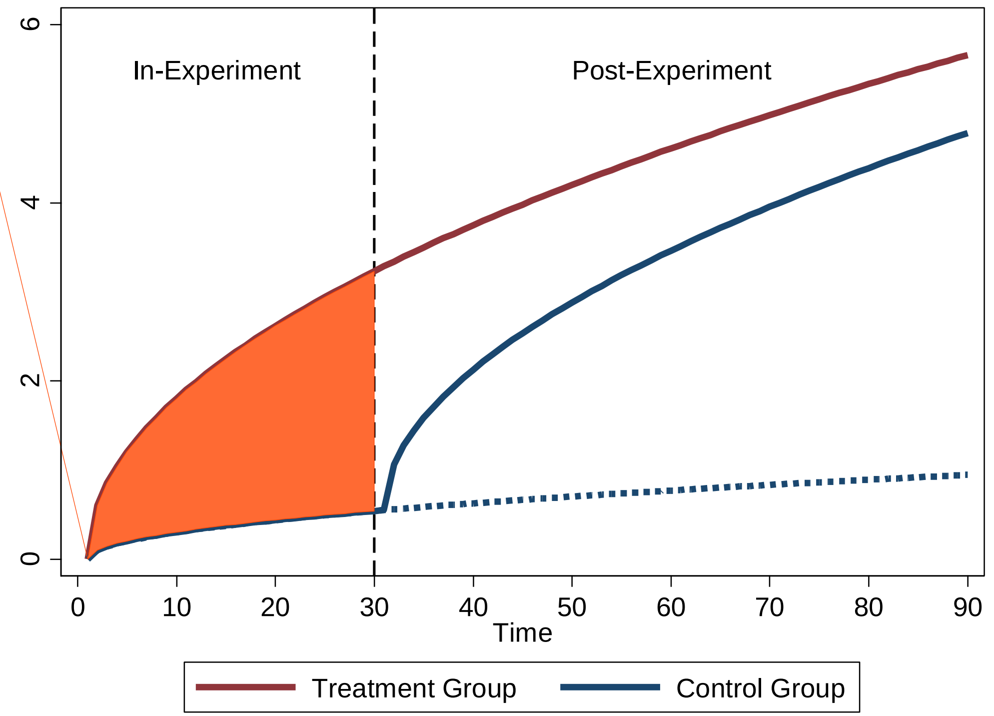


Identification and Estimation (Simplified Example)

- Days 1-30: experiment running
  - $T1$ exposed to treatment
  - C not exposed to treatment

- $T1 - C$ differences in average outcomes measure ATEs for 1-30 days of exposure

$$\bar{Y}_t^{T1} - \bar{Y}_t^{C} = ATE_t^{T1}, t = 1, \dots, 30$$



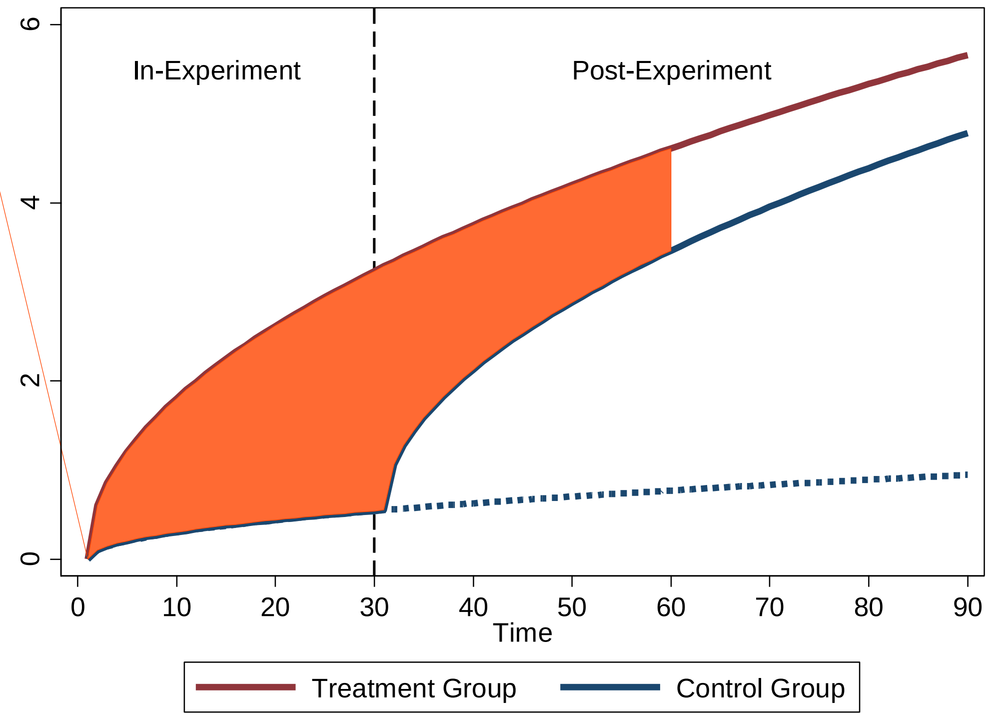
Identification and Estimation (Simplified Example)

- Days 31-60: experiment dialed up
  - $T1$ continues to be exposed to treatment
  - C starts being exposed to treatment

- $T1 - C$ differences in average outcomes no longer measure ATEs for 31-60 days of exposure

$$\bar{Y}_t^{T1} - \bar{Y}_t^{C} = ATE_t^{T1} - ATE_{t-30}^{C}, t = 31, \ldots, 60$$

- Need to uncover average outcomes of C had they continued to not be exposed to treatment



Identification and Estimation (Simplified Example)

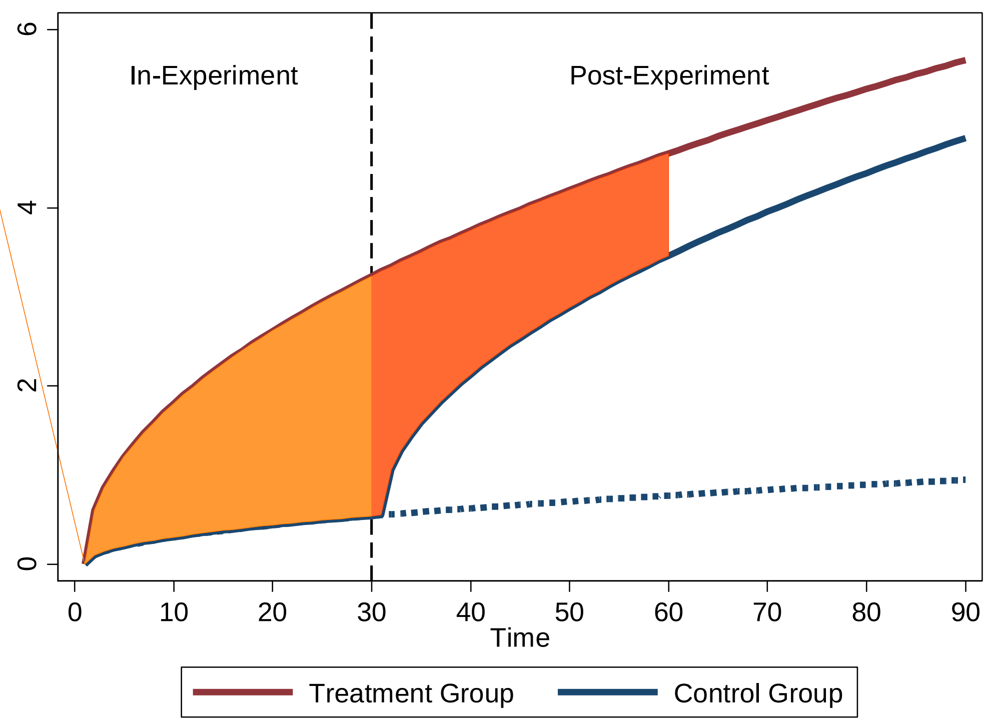
- Assume that ATEs are stable in calendar time

$$ATE_t^{T1} = ATE_t^{C} \text{ for all } t$$

- Correct average outcomes of C by subtracting out ATE_{t-30}

$$\bar{Y}_t^{T1} - (\bar{Y}_t^{C} - ATE_{t-30}^{T1}) = ATE_t^{T1}, t = 31, \ldots, 60$$

- Allows us to continue measuring ATEs for 31-60 of exposure



Identification and Estimation (Simplified Example)

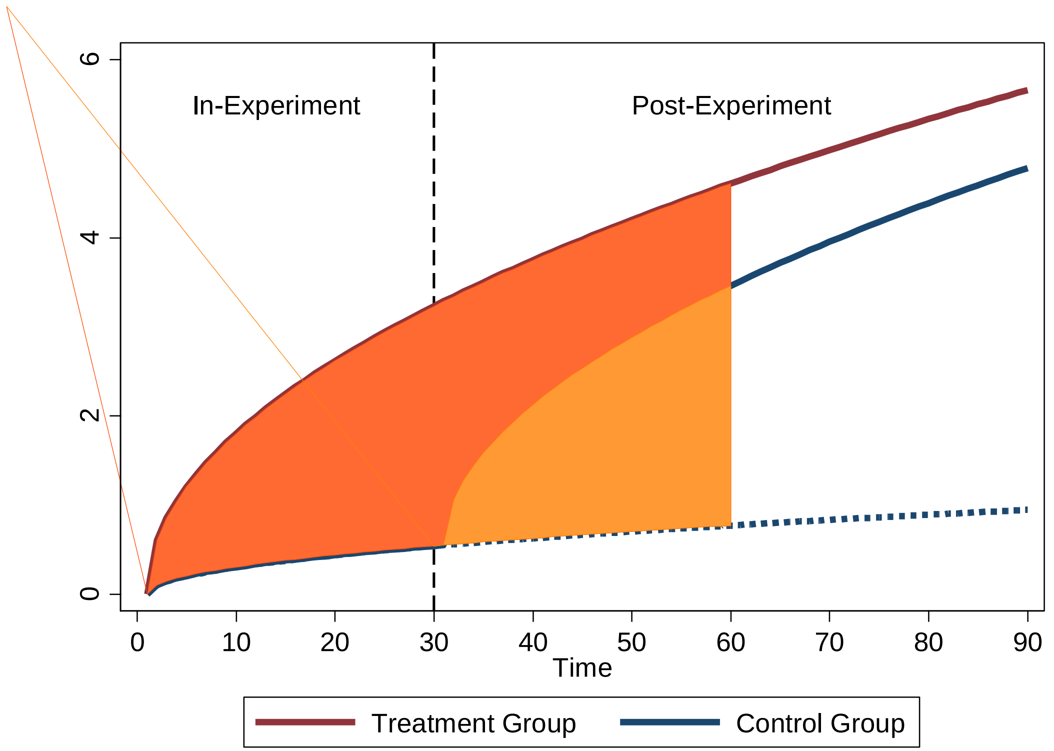
- Assume that ATEs are stable in calendar time

$$ATE_t^{T1} = ATE_t^{C} \text{ for all } t$$

- Correct average outcomes of C by subtracting out ATE_{t-30}

$$\bar{Y}_t^{T1} - (\bar{Y}_t^{C} - ATE_{t-30}^{T1}) = ATE_t^{T1}, t = 31, \dots, 60$$

- Allows us to continue measuring ATEs for 31-60 of exposure



Identification and Estimation (Simplified Example)

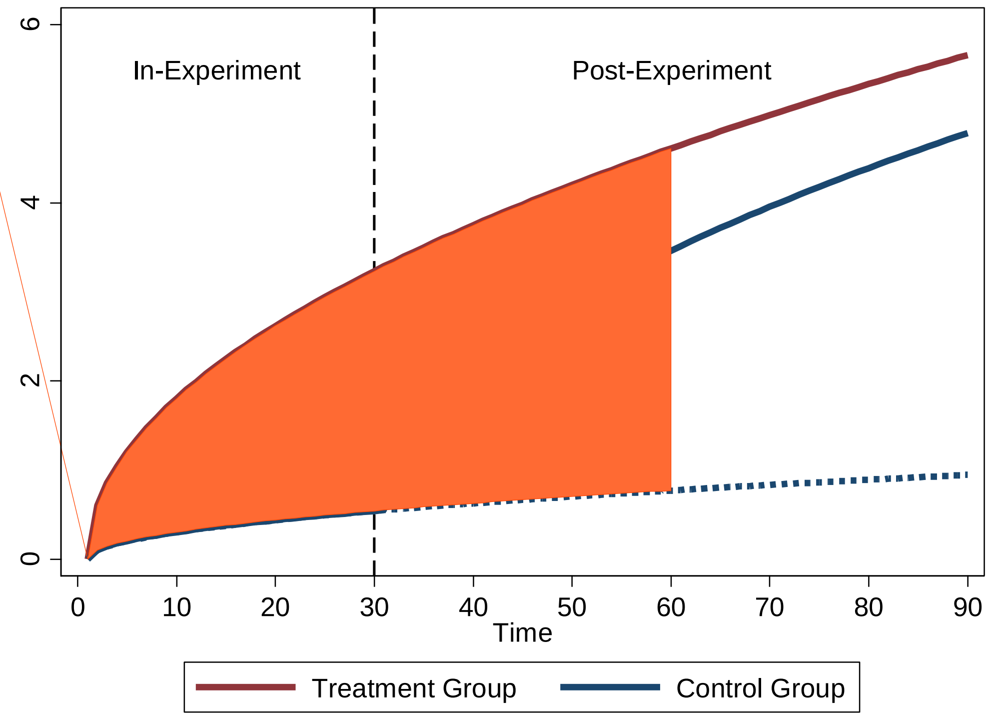
- Assume that ATEs are stable in calendar time

$$ATE_t^{T1} = ATE_t^C \text{ for all } t$$

- Correct average outcomes of C by subtracting out ATE_{t-30}

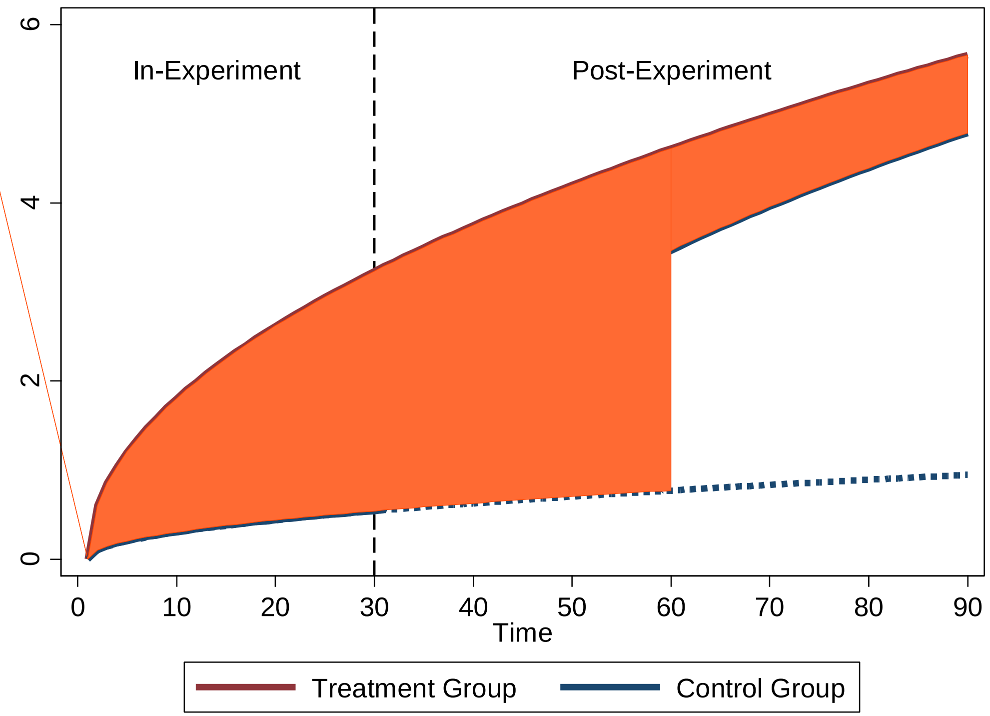
$$\bar{Y}_t^{T1} - (\bar{Y}_t^C - ATE_{t-30}^{T1}) = ATE_t^{T1}, t = 31, \ldots, 60$$

- Allows us to continue measuring ATEs for 31-60 of exposure



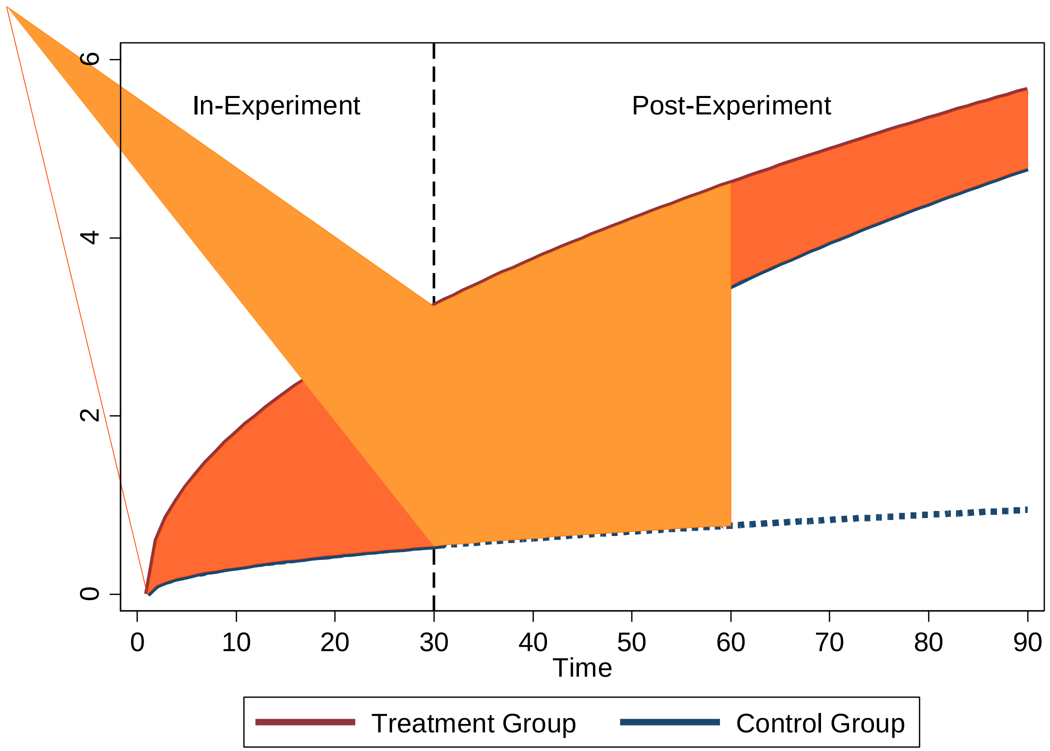
Identification and Estimation (Simplified Example)

- We continue to face the same issue with $T1-C$ differences in average outcomes in days 61-90
- But we just measured the ATEs for 31-60 days of exposure using our method
- We can now correct the average outcomes of C by subtracting out these ATEs
- This allows us to continue measuring ATEs for 61-90 of exposure (and further)



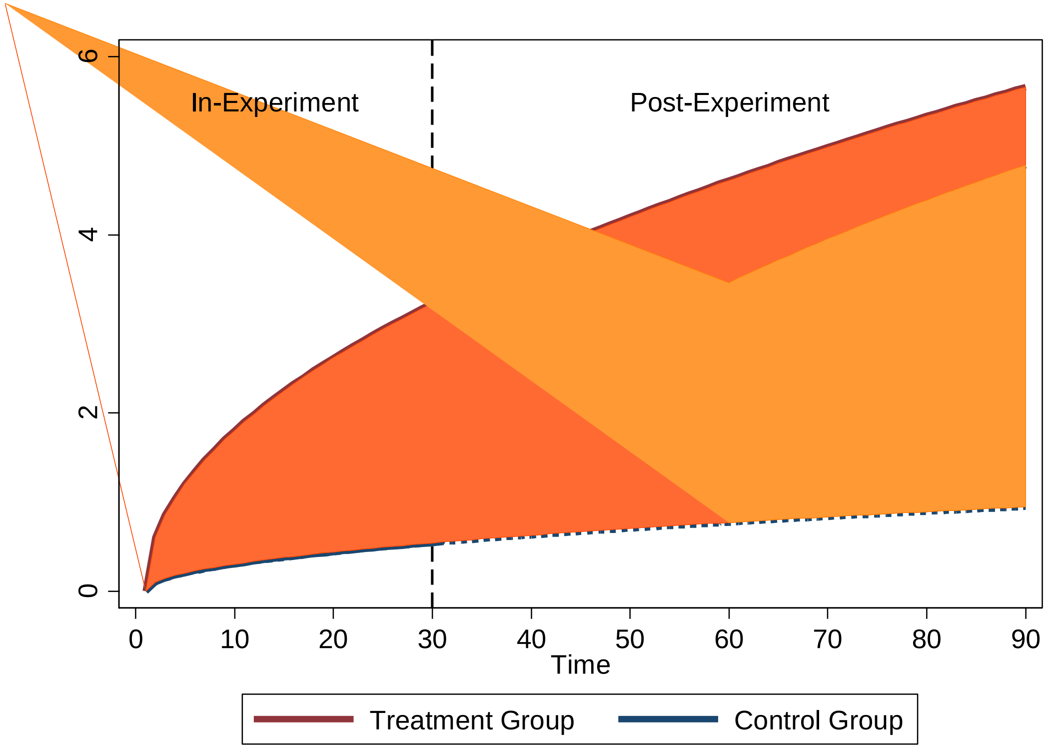
Identification and Estimation (Simplified Example)

- We continue to face the same issue with $T1 - C$ differences in average outcomes in days 61-90
- But we just measured the ATEs for 31-60 days of exposure using our method
- We can now correct the average outcomes of C by subtracting out these ATEs
- This allows us to continue measuring ATEs for 61-90 of exposure (and further)



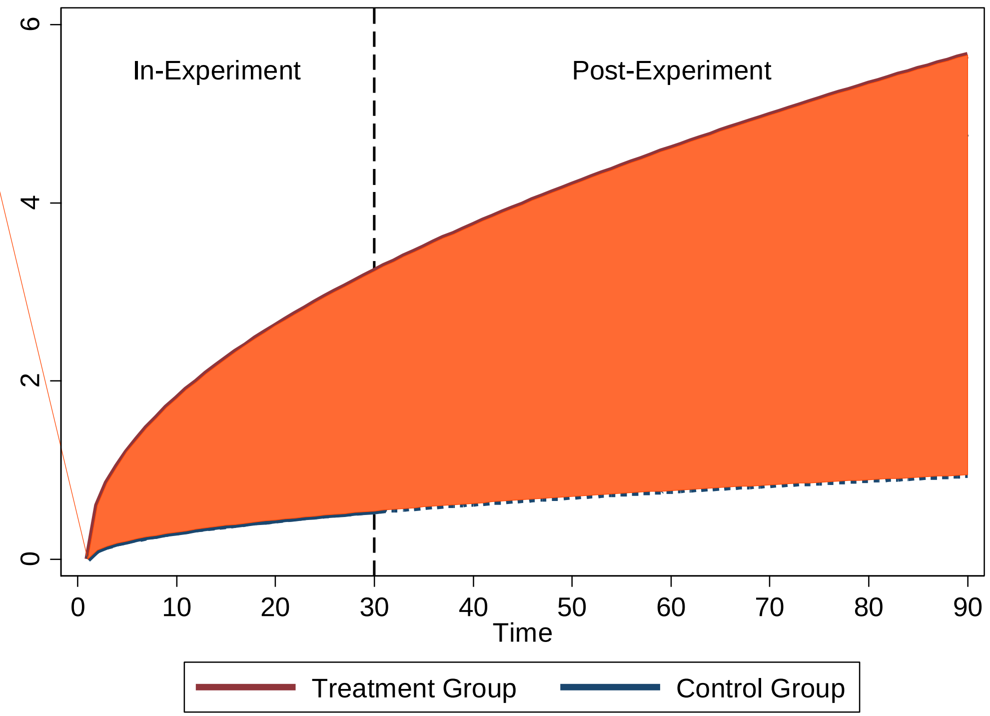
Identification and Estimation (Simplified Example)

- We continue to face the same issue with $T1 - C$ differences in average outcomes in days 61-90
- But we just measured the ATEs for 31-60 days of exposure using our method
- We can now correct the average outcomes of C by subtracting out these ATEs
- This allows us to continue measuring ATEs for 61-90 of exposure (and further)



Identification and Estimation (Simplified Example)

- We continue to face the same issue with $T1-C$ differences in average outcomes in days 61-90
- But we just measured the ATEs for 31-60 days of exposure using our method
- We can now correct the average outcomes of C by subtracting out these ATEs
- This allows us to continue measuring ATEs for 61-90 of exposure (and further)



Theoretical Contribution

- We show theoretical results for our estimator (closed form and simulation)
- Our estimation technique is unbiased for the correct long-term effect under a key assumption
  - It allows any pattern in outcome unrelated to treatment (e.g., seasonality)
  - Treatment effects can follow any function in terms of duration of exposure

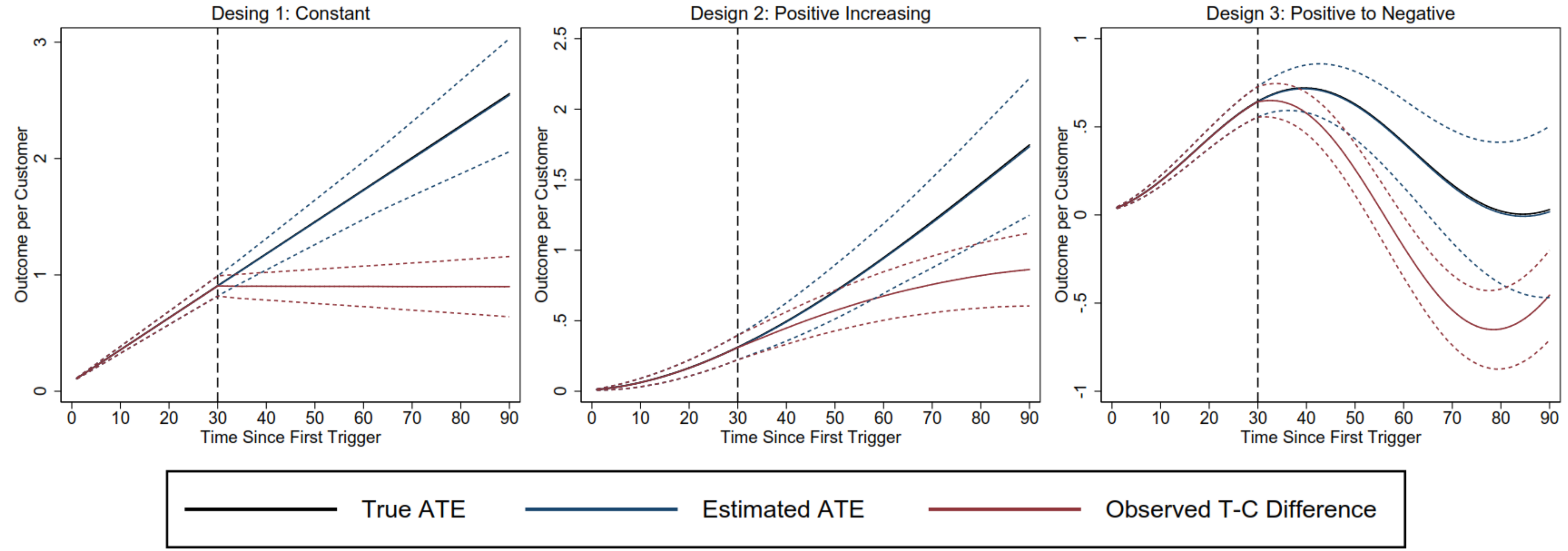
Key Assumption: time-stable causal effects

- **Assumption**: causal effects are the same for the time period of A/B start and dial-up for a similar population
- Violated if there are un-modeled seasonality or trends in effects
- We need this assumption to use A/B tests anyway
- Experiments always look forward: we learn an effect today to predict the impact of changing a policy tomorrow
- We only need to perform better than the baseline approach (simple T/C difference for all periods)

Types of Effects

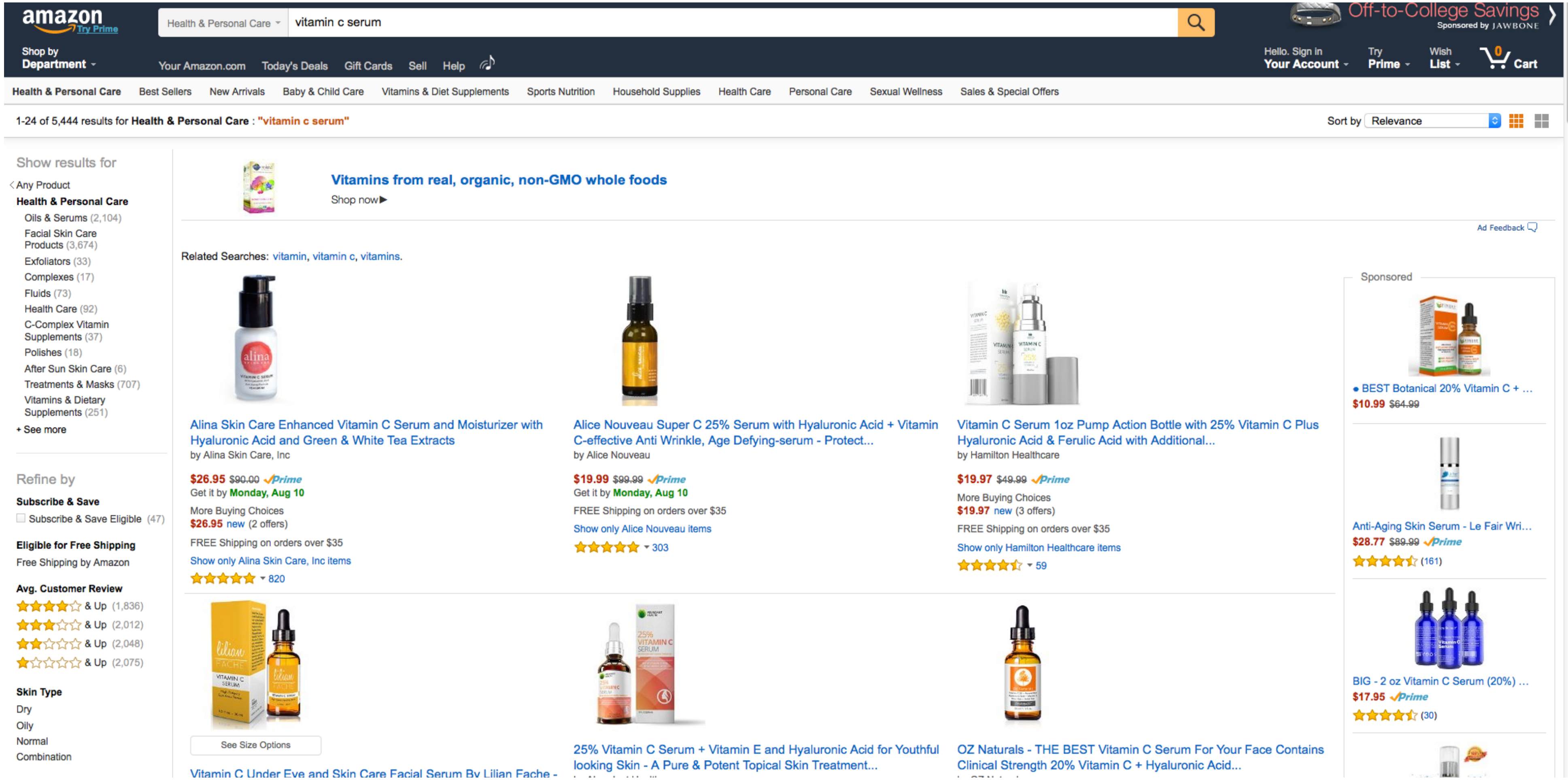
- Constant effect
- Effect that grows over time
- Effect smaller over time (“novelty effect”)
- Treatment displaces spend over time (e.g., “Pull forward” effect)

Types of Effects: Monte Carlo Simulation

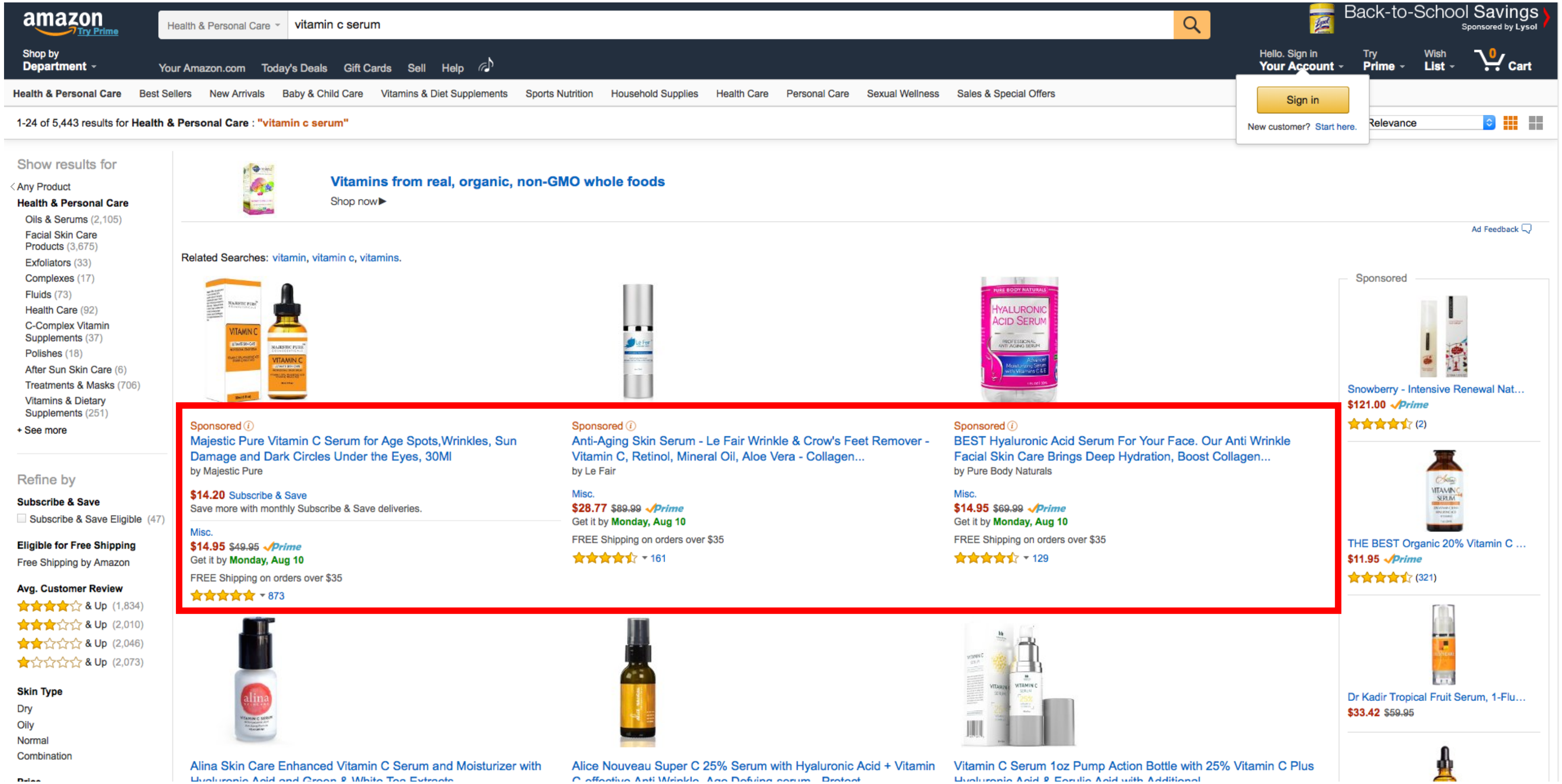


Note: Dashed lines represent simulation-based 95% confidence intervals

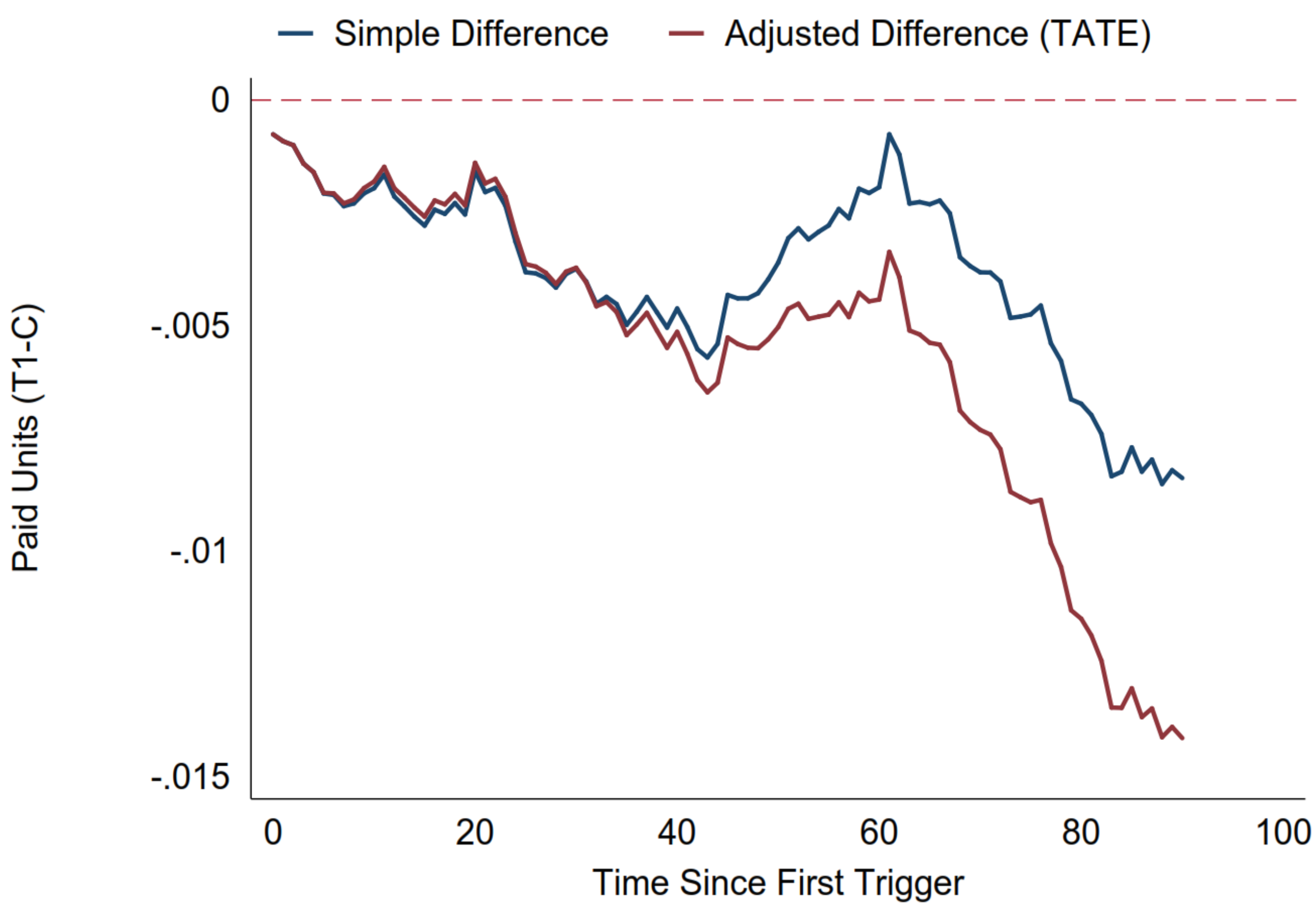
Example: Sponsored Products Weblab (Control)



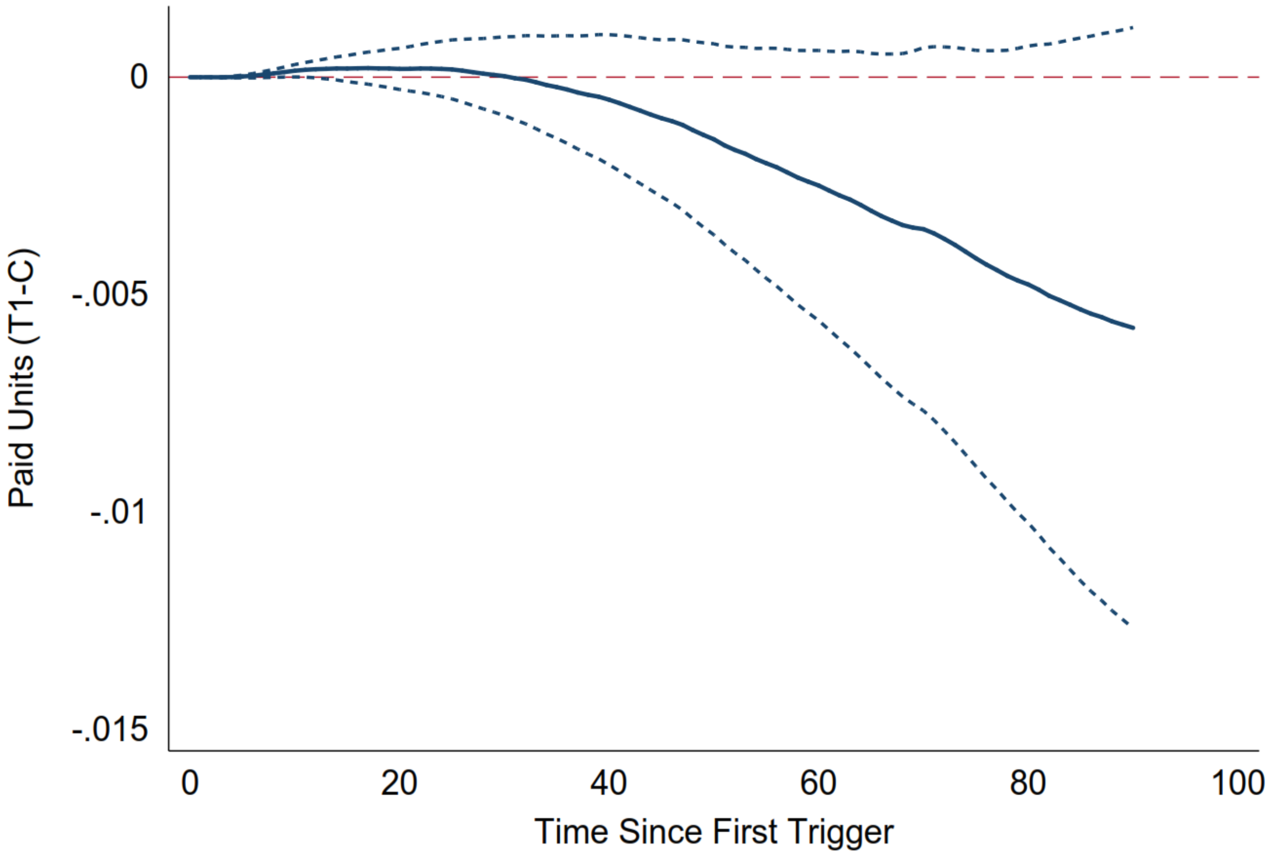
Example: Sponsored Products Weblab (Treatment)



Estimates of Long-Term Cumulative Effect

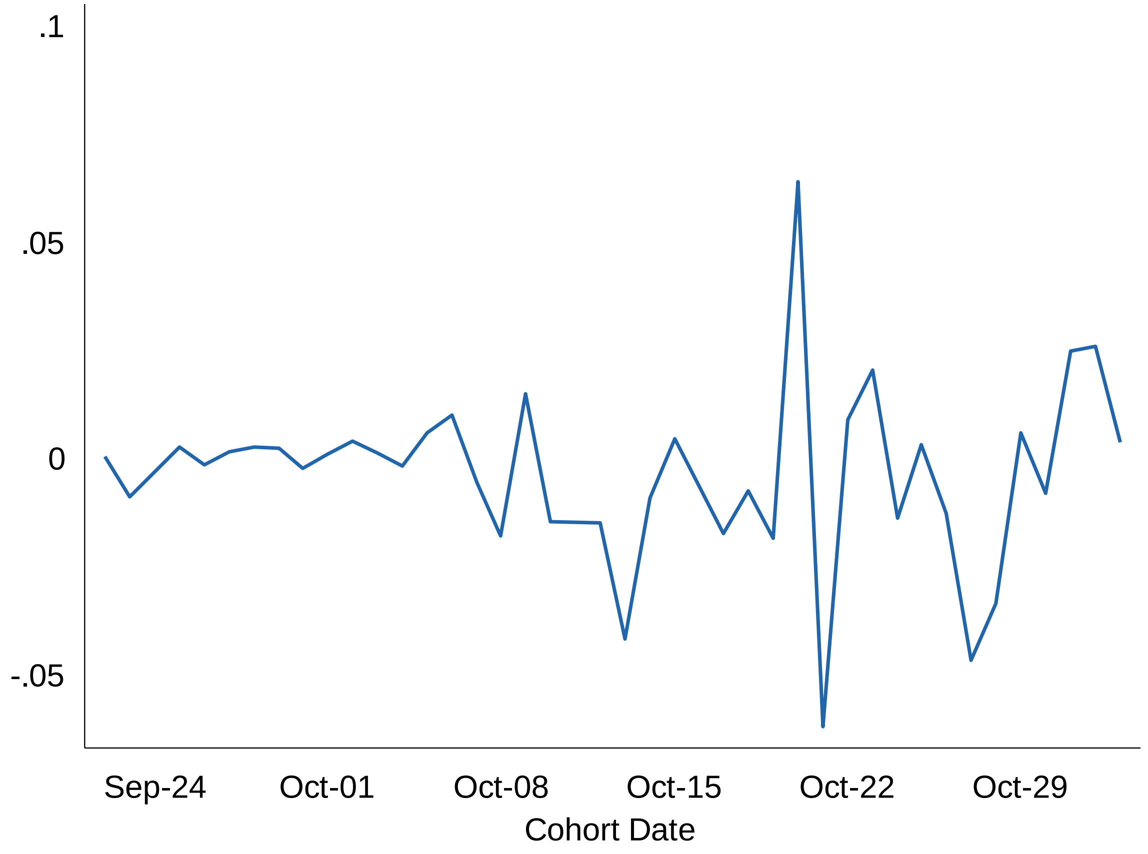


Difference between estimates of cumulative effect (Simple Difference - Adjusted Difference)

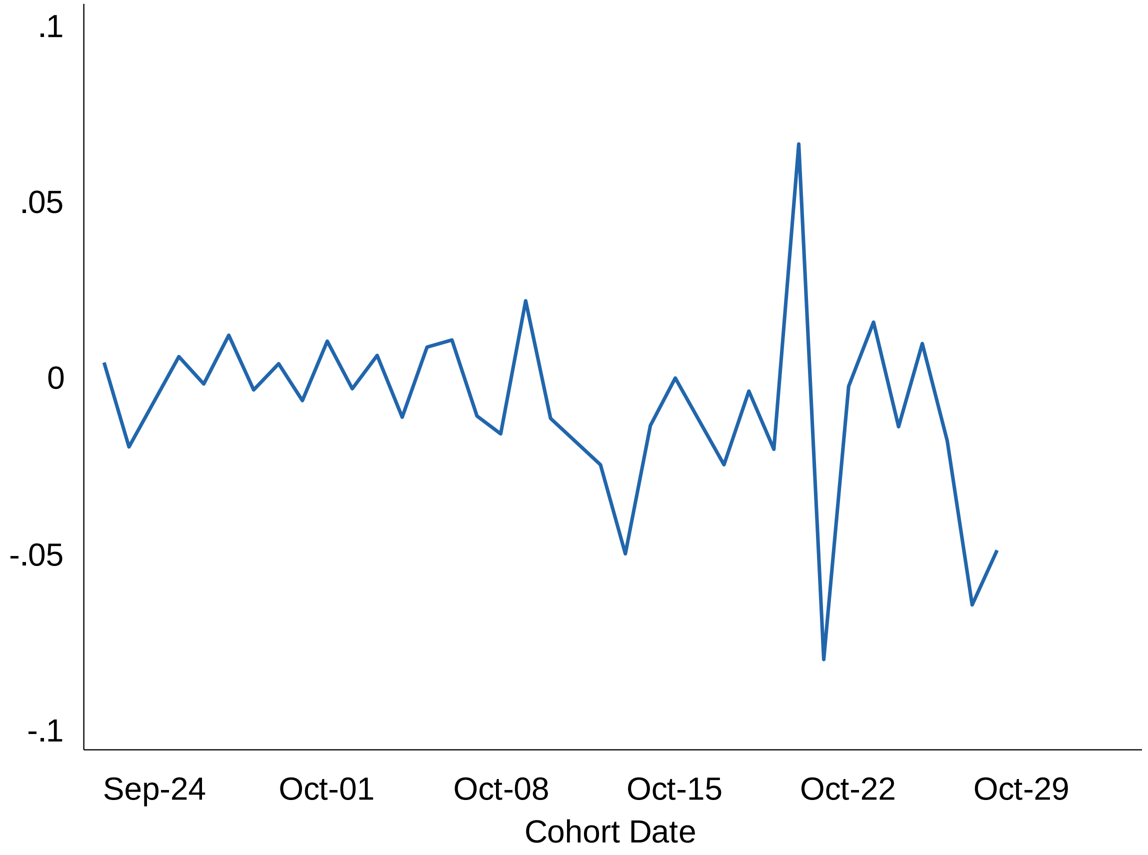


Dashed lines represent 95% confidence interval.

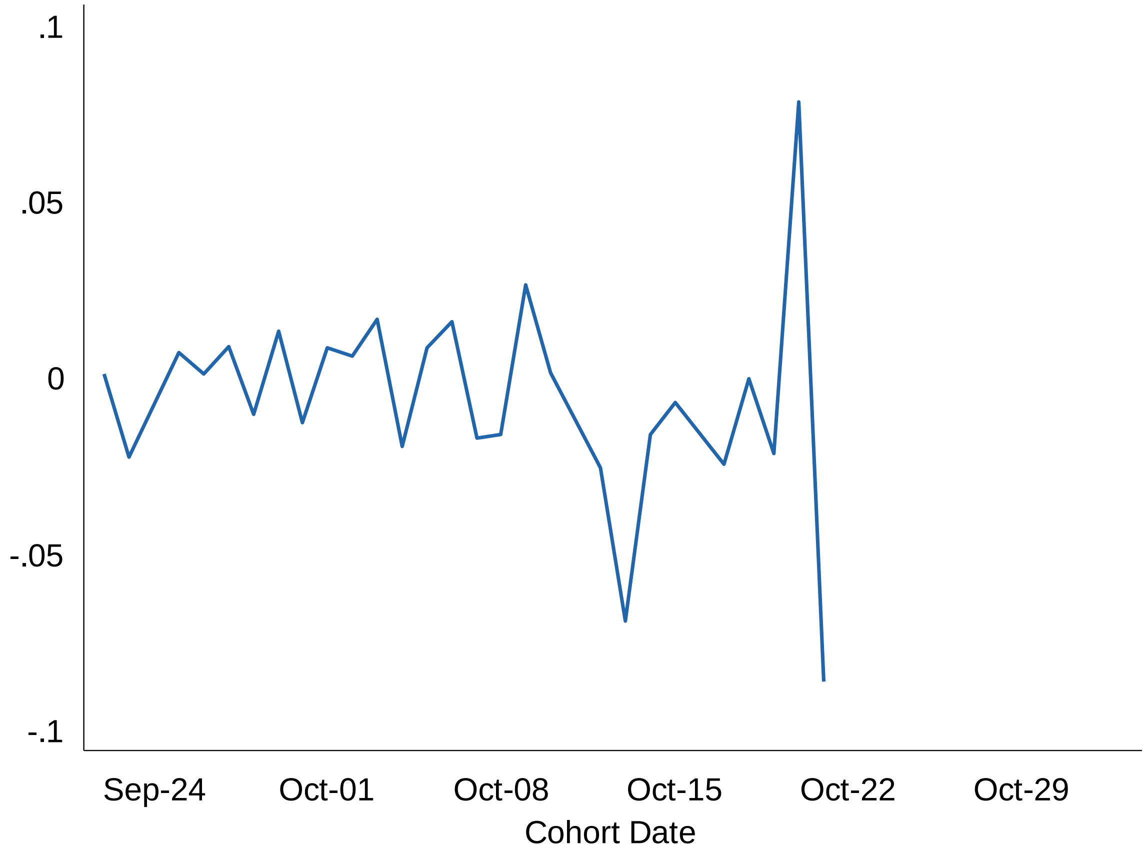
Validation: time constant effects assumption (1-Day Effects)



Validation: time constant effects assumption (7-Day Effects)



Validation: time constant effects assumption (14-Day Effects)



Future Work

- Improve estimation strategy to reduce variance
- Reduce noise using high-quality ML estimators
- Re-analyze more Weblabs
- Use long-term holdouts to validate
- Build as monitoring metric into APT/Weblab