The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely

Susan Athey, Stanford

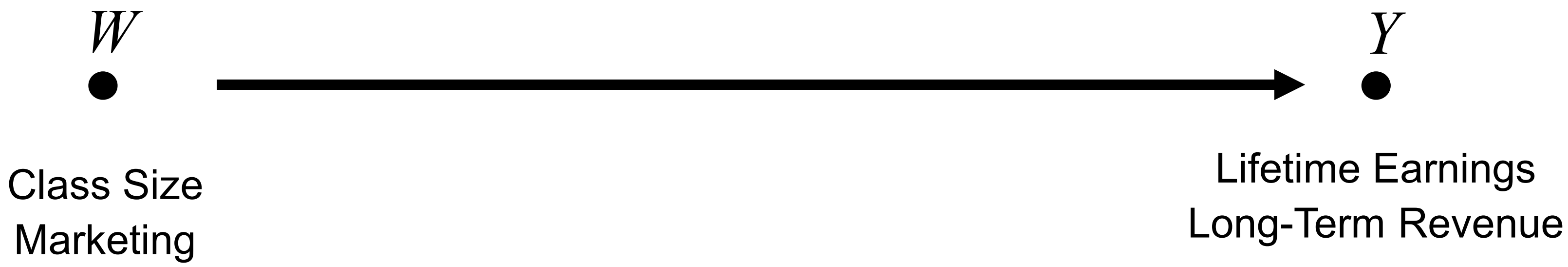
Raj Chetty, Harvard

Guido Imbens, Stanford

Hyunseung Kang, UW-Madison

November 2019

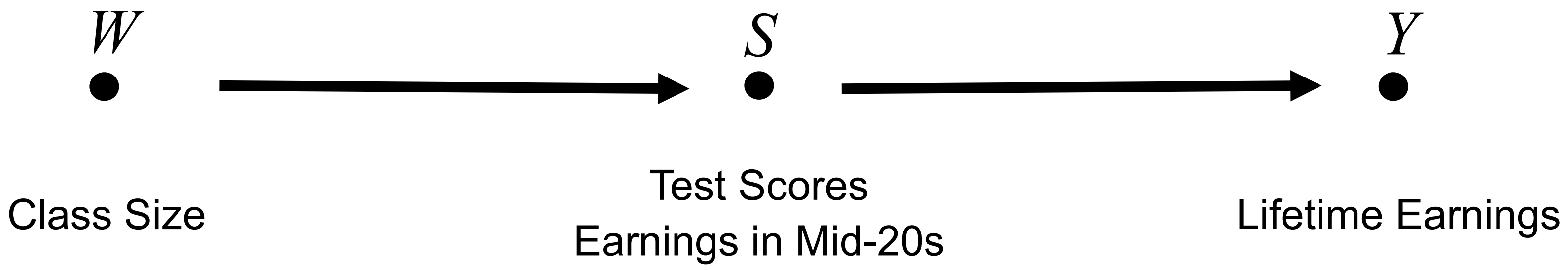
Problem: Estimating Long-Term Impacts of Interventions



- Estimating long-term impacts of treatments is central in many fields, from economics to marketing
- Two key challenges in estimating long-term treatment effects using conventional experimental/quasi-experimental methods

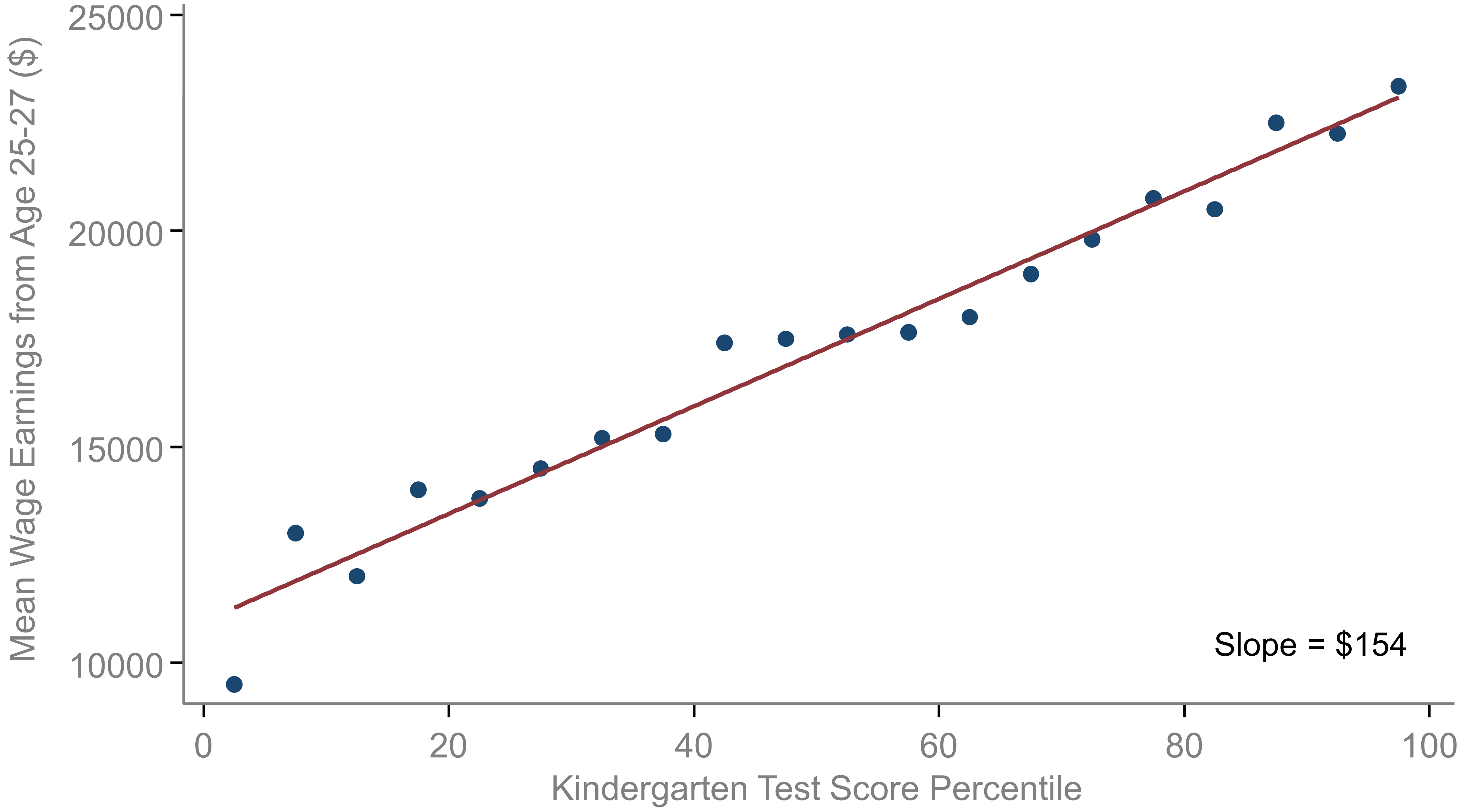
    1. Long delays in observing impacts
    2. Experimental estimates are often very imprecise

Using Short-Term Outcomes as Proxies

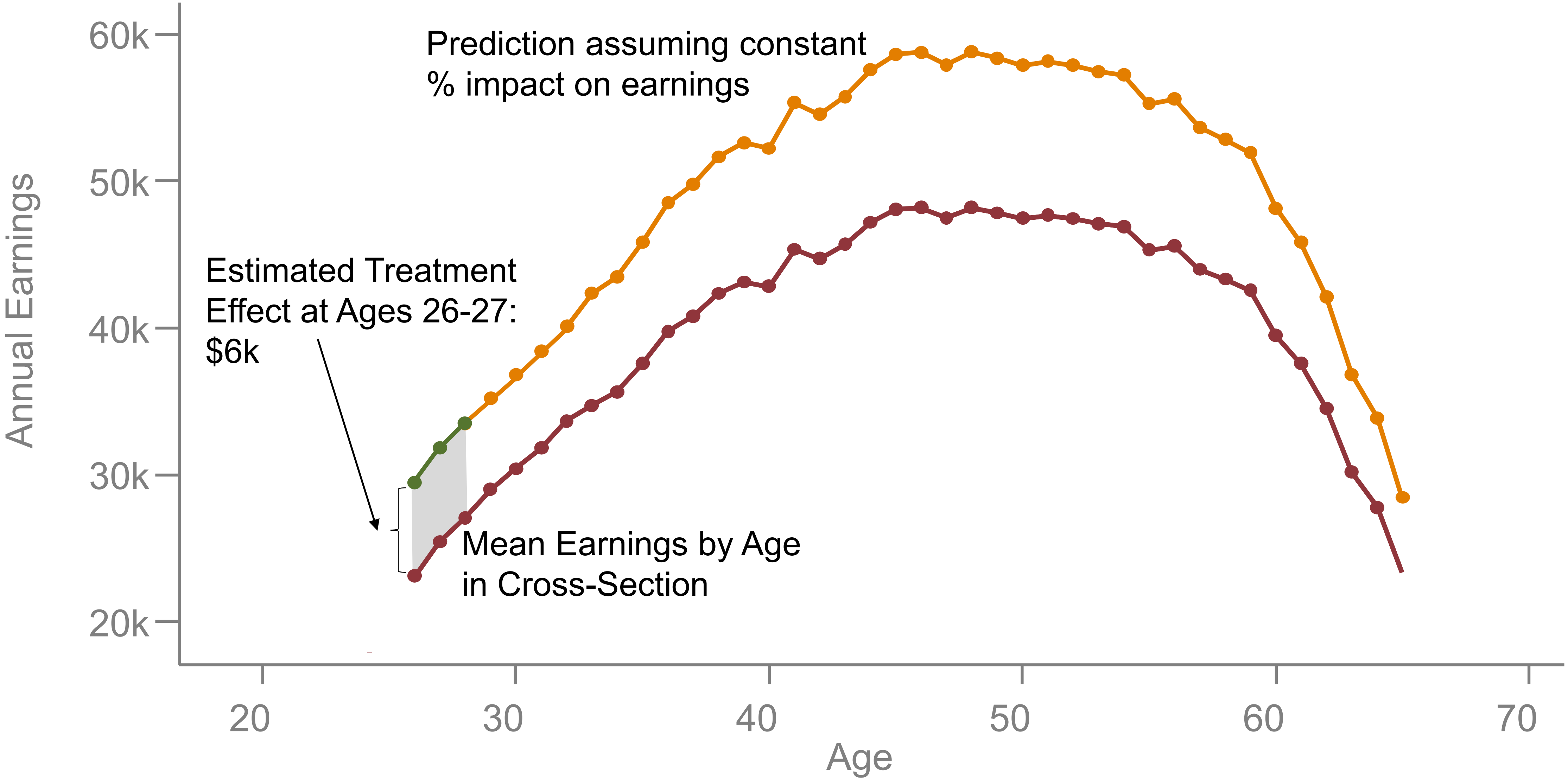


- One intuitive solution: use short-term proxies to predict long-term impacts
  - Estimate effect of treatment on an intermediate outcome **S**
  - Regress **Y** on **S** in observational data and multiply treatment effect on **S** by this regression coefficient to predict long-term impact
  - This is common in the social sciences…

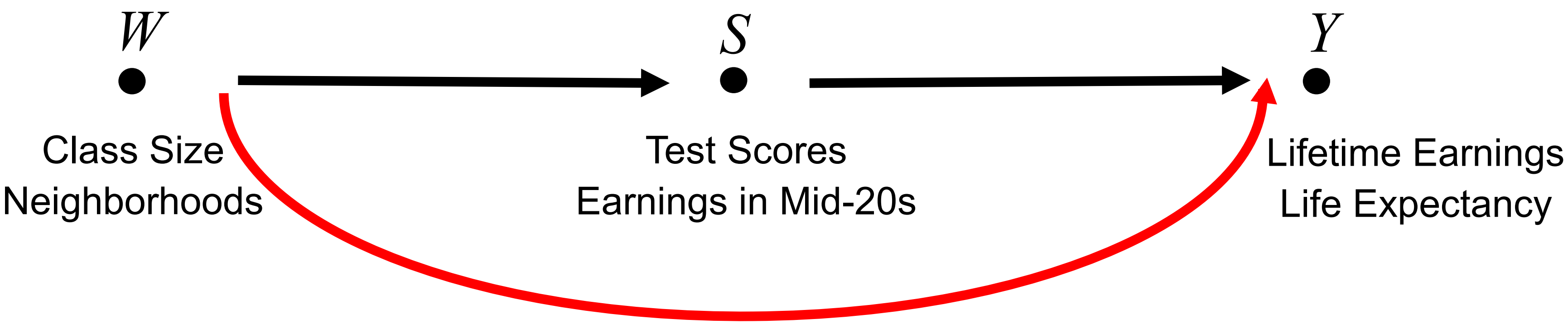
Predicting Earnings from Early Childhood Test Scores



Predicting Lifetime Earnings Impacts
Using Treatment Effect Estimates on Earnings in Early Adulthood



Potential Solution: Surrogates



- Prentice (1989) formalized this approach in biostatistics, labeling an intermediate outcome a *surrogate* if **Y** is independent of **W** conditional on **S**

- Problem: validity of this assumption is often unclear in applications
  - Do test scores fully capture impacts on earnings by themselves?
  - Do short-term impacts on earnings accurately reflect lifetime earnings impacts?

This Paper: Combining Multiple Short-Term Proxies

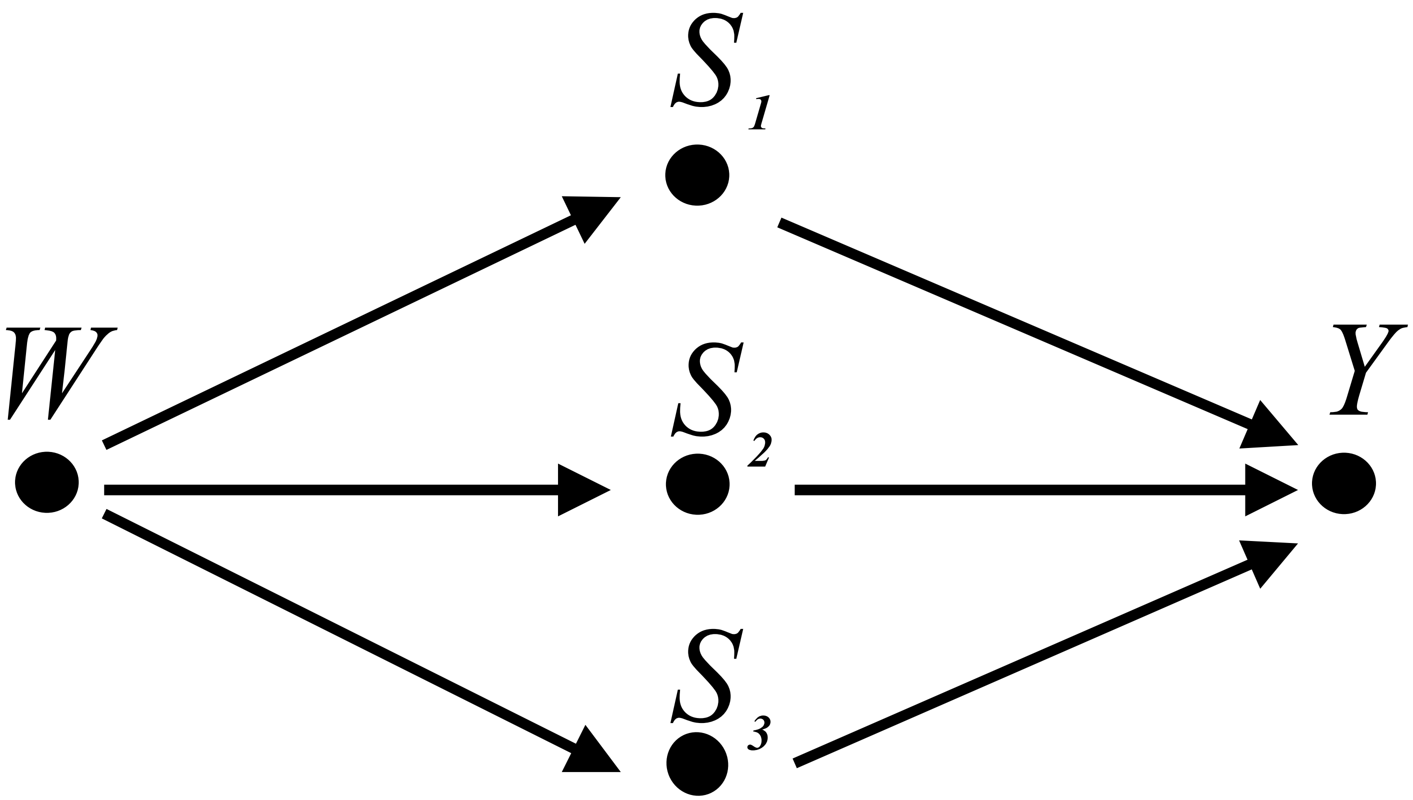
- How can we estimate long-term treatment effects when we don't necessarily have a valid surrogate?

- We show how we can make progress on these issues in the era of big data, where we typically have many intermediate outcomes, not just one potential surrogate

- Rather than debating whether any *one* variable is a valid statistical surrogate, combine many short-term proxies to create a "surrogate index"

  - Combining many variables makes it more likely that we span all the causal pathways from treatment to long-term outcome

Combining Multiple Surrogates



This Paper

- Simple idea: form predicted value of long-term outcome using multiple surrogates (e.g., via linear regression) and estimate treatment effects on that predicted value
  - This can allow us to estimate long-term treatment effects more quickly *and* more precisely (smaller standard errors)
- Approach is intuitive, but most work still uses a single variable as a candidate surrogate

This Paper

- Contributions of this paper:

  1. [Identification] Formalize assumptions required for identification using surrogate index

  2. [Bias] Bound bias from violations of these assumptions and show how they can be validated

  3. [Precision] Characterize gains in precision from using surrogate index instead of long-term outcome

  4. [Application] Apply method to show practical value of combining proxies for problems we work on

- Illustrate method and key results primarily focusing on empirical application here

Setup

- Assume researcher has two different datasets:
  - *Experimental dataset* (**E**): data on **W** (treatment) and **S** (intermediate outcome), with **W** randomly assigned
    - Example: Tennessee STAR experiment that varied class size randomly
  - *Observational dataset* (**O**)**:** data on **S** and **Y** (long-term outcome), and possibly **W**, with **W** not randomly assigned
    - Example: standard school district dataset linked to long-term outcome data

The Surrogate Index

- Surrogate index is the conditional expectation of long-term outcome given the intermediate outcomes (and any pre-treatment covariates) in the observational dataset

$$\mathbb{E}[Y_i \mid S_i = s, X_i = x, P_i = O]$$

- In a linear model, can be estimated as the predicted value from a regression of the long-term outcome on the intermediate outcomes

Identification Using the Surrogate Index

- Treatment effect on the surrogate index in the experimental sample is an unbiased estimate of treatment effect on the long-term outcome under three assumptions:

Assumption 1 (Unconfounded Treatment Assignment):

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1), S_i(0), S_i(1)\Big) \Big| P_i = \mathrm{E}$$

Assumption 2 (Surrogacy):

$$W_i \perp\!\!\!\perp Y_i \Big| S_i, P_i = \mathrm{E}.$$

Assumption 3 (Comparability):

$$Y_i \Big| S_i, P_i = \mathrm{O} \quad \sim \quad Y_i \Big| S_i, P_i = \mathrm{E}$$

Empirical Application: California GAIN Training Program

- California Greater Avenues to Independence program: job assistance program implemented in late 1980s to help welfare (AFDC) recipients find work

- MDRC conducted a randomized trial of GAIN in four urban counties: Alameda (Oakland), Los Angeles, Riverside, and San Diego

- Focus first on Riverside program, which was widely heralded as being the most successful program that had the largest impacts on employment and earnings
  - Riverside emphasized a “jobs first” approach to re-entry into labor force (rather than human capital development/training to find ideal match)

- Then return to other sites, which we hold out and use for out-of-sample validation

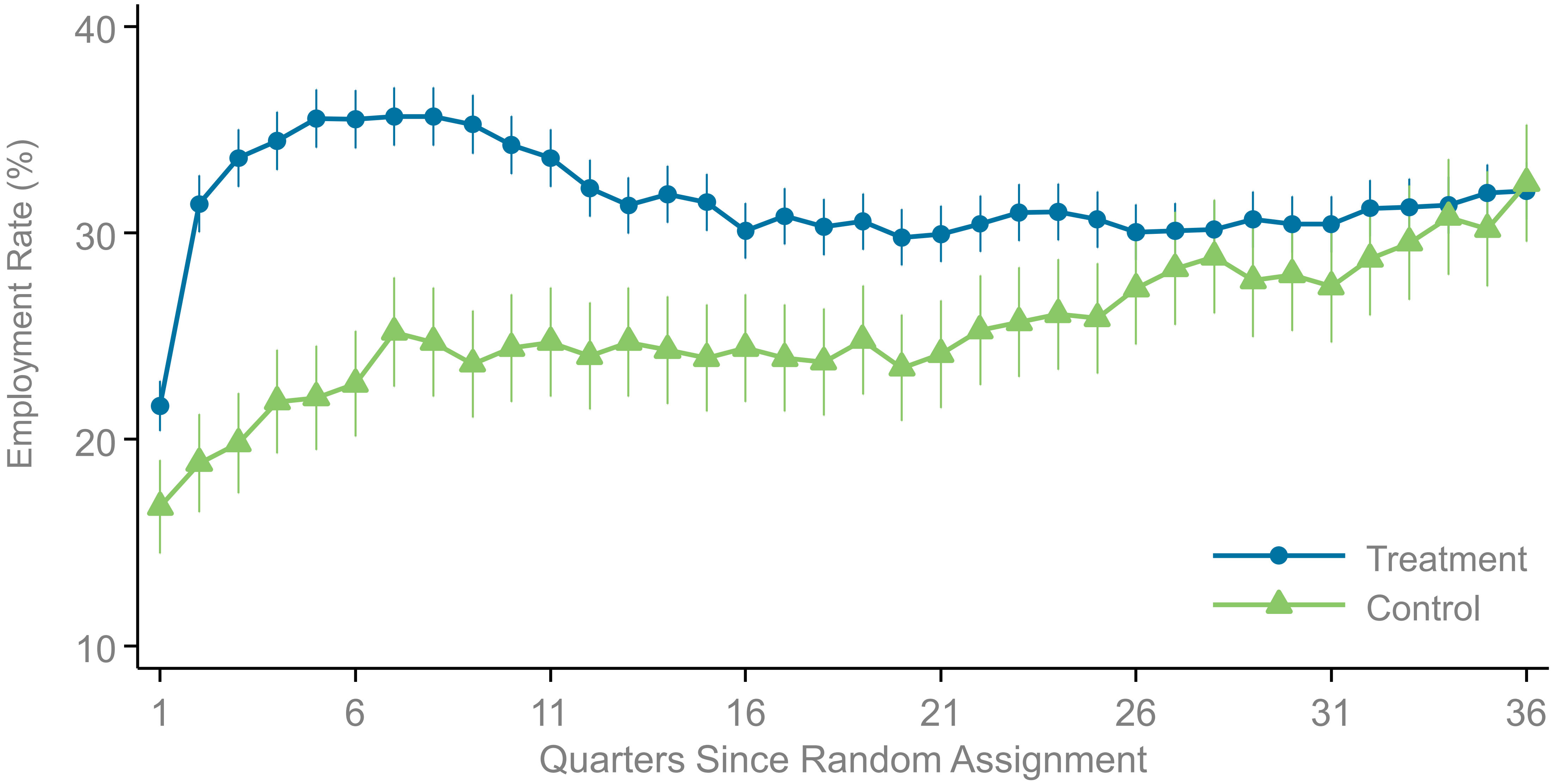
Riverside GAIN Program: Experimental Analysis

- Use data from Hotz, Imbens, and Klerman (2006), who conducted a nine-year follow-up using data from UI records

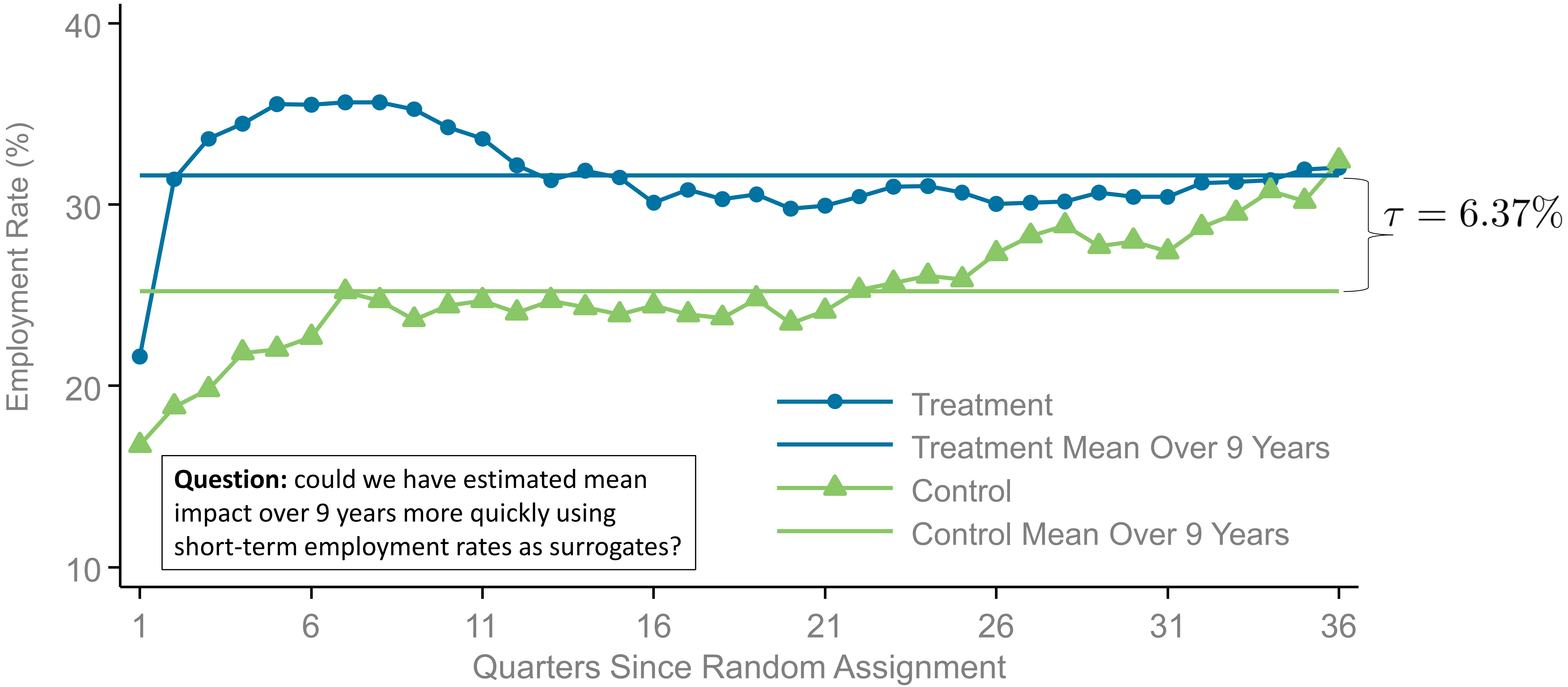
- 5,445 individuals participated in program in Riverside, randomly assigned to treatment and control

  - At baseline: 22% employed; mean quarterly earnings of $452

Employment Rates in Treatment vs. Control Group, by Quarter



Employment Rates in Treatment vs. Control Group, by Quarter



Question: could we have estimated mean impact over 9 years more quickly using short-term employment rates as surrogates?

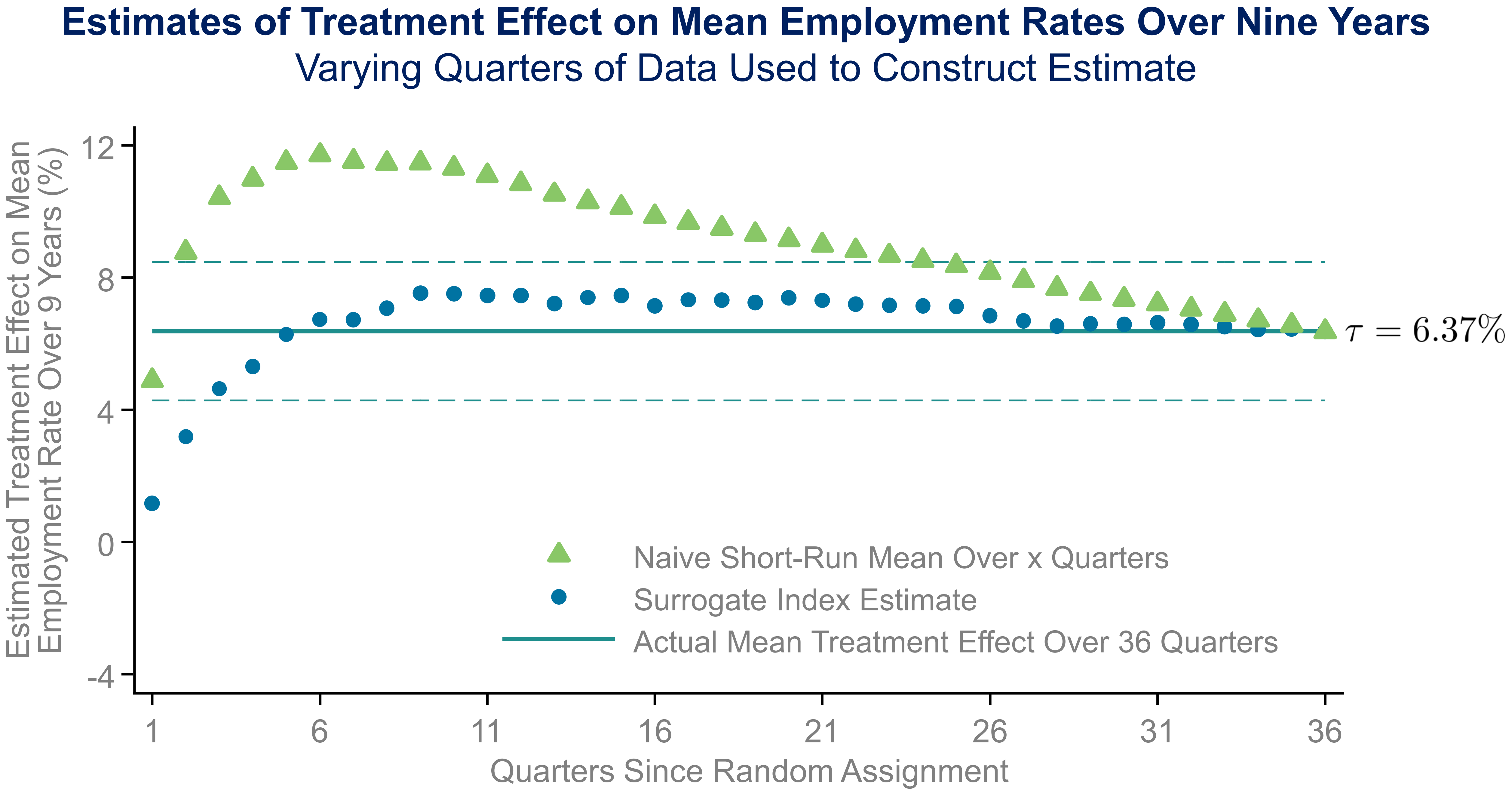
Construction of Surrogate Index

- Construct surrogate index by regressing mean employment rate over 36 quarters on employment indicators from quarter 1 to quarter *S:*

$$\bar{Y}_{iT} = \beta_0 + \sum_{t=1}^{S} \beta_t Y_{it} + \varepsilon_i$$

- Then estimate treatment effect on surrogate index based on employment rates up to quarter *S*

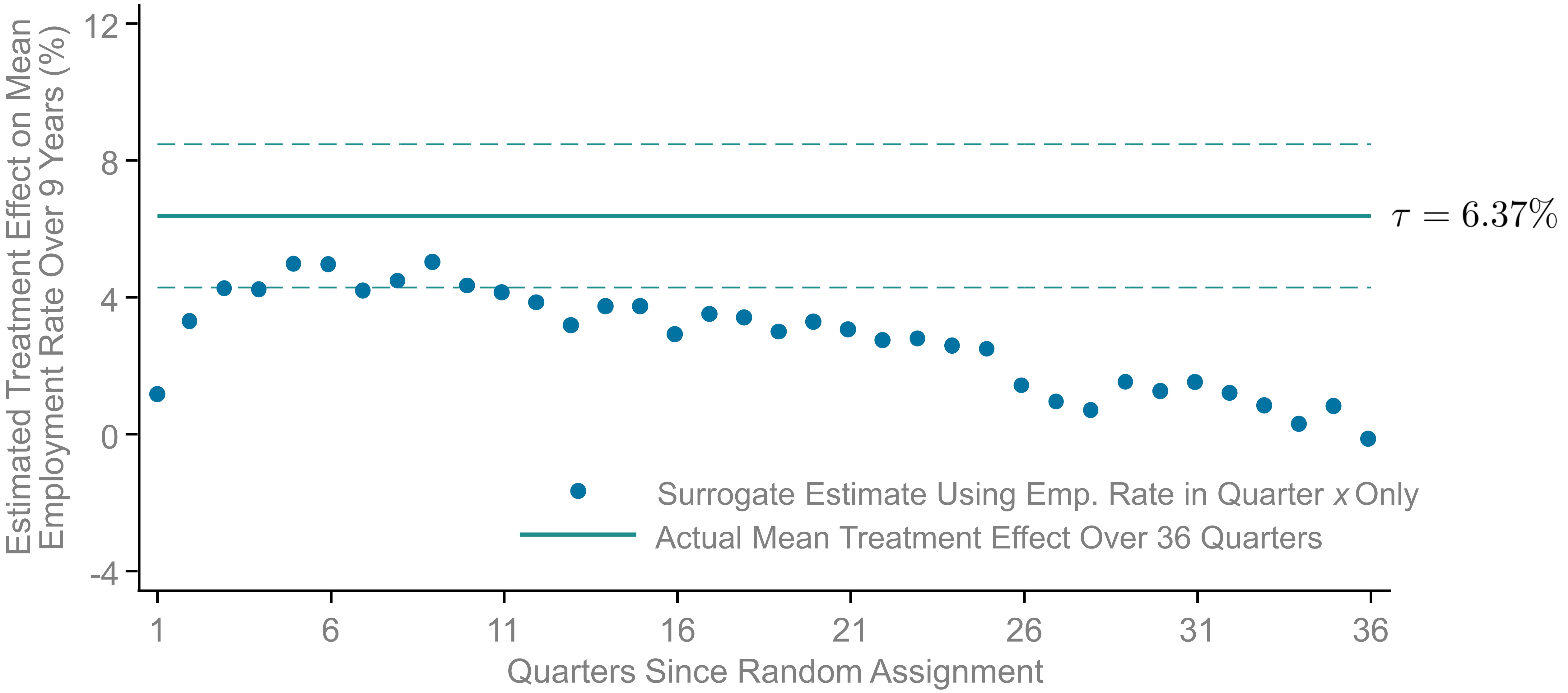
- Assess how quickly (at what value of *S*) we can estimate nine-year mean impact accurately

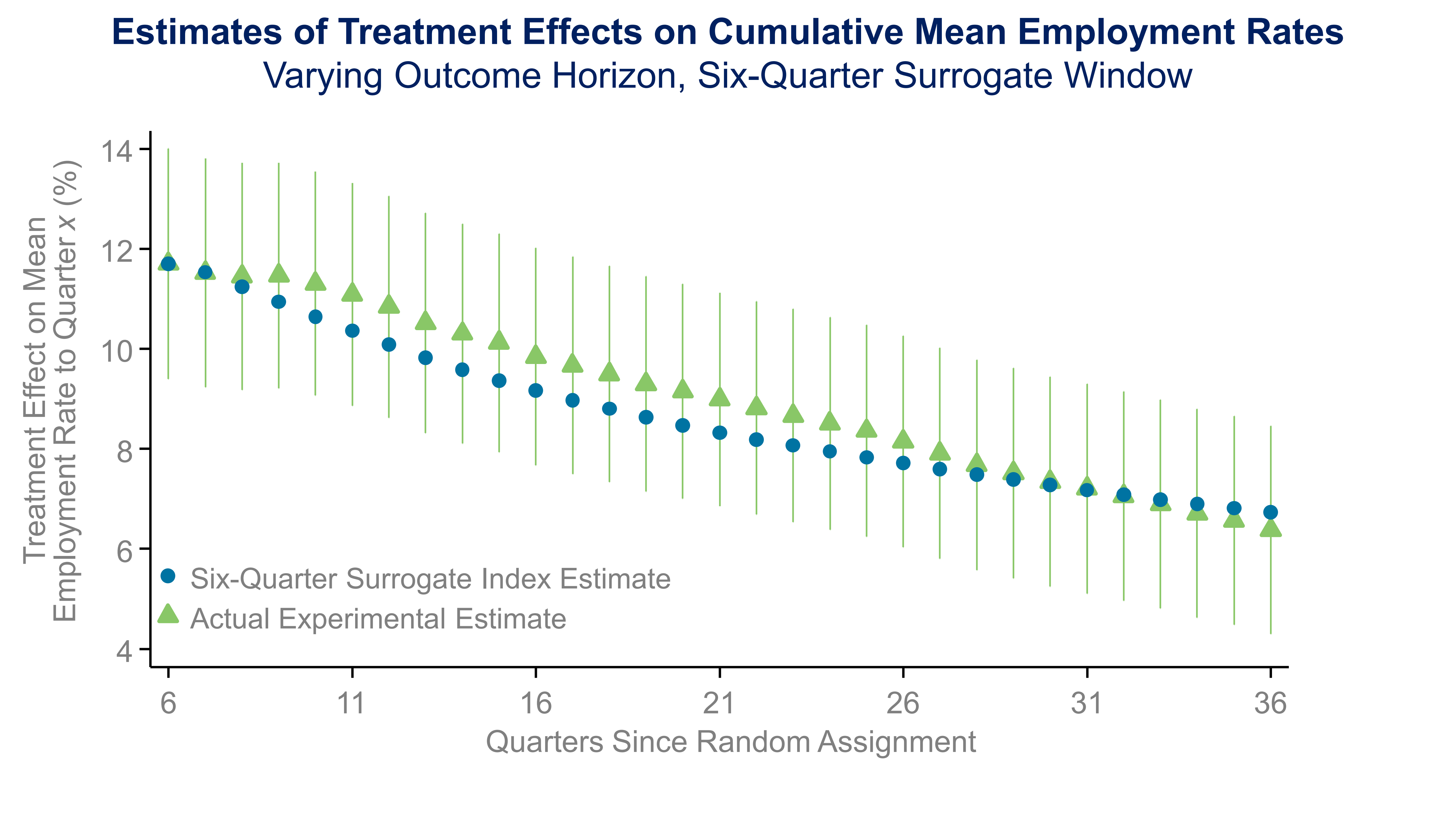


Estimates of Treatment Effect on Mean Employment Rates Over Nine Years
Varying Quarters of Data Used to Construct Estimate
Estimated Treatment Effect on Mean Employment Rate Over 9 Years (%)
12
8
4
0
-4
1
6
11
16
21
26
31
36
Quarters Since Random Assignment
$\tau = 6.37\%$
Naive Short-Run Mean Over x Quarters
Surrogate Index Estimate
Actual Mean Treatment Effect Over 36 Quarters

Estimates of Treatment Effect on Mean Employment Rates Over Nine Years

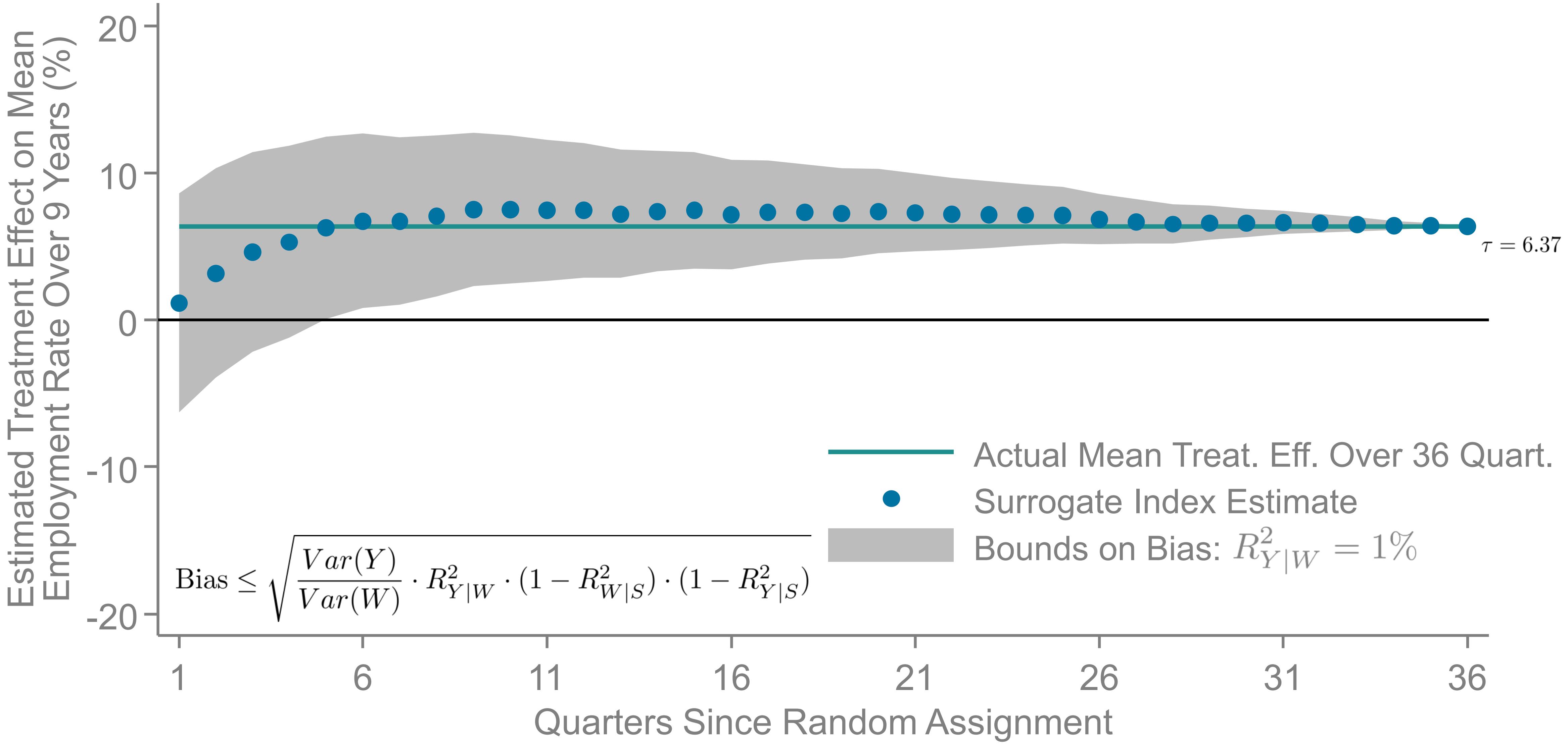
Varying Quarters of Data Used to Construct Estimate



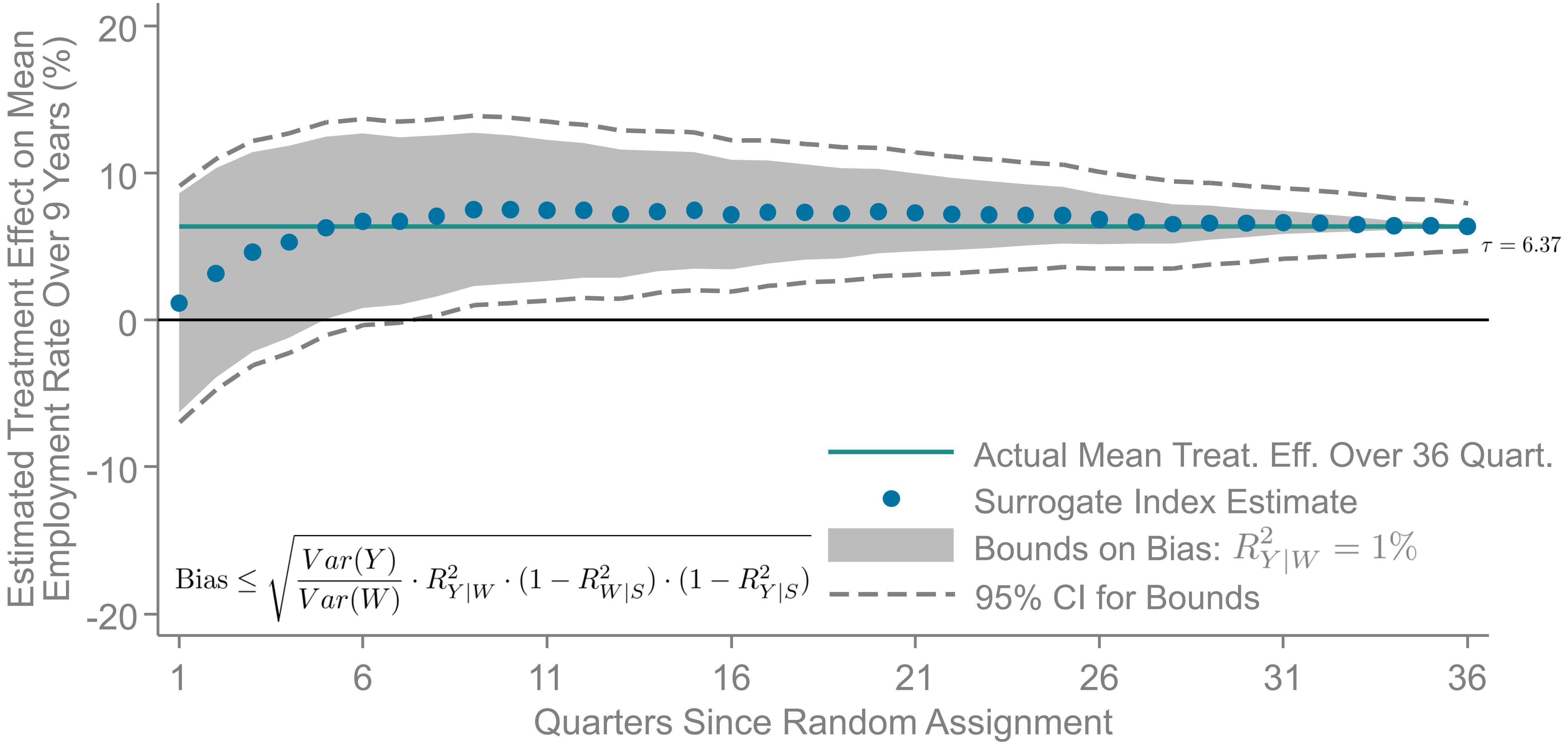

Estimates of Treatment Effects on Cumulative Mean Employment Rates
Varying Outcome Horizon, Six-Quarter Surrogate Window
Treatment Effect on Mean
Employment Rate to Quarter x (%)
14
12
10
8
6
4
6
11
16
21
26
31
36
Quarters Since Random Assignment
Six-Quarter Surrogate Index Estimate
Actual Experimental Estimate

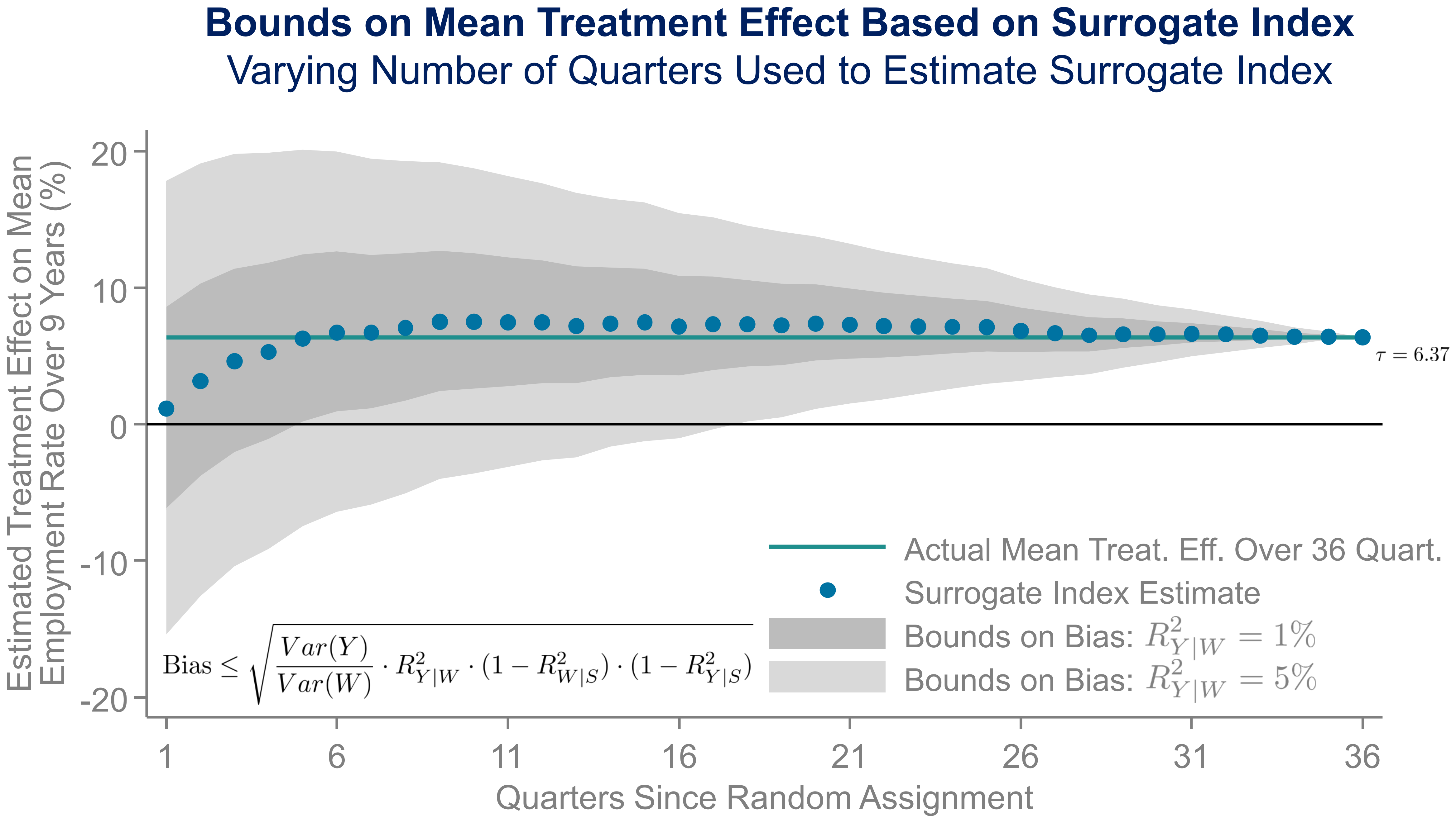
Bounds on Mean Treatment Effect Based on Surrogate Index

Varying Number of Quarters Used to Estimate Surrogate Index

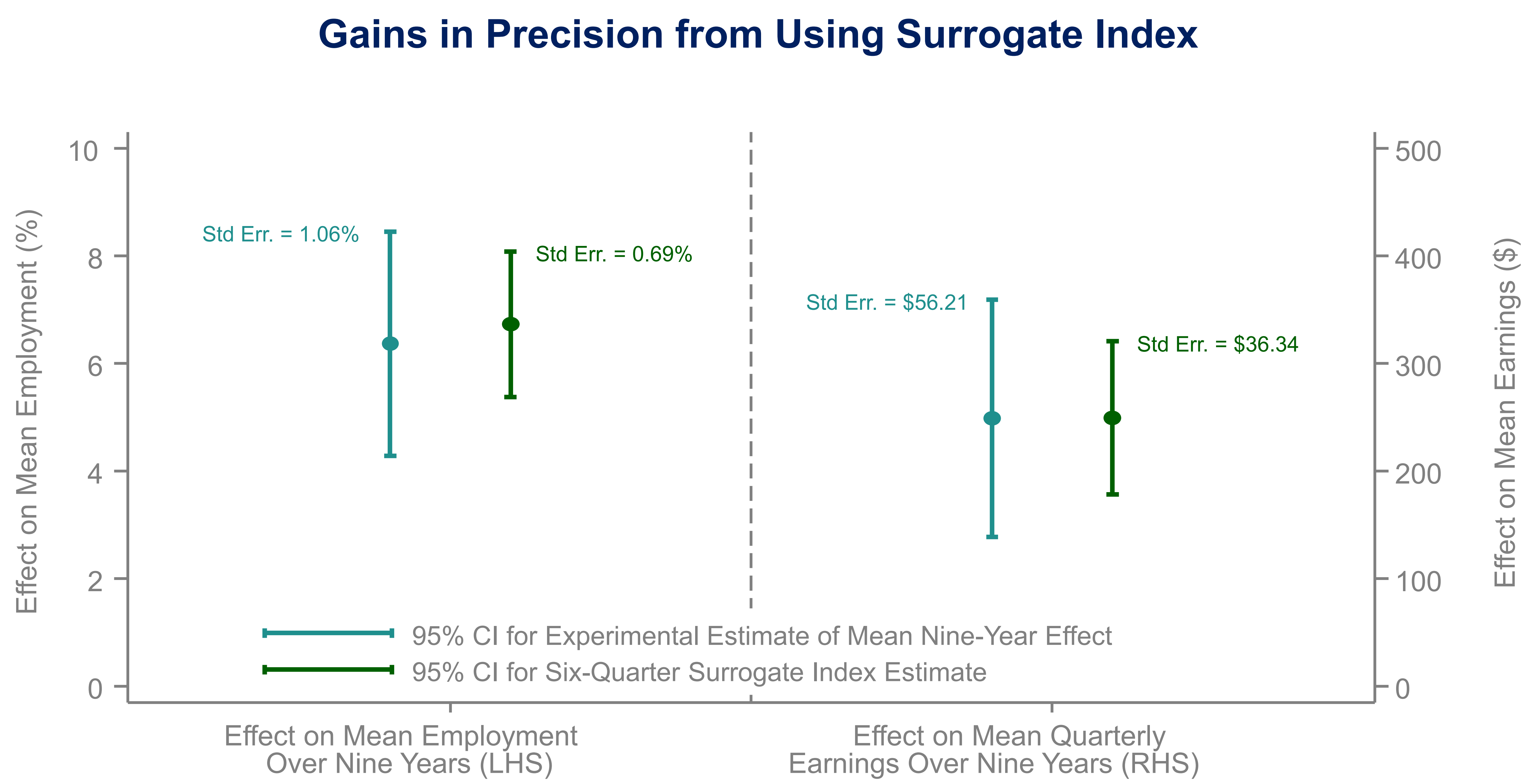


Bounds on Mean Treatment Effect Based on Surrogate Index
Varying Number of Quarters Used to Estimate Surrogate Index





Bounds on Mean Treatment Effect Based on Surrogate Index
Varying Number of Quarters Used to Estimate Surrogate Index
Estimated Treatment Effect on Mean Employment Rate Over 9 Years (%)
20
10
0
-10
-20
1
6
11
16
21
26
31
36
Quarters Since Random Assignment
$\tau = 6.37$
Actual Mean Treat. Eff. Over 36 Quart.
Surrogate Index Estimate
Bounds on Bias: $R^2_{Y|W} = 1\%$
Bounds on Bias: $R^2_{Y|W} = 5\%$
$\text{Bias} \leq \sqrt{\frac{Var(Y)}{Var(W)} \cdot R^2_{Y|W} \cdot (1 - R^2_{W|S}) \cdot (1 - R^2_{Y|S})}$


Gains in Precision from Using Surrogate Index
Effect on Mean Employment (%)
Effect on Mean Earnings ($)
10
8
6
4
2
0
500
400
300
200
100
0
Std Err. = 1.06%
Std Err. = 0.69%
Std Err. = $56.21
Std Err. = $36.34
95% CI for Experimental Estimate of Mean Nine-Year Effect
95% CI for Six-Quarter Surrogate Index Estimate
Effect on Mean Employment Over Nine Years (LHS)
Effect on Mean Quarterly Earnings Over Nine Years (RHS)

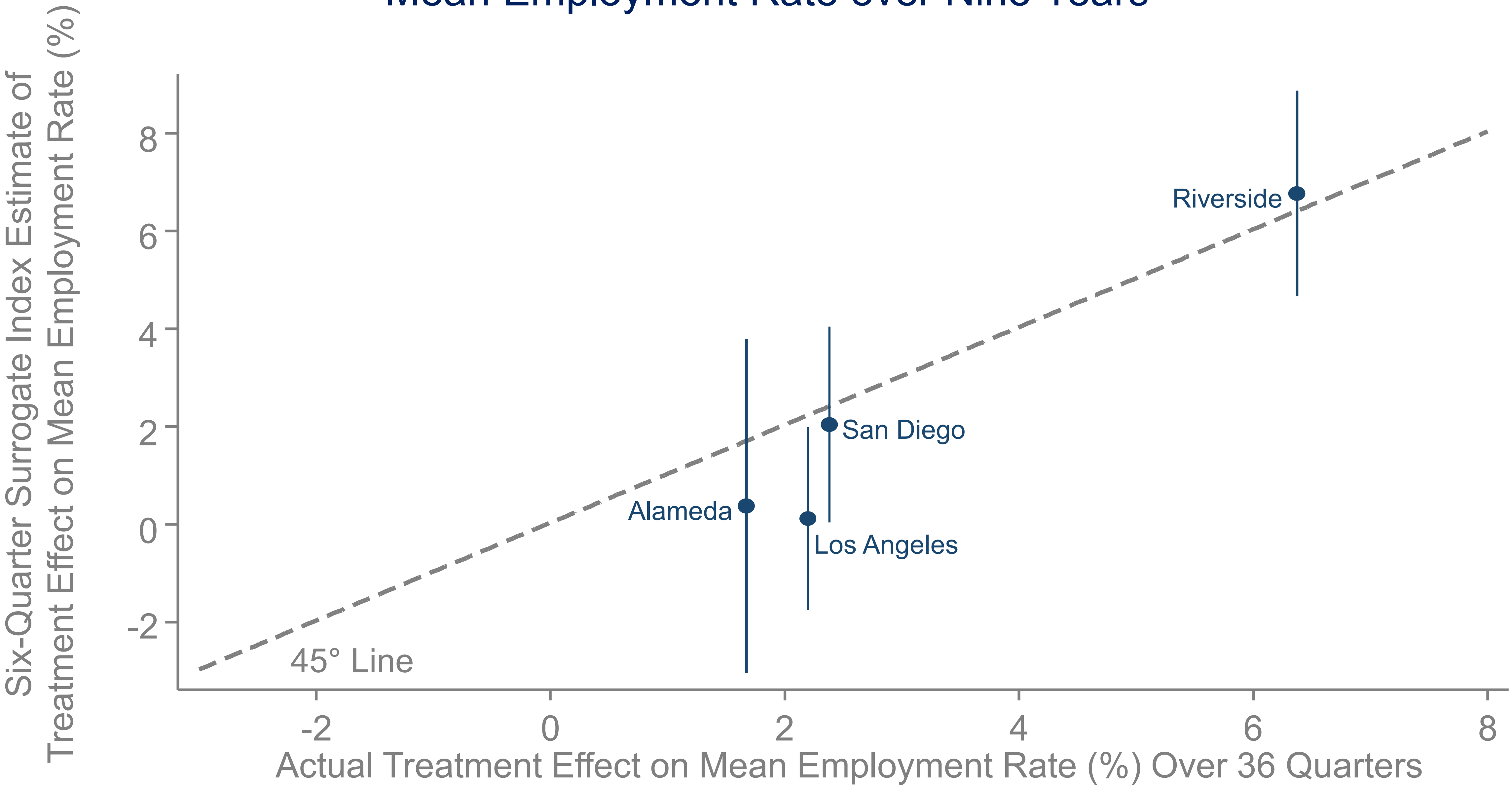
Predicting Cross-Site Heterogeneity

- Now turn to data from the other three sites: Oakland, LA, San Diego

- Use six-quarter surrogate index estimated in Riverside and ask how well it performs in predicting heterogeneity in treatment effects across sites
    - Joint test of surrogacy and comparability assumptions

Surrogate Index Estimates vs. Actual Experimental Estimates, by Site

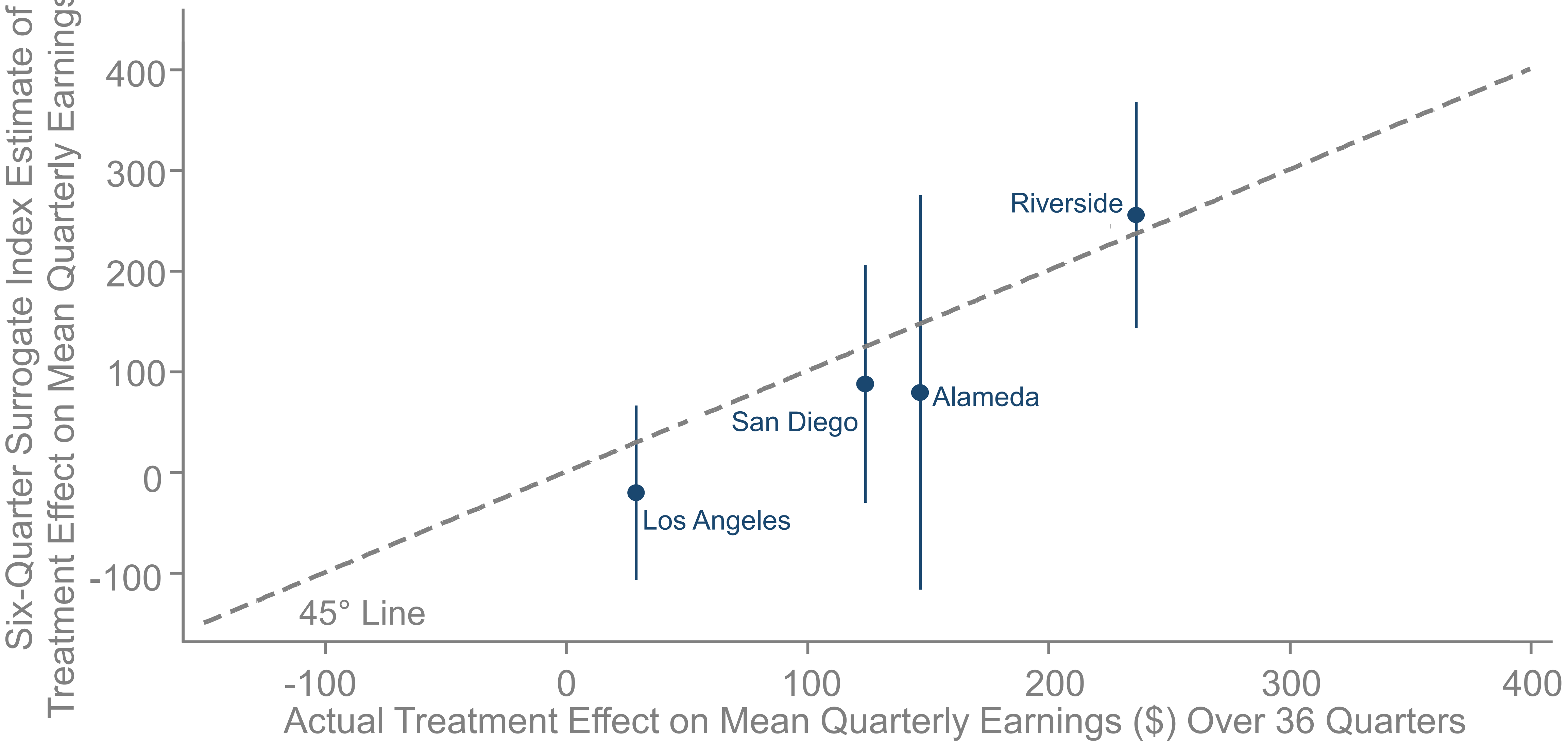
Mean Employment Rate over Nine Years



Note: Surrogate Index Estimates are based on a Six-Quarter Surrogate Index Estimated Using Data from Riverside

Surrogate Index Estimates vs. Actual Experimental Estimates, by Site

Mean Employment Rate over Nine Years



Note: Surrogate Index Estimates are based on a Six-Quarter Surrogate Index Estimated Using Data from Riverside

Conclusion

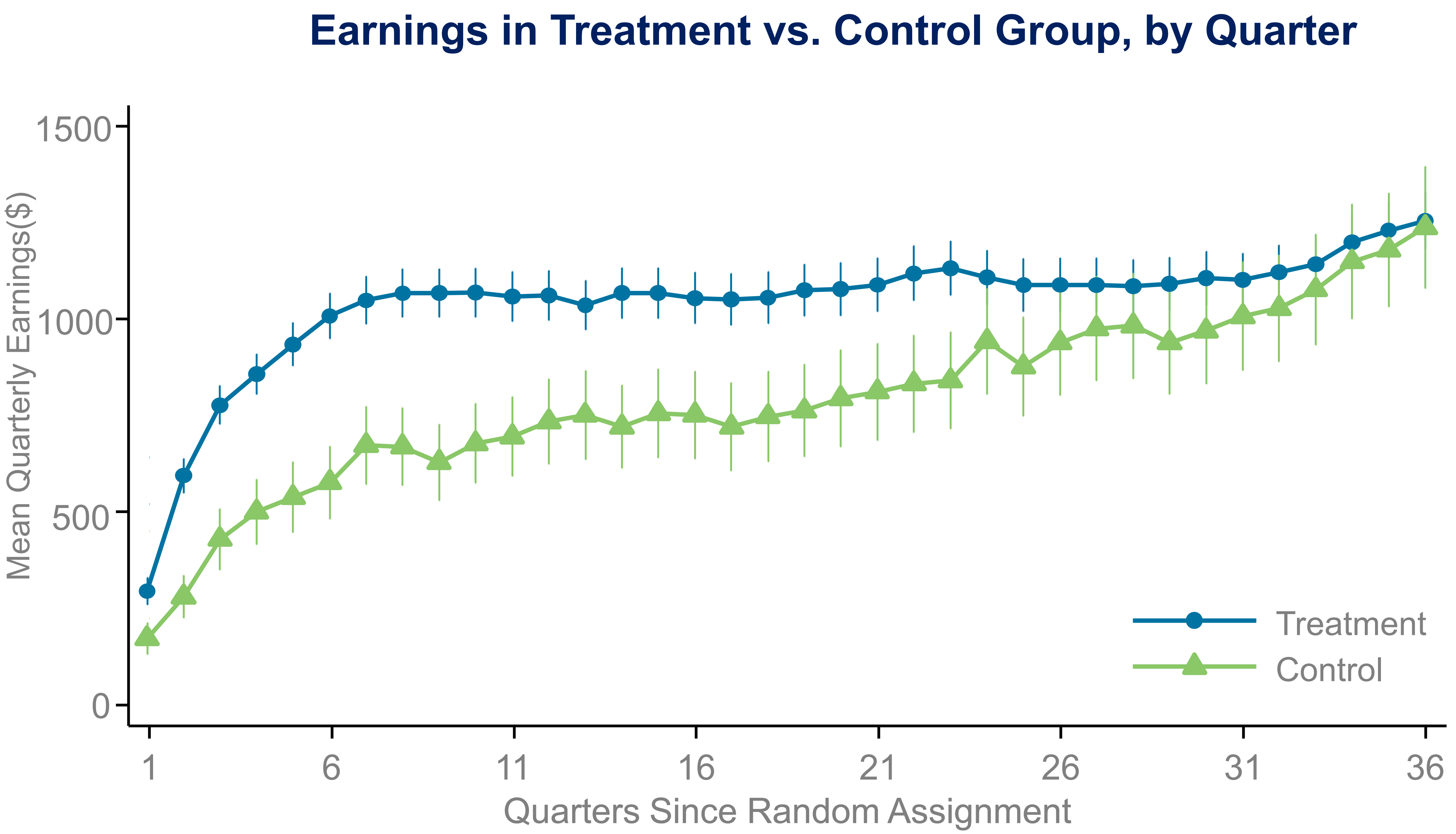
- Surrogate indices can be used to expedite and improve the precision of estimation of long-term treatment effects under empirically plausible assumptions

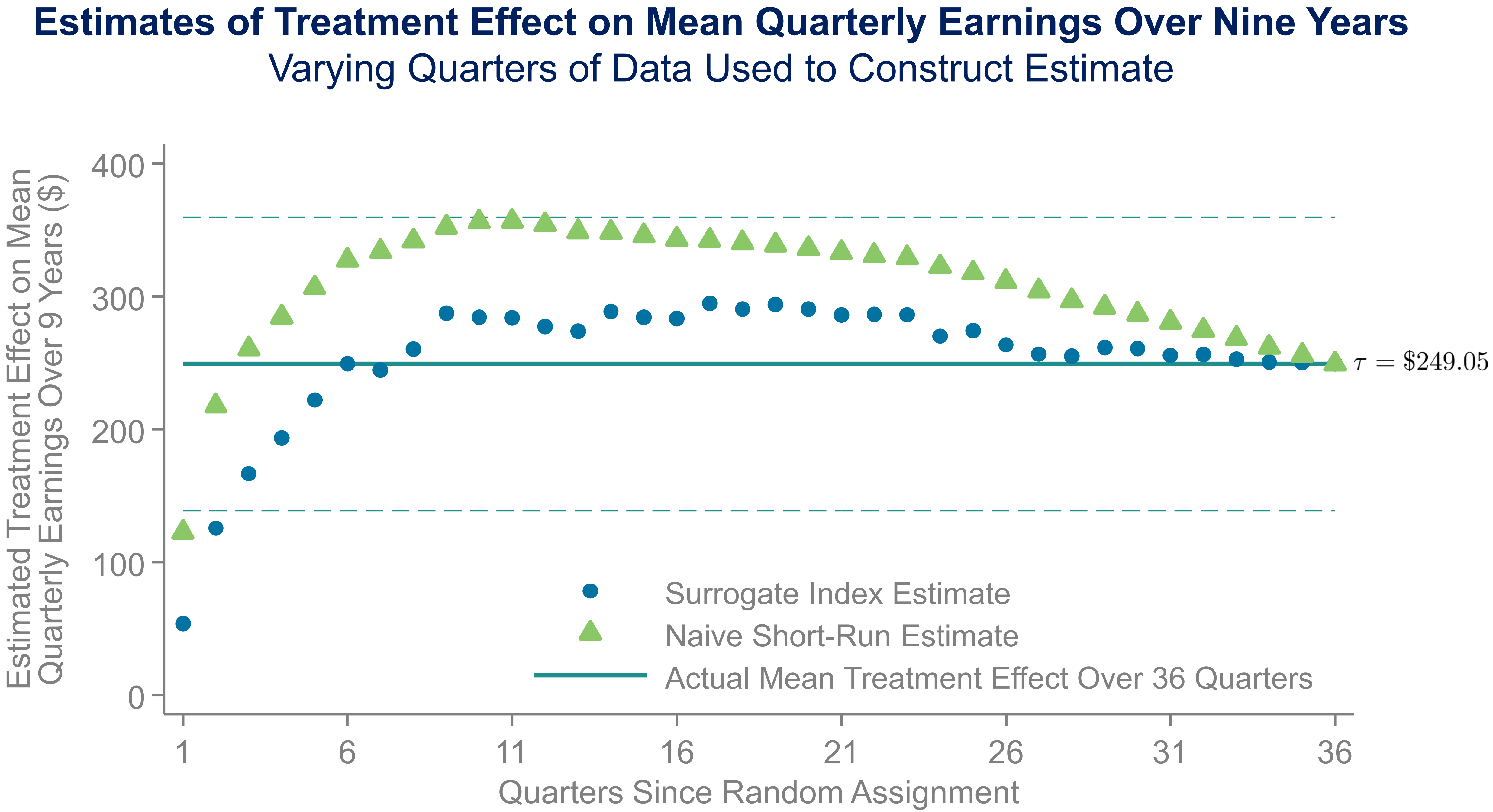
  - Impacts of economic programs on lifetime earnings to early childhood interventions on health to marketing impacts on downstream revenue

Future Work: Building a Surrogate Library

- Over time, we can develop guidance on which surrogates are adequate by analyzing other experiments, as we did across sites in the GAIN job training program
  - Ex: how many years of earnings, college attendance, other measures are needed to reliably predict lifetime income?
- Identifying surrogates that match long-term outcomes in existing/ongoing empirical studies would help us build a "surrogate library"
  - These surrogate indices can then be used in future work to increase precision and speed of program evaluation

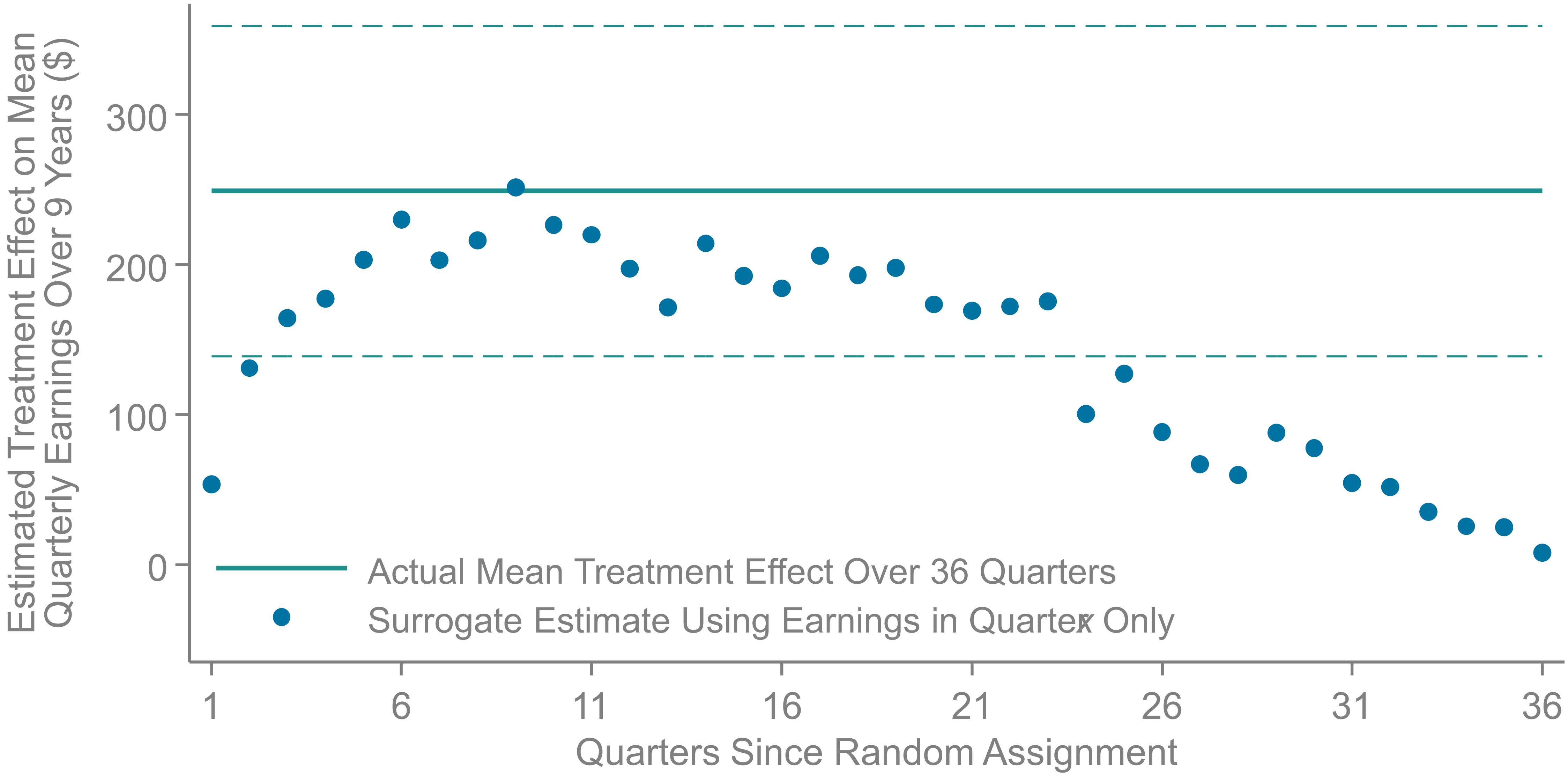
Supplementary Results

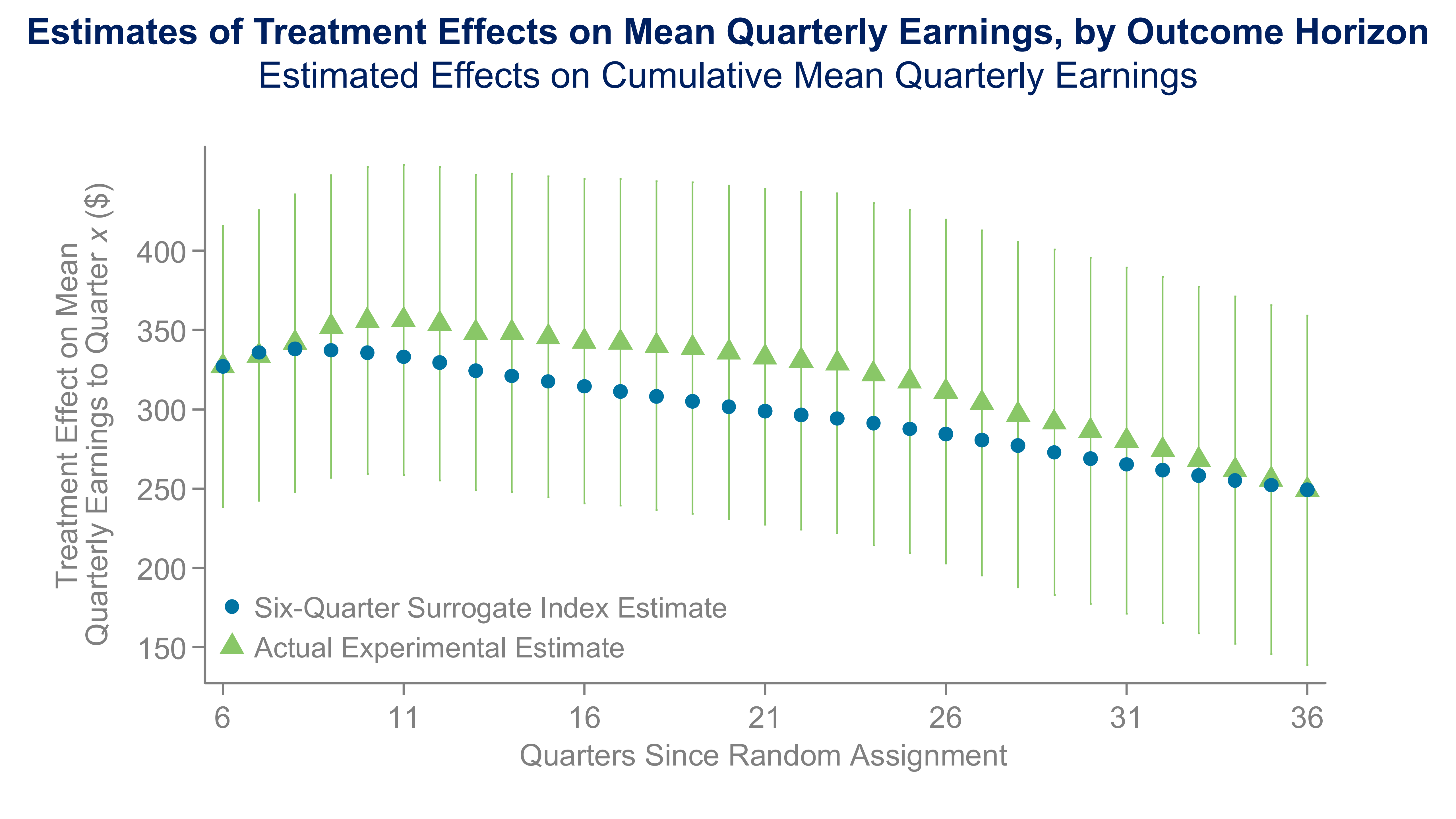

Earnings in Treatment vs. Control Group, by Quarter
Mean Quarterly Earnings($)
1500
1000
500
0
1
6
11
16
21
26
31
36
Quarters Since Random Assignment
Treatment
Control



Estimates of Treatment Effect on Mean Quarterly Earnings Over Nine Years
Varying Quarters of Data Used to Construct Estimate
Estimated Treatment Effect on Mean Quarterly Earnings Over 9 Years ($)
0
100
200
300
400
1
6
11
16
21
26
31
36
Quarters Since Random Assignment
$\tau = \$249.05$
Surrogate Index Estimate
Naive Short-Run Estimate
Actual Mean Treatment Effect Over 36 Quarters

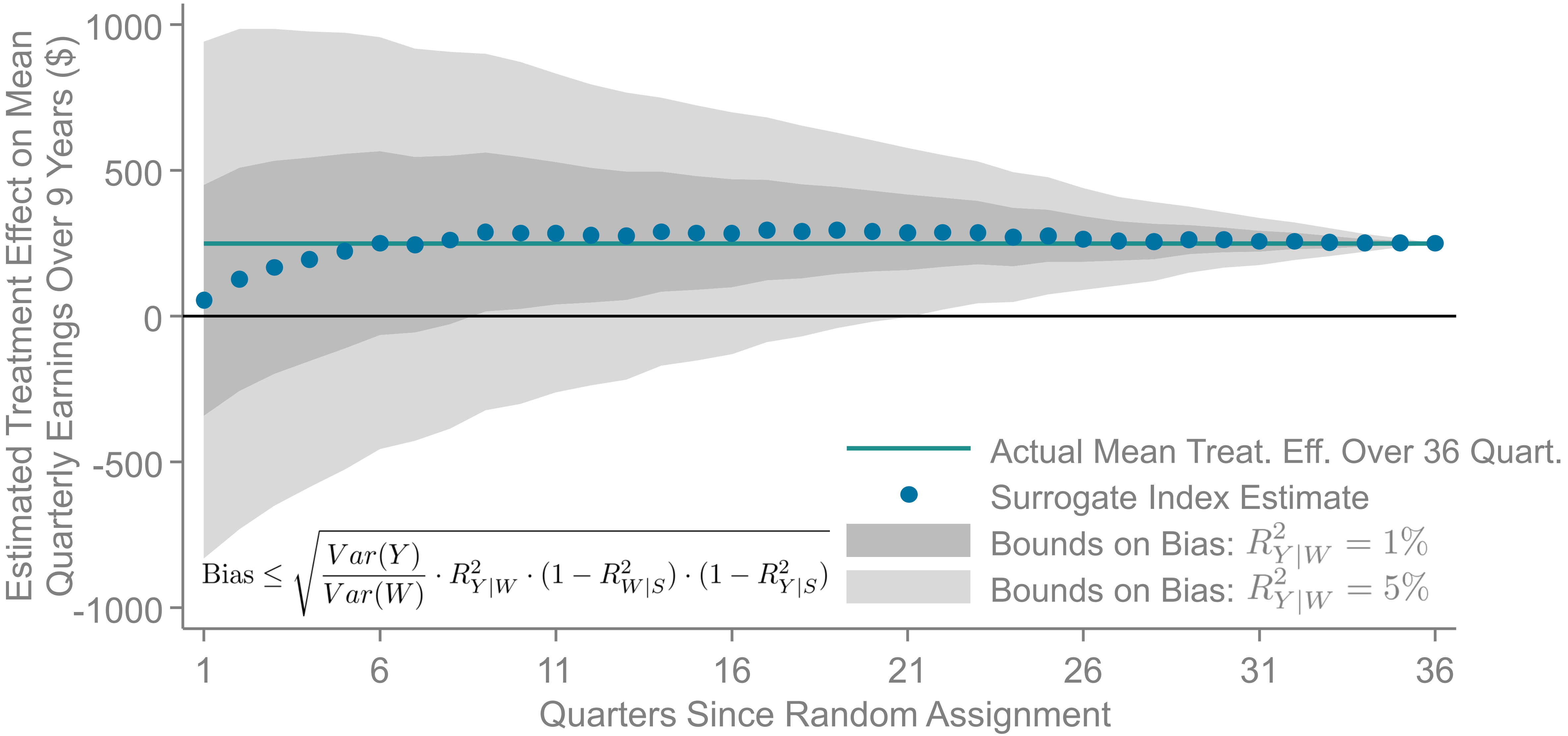
Estimates of Treatment Effect on Mean Quarterly Earnings Over Nine Years
Using Earnings in a Single Quarter as a Surrogate





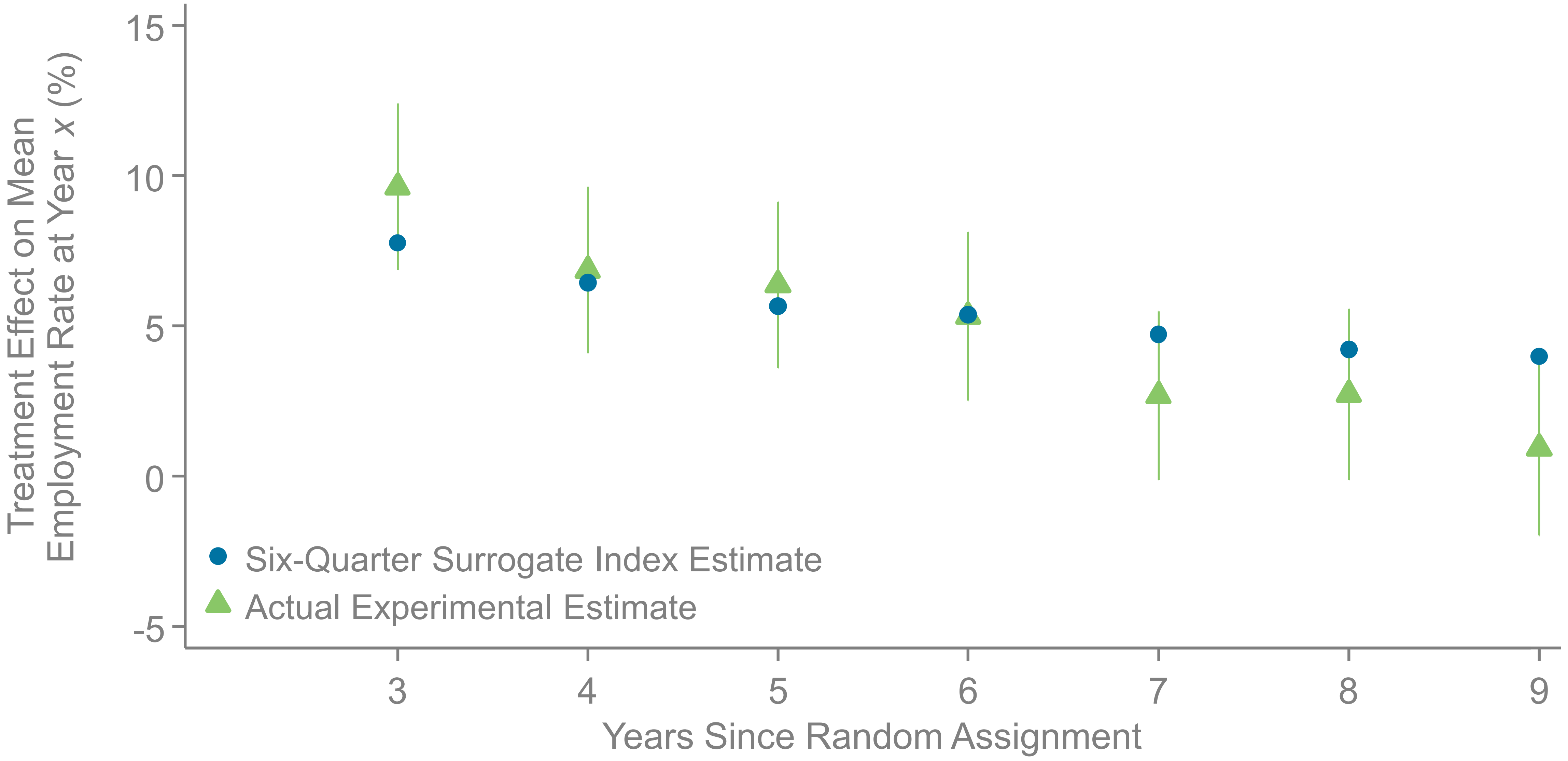
Estimates of Treatment Effects on Mean Quarterly Earnings, by Outcome Horizon
Estimated Effects on Cumulative Mean Quarterly Earnings
Treatment Effect on Mean Quarterly Earnings to Quarter x ($)
150
200
250
300
350
400
6
11
16
21
26
31
36
Quarters Since Random Assignment
Six-Quarter Surrogate Index Estimate
Actual Experimental Estimate

Bounds on Mean Treatment Effect on Earnings Based on Surrogate Index
Varying Number of Quarters Used to Estimate Surrogate Index



Estimates of Treatment Effects on Mean Employment Rates by Year

Actual Estimates by Year vs. Six-Quarter Surrogate Index Estimate



Estimates of Treatment Effects on Mean Quarterly Earnings by Year

Actual Estimates by Year vs. Six-Quarter Surrogate Index Estimate