Experimentally Validated Digital Twins

Evaluating Workforce Training Programs in the Age of Big Data

Raj Chetty, Harvard University and Opportunity Insights
Jamie Fogel, Harvard University and Opportunity Insights
Lawrence F. Katz, Harvard University and Opportunity Insights
Kadeem Noray, Harvard Business School
Sonya Porter, U.S. Census Bureau
James Reisinger, Meta

November 2026

We are very grateful to Ruhi Baichwal, Corey Gelb-Bicknell, Juhui Jin, Josephine Kim, and Samir Rein for exceptional research assistance. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed the results for disclosure avoidance protection (Project 7519874: CBDRB-FY22-CES014-006, CBDRB-FY23-CES004-039, CBDRB-FY25-CES027-004, CBDRB-FY25-CES027-008, CBDRB-FY25-CES027-010, CBDRB-FY26-CES004-005, CBDRB-FY26-CES004-007, CBDRB-FY26-CES027-005, CBDRB-FY26-CES027-002, CBDRB-FY26-CES027-003, CBDRB-FY26-CES027-004, CBDRB-FY26-CES030-007).

OPPORTUNITY INSIGHTS



Source: Workforce Almanac

5,000+

job training programs in America in 2022

Jewish Vocational Services Bay Area
Western Governors University
Texas Association of Community Colleges
General Assembly
Lorain County Community College
Management Leadership for Tomorrow
Sponsors for Educational Opportunity (SEO)
Job Corps
TechHire
Project Quest
Upwardly Global
Apprenti
Colorado Mountain College
Year Up
PowerCorpsPHL
St. Nick's Alliance
Urban Ed
Madison Strategies
Jewish Vocational Services Boston
Activate Work
Tech Impact
First Step Staffing
Vehicles for Change
Clinical Research Fastrack
ForgeNow
WRTP
JobTrain
Towards Employment
Cross Purpose
Free World
TWP Platform to Employment
NPower
NetWerk®
TWP Health Career Academy
JVS Bay Area
Generation USA
The Mission Continues
Valencia College
PeopleShores
EWIB
Career Resources Inc Reentry Employment
Per Scholas
i.c.Stars
Learning Alliance Corporation
SkillUp
Activate Work
Great Jobs KC
TechBridge
CareerWise USA
Creating Coding Careers
Upwardly Global
Merit America
Tech Impact
CWP BEST Chance Program
Rogue Community College
Dallas College
Pursuit
League for Innovation in the Community College
Community Learning Center Inc.

Source: Workforce Almanac

***Evaluated Using Randomized Trial**

How Can We Evaluate Program Impacts at Scale?

- Randomized trials provide gold standard evidence
- But it is infeasible to run RCTs for every program
- One practical alternative: use observational (non-randomized) data to compare program participants to similar non-participants

History of Observational Estimators

- Many attempts to develop observational estimators over the past 40 years
  [e.g., Ashenfelter & Card 1985, LaLonde 1986, Heckman et al. 1998, Dehejia and Wahba 1999]
  - General conclusion: observational estimates (matching, reweighting, regression) are unstable and often biased relative to estimates from RCTs [Imbens and Xu 2025]

- Prior work on observational estimators used small datasets with limited information
  - Ex: canonical LaLonde (1986) study sought to match job training participants to similar individuals in Current Population Survey (N = 16k)
  - Problem: difficult to find close matches on key dimensions (e.g., exact age, employment history, location)

Observational Estimators in the Age of Big Data

- Today, we have much larger administrative datasets with richer information

- Can we develop observational methods that work better in the age of big data using machine learning models?



Goals of This Project

1. Develop machine learning estimators of workforce training programs' impacts

    - Measure bias in these estimates by comparing them to existing RCT estimates

2. Apply observational estimator to estimate impacts of programs where no trial exists

    - Use estimates to learn about mechanisms that make workforce training programs effective

Outline

1. Observational Estimator

2. Experimental Validation

3. Estimating Treatment Effects Beyond the Experimental Sample

4. Mechanisms Underlying Effective Training Programs



Part 1: Observational Estimator



Data Sources

1. Tax + Census records covering the U.S. population (~327 million adults) from 1999-2025

    - Earnings measured from W-2 forms filed by employers; all monetary values reported in this talk are in 2025 (inflation-adjusted) dollars

    - Other information from 1040 returns, information returns, and decennial Census: age, sex, Census tract, employers, industry, college attended, investment income

2. Enrollment data from workforce training and higher education programs linked to Tax + Census data using secure protocols by Census Bureau staff

→ Linked longitudinal analysis dataset that is stripped of identifiers and anonymized

Estimating Treatment Effects in Observational Data

- Target: average treatment effect of a program on the treated (ATT)

$$\tau = \mathrm{E}[Y_i(1) \mid D_i = 1] - \mathrm{E}[Y_i(0) \mid D_i = 1]$$

- Fundamental problem of causal inference: counterfactual (untreated) earnings, $\mu_0 = \mathrm{E}[Y_i(0) \mid D_i = 1]$, is not observed

Estimating Counterfactual Earnings from Observational Data

- Estimate counterfactual μ_0 using non-participants with same pre-enrollment characteristics X_i as participants
  - First estimate a model of earnings outcomes among non-participants:

$$m_0(x) = \mathrm{E}[Y_i \mid X_i = x, D_i = 0]$$

  - Then estimate counterfactual using prediction from outcome model:

$$\mu_0 = \mathrm{E}[m_0(X_i) \mid D_i = 1]$$

Estimating Treatment Effects in Observational Data

- Observational estimator of treatment effect:

$$\tau^{OBS} = \mathrm{E}[Y_i \mid D_i = 1] - \mathrm{E}[m_0(X_i) \mid D_i = 1]$$

- Under selection on observables and overlap, $\tau^{OBS} = \tau$
  [Rosenbaum and Rubin 1983]
  - Overlap holds in our setting because we use full U.S. population data
  - Selection on observables may not hold; key goal of the paper is to measure the extent to which it fails

Estimating the Outcome Model

- Simplest estimator of outcome model $m_0(X_i)$: exact matching on observable characteristics
  - Not feasible: 5 years of earnings deciles, 2 years of 4-digit industry, county gives 10 trillion possible cells
- Instead use a flexible machine learning estimator
  - Gradient-boosted trees (LightGBM) trained on non-participants, separately for each program × site × enrollment-year cell

Plug-In ML Estimator

- Plug-in estimator: estimate program impacts as participants' average realized earnings minus predicted earnings without the program

$$\hat{\tau} = \frac{1}{N_1} \sum_{i:D_i=1} \{Y_i - \hat{m}_0(X_i)\}$$

Bias in Plug-In ML Estimator

- Regularization in ML model shrinks predictions $\hat{m}_0$ toward the mean, generating bias [Chernozhukov et al. 2018]

- Measure this bias using average prediction errors among non-participants, reweighted to match treated group using a propensity score model:

$$\frac{1}{N_1} \sum_{i:D_i=0} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} \{Y_i - \hat{m}_0(X_i)\}$$

- Estimate predicted probability of program participation $\hat{e}(X_i)$ using gradient boosted tree

Double/Debiased Machine Learning (DML) Estimator

- DML estimator combines outcome and propensity model estimates:

$$\hat{\tau}^{DML} = \frac{1}{N_1} \sum_{i:D_i=1} \{Y_i - \hat{m}_0(X_i)\} - \frac{1}{N_1} \sum_{i:D_i=0} \frac{\hat{e}(X_i)}{1 - \hat{e}(X_i)} \{Y_i - \hat{m}_0(X_i)\}$$

Outcome model estimate

Propensity model estimate

- DML is root-*N* consistent and asymptotically normal under selection on observables, overlap and regularity conditions [Chernozhukov et al. 2018]

- Use five-fold cross-fitting for both outcome and propensity models, with standard errors from cross-fitted influence scores

Covariates Used in DML Model

- Covariates X_i, all measured before enrollment in training program:
  - **Work history**: earnings, employment, and industry in five years preceding enrollment
  - **Demographics**: age, sex, marital status, children
  - **Residential history**: state and county indicators and Census tract characteristics (poverty, unemployment, HS graduation rates, racial composition, test scores)



Part 2: Experimental Validation

Case Study: Year Up Job Training Program

- Begin with a case study of Year Up job training program

- Revisit impacts of 2013-2014 randomized trial involving 2,544 people who applied to Year Up

Learn new skills to launch your career.

Year Up's job training program is tuition free, and offers access to today's best companies and a proven path to career success.

Zip Code

FIND TRAINING

Our approach

Three steps to a new job

1 Learn



The Learning & Development phase of our program provides you with skills based training from expert

2 Intern



Put the skills you learned into practice! We'll match you to a hands-on internship at a top

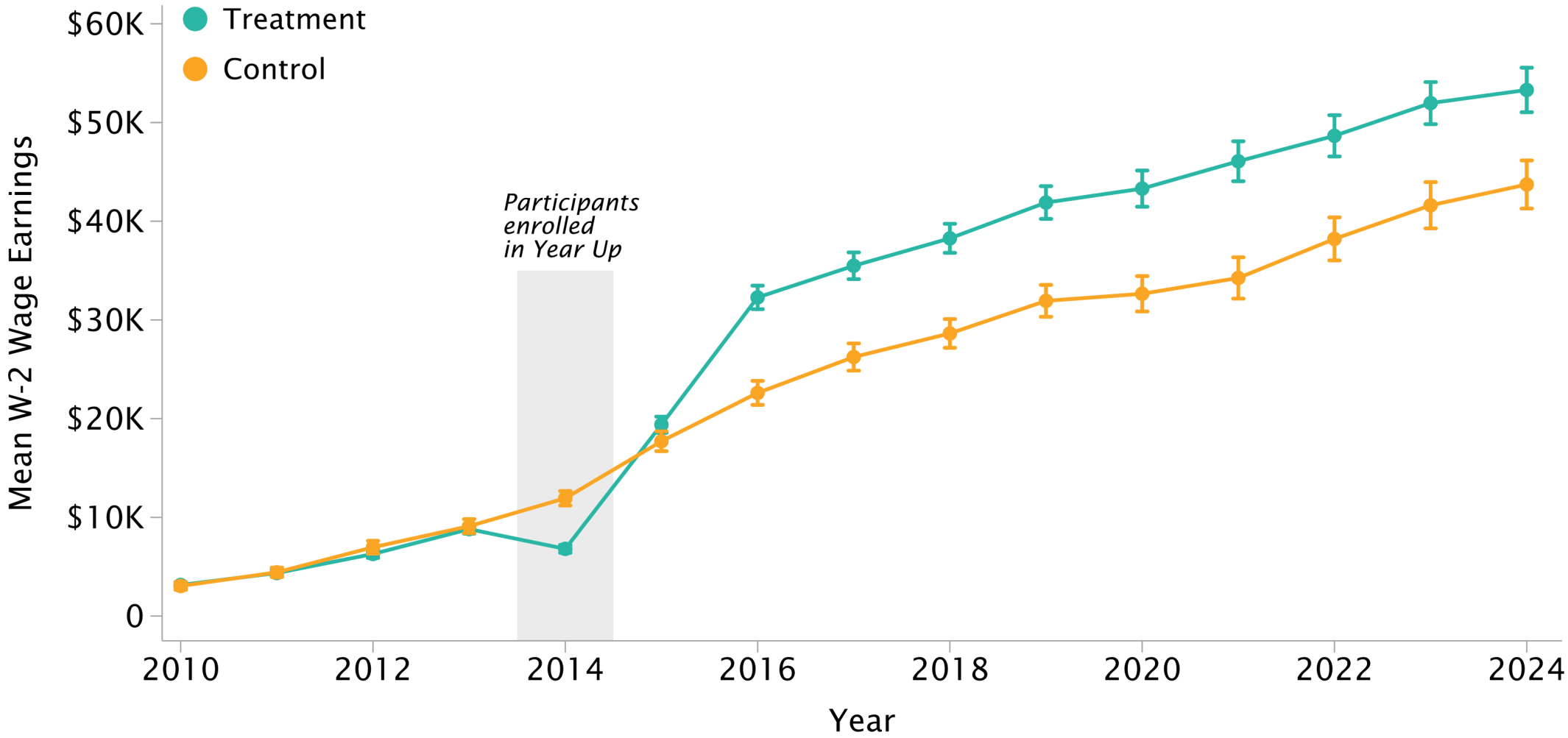
3 Land a job



80% of alumni are employed and/or enrolled in college within 4 months of graduation. Our support staff will

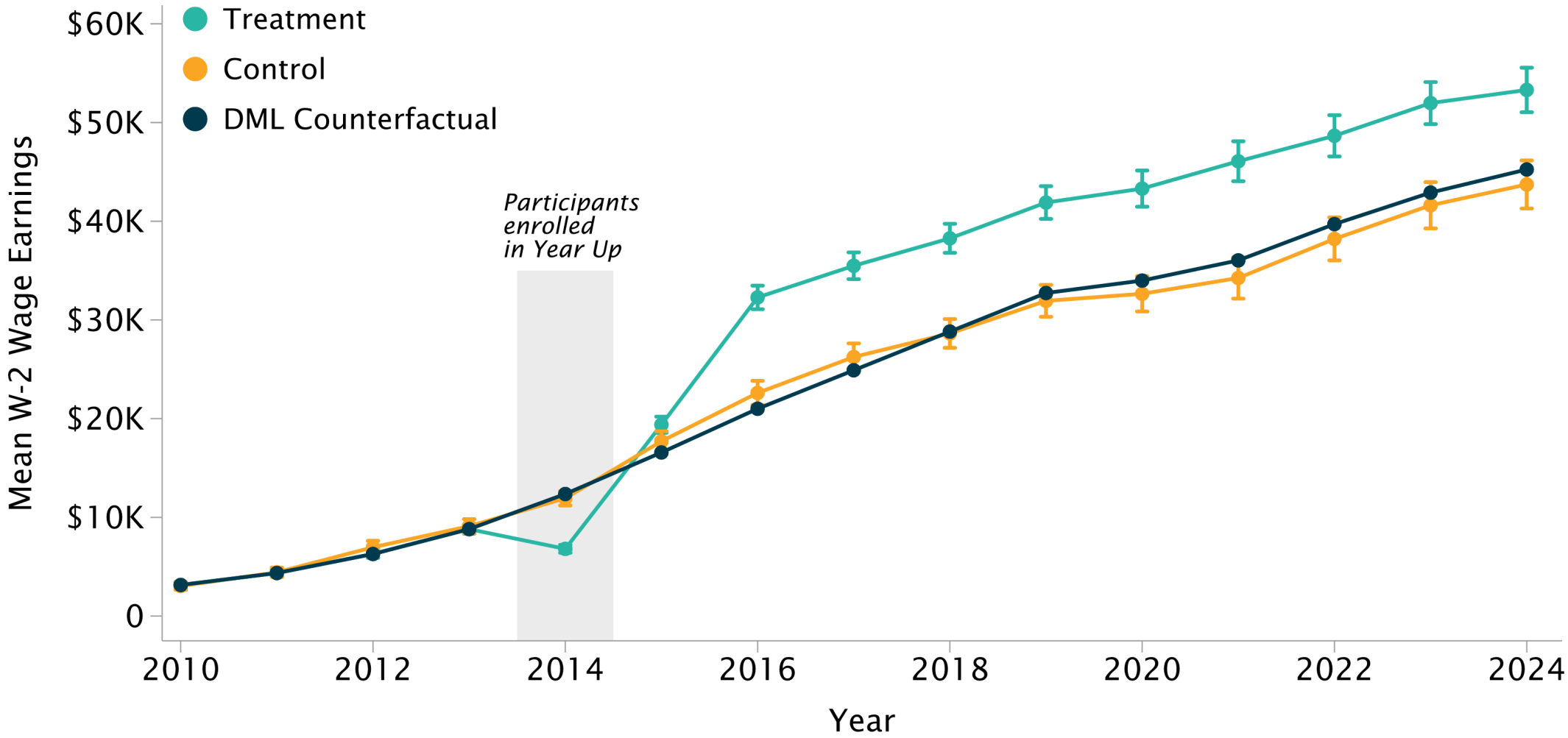
Effect of Year Up United Job Training Program on Earnings

Evidence from PACE Randomized Trial in 2014



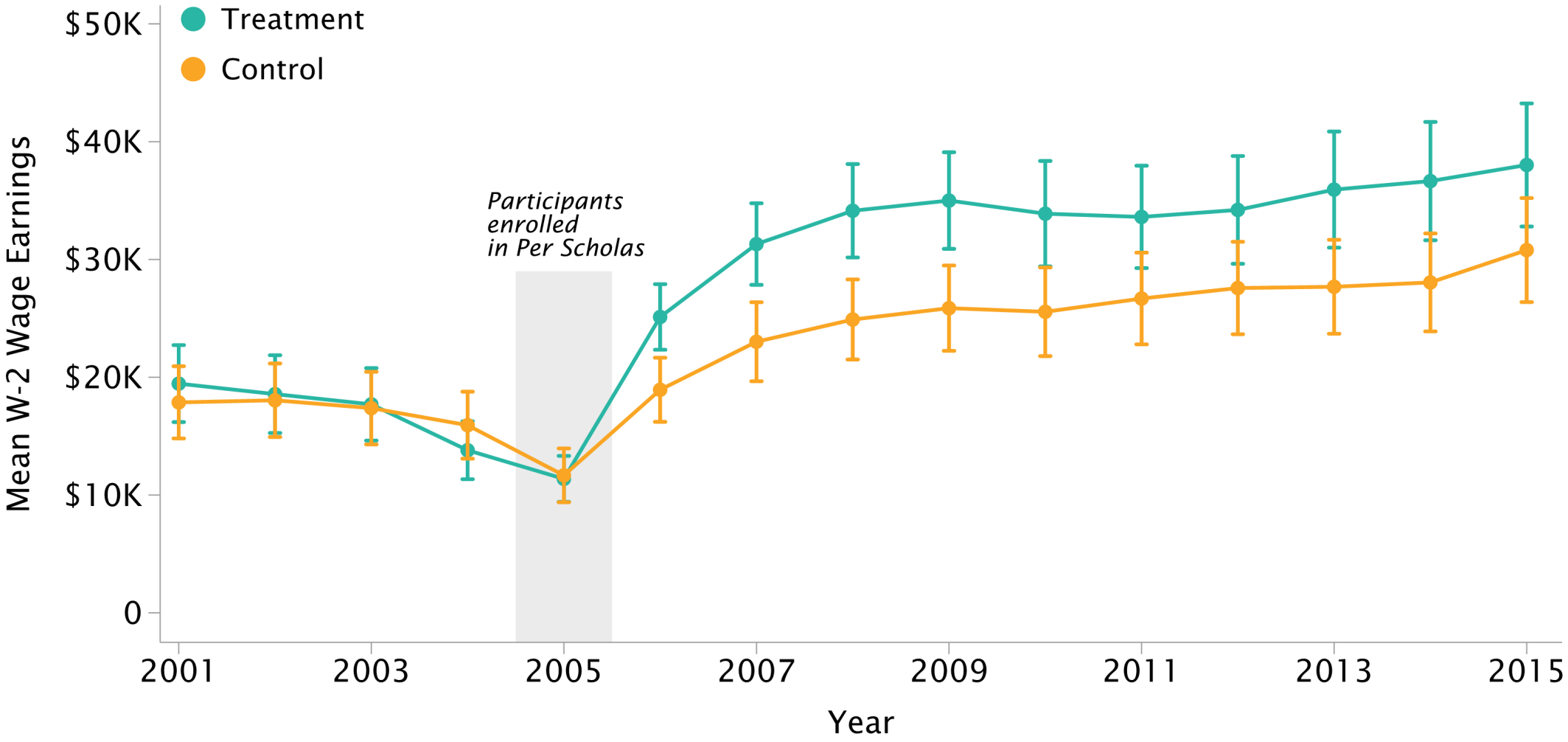
Effect of Year Up United Job Training Program on Earnings

Evidence from PACE Randomized Trial in 2014



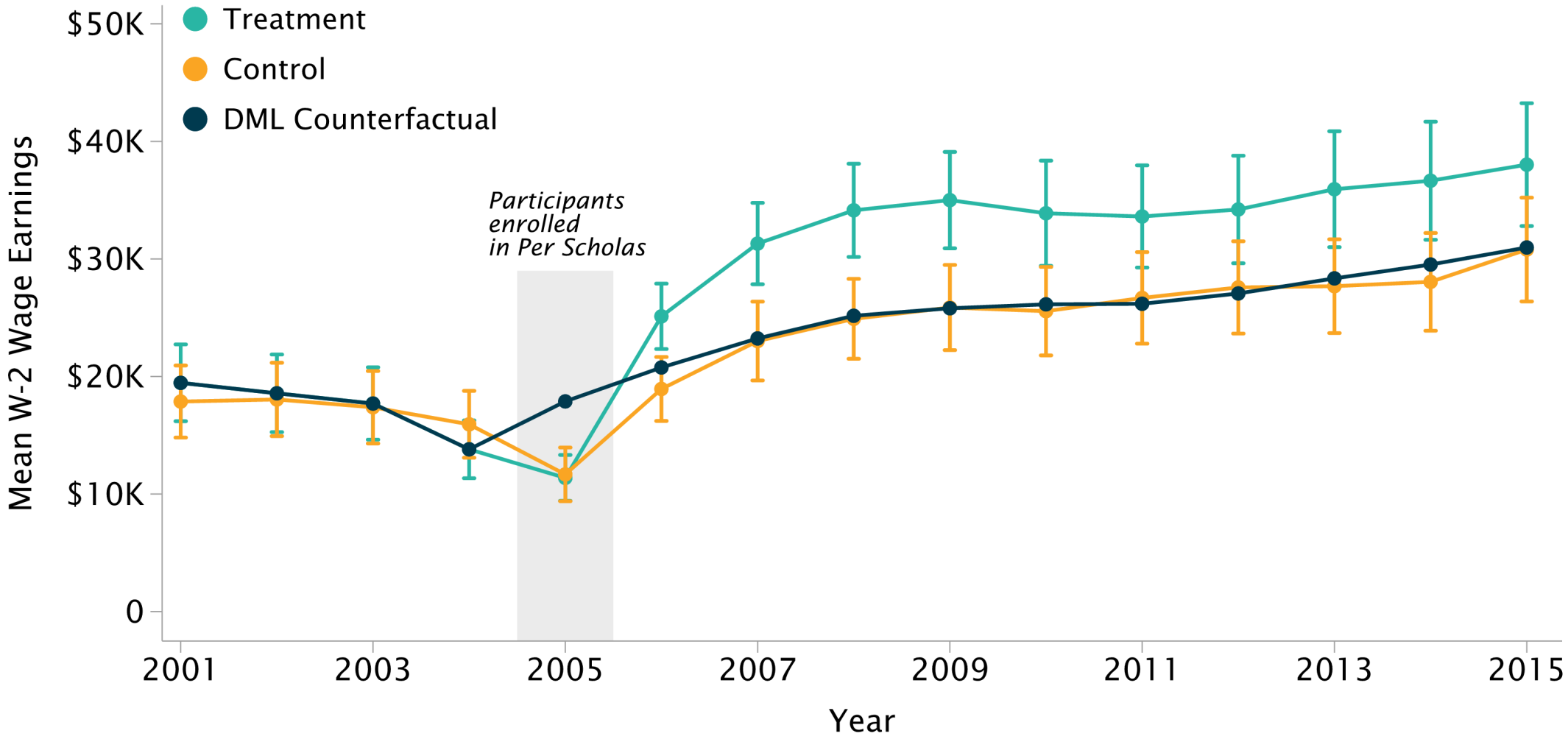
Effect of Per Scholas Job Training Program on Earnings

Evidence from SEIS Randomized Trial in 2005



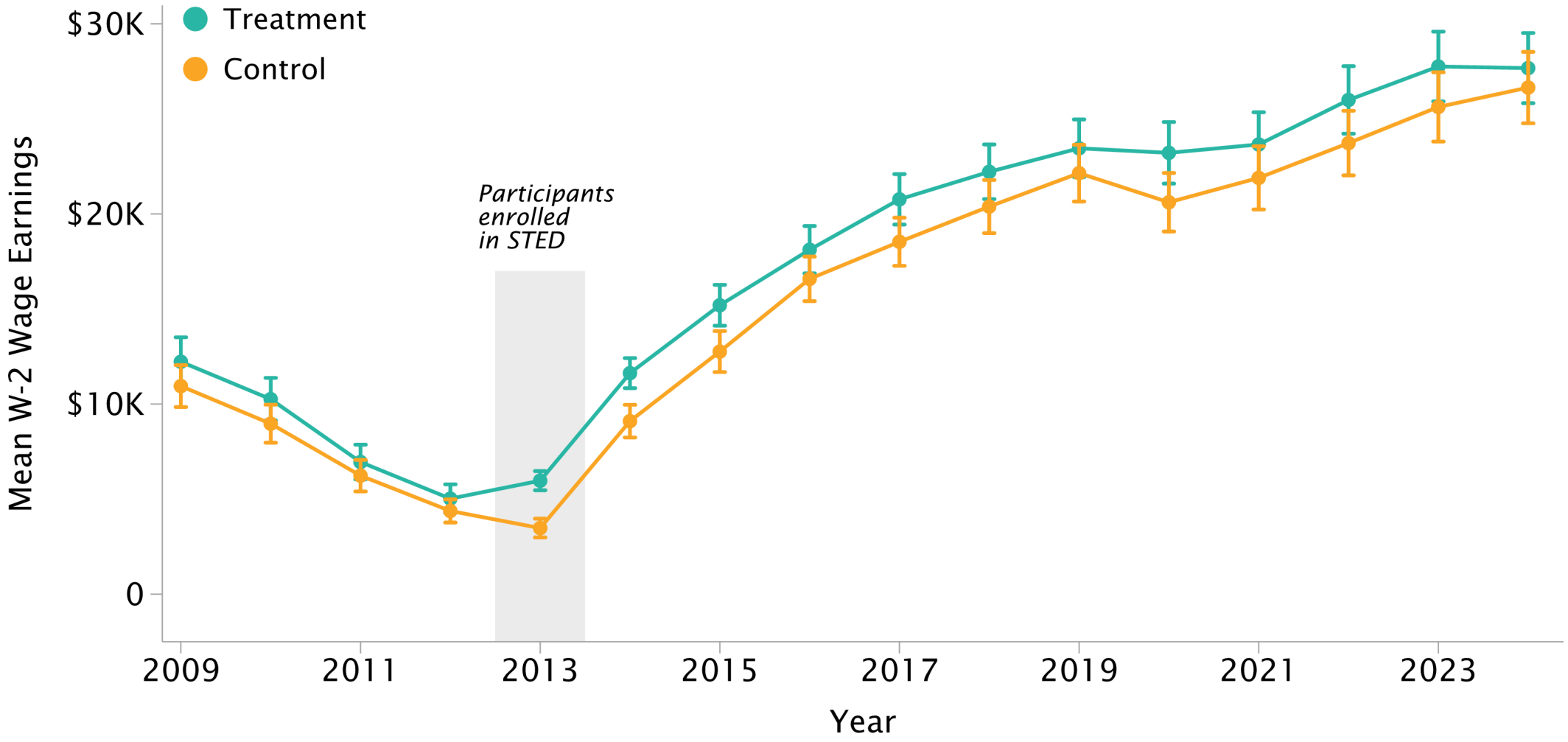
Effect of Per Scholas Job Training Program on Earnings

Evidence from SEIS Randomized Trial in 2005



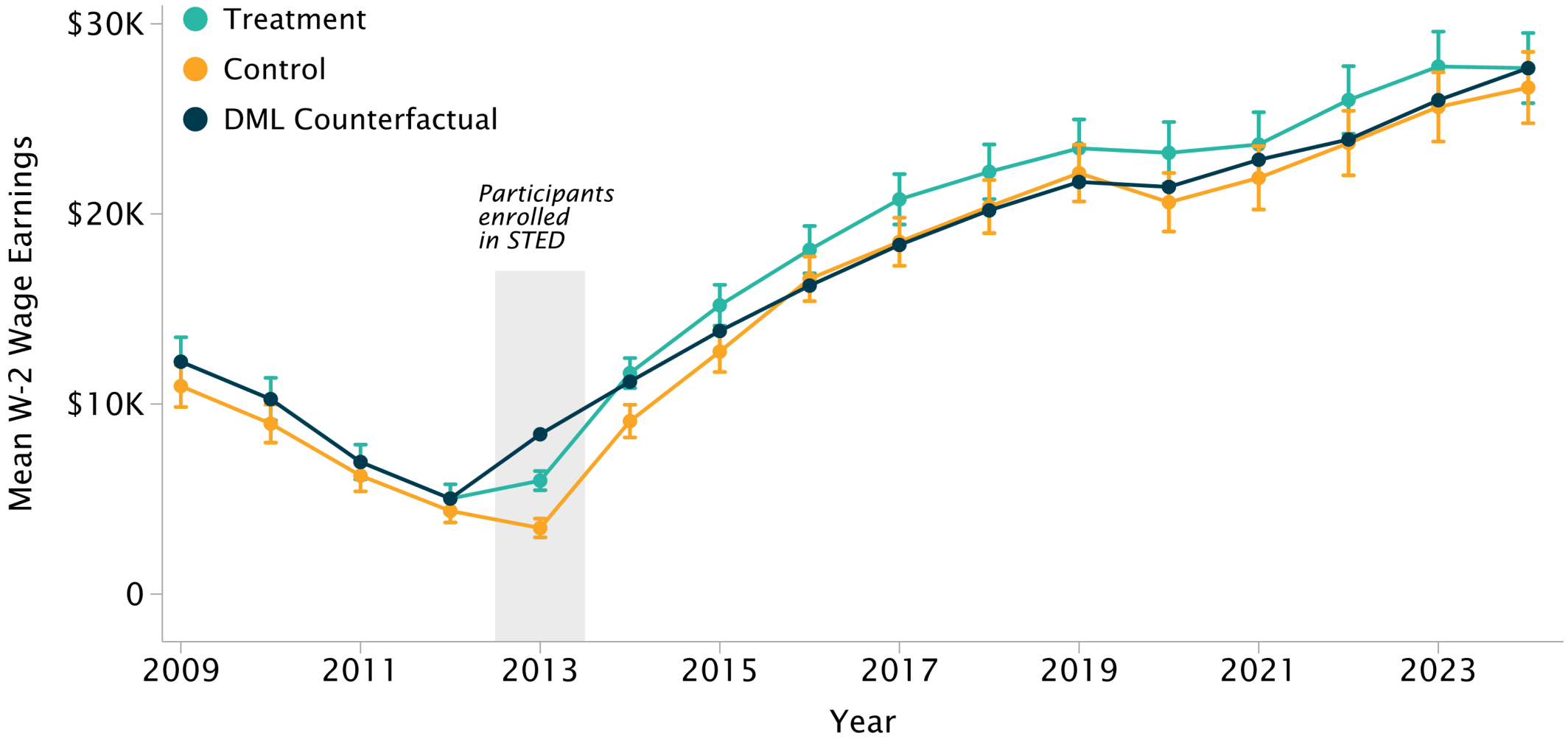
Effect of STED Program on Earnings

Evidence from LA Paid Work Experience Trial in 2013



Effect of STED Program on Earnings

Evidence from LA Paid Work Experience Trial in 2013

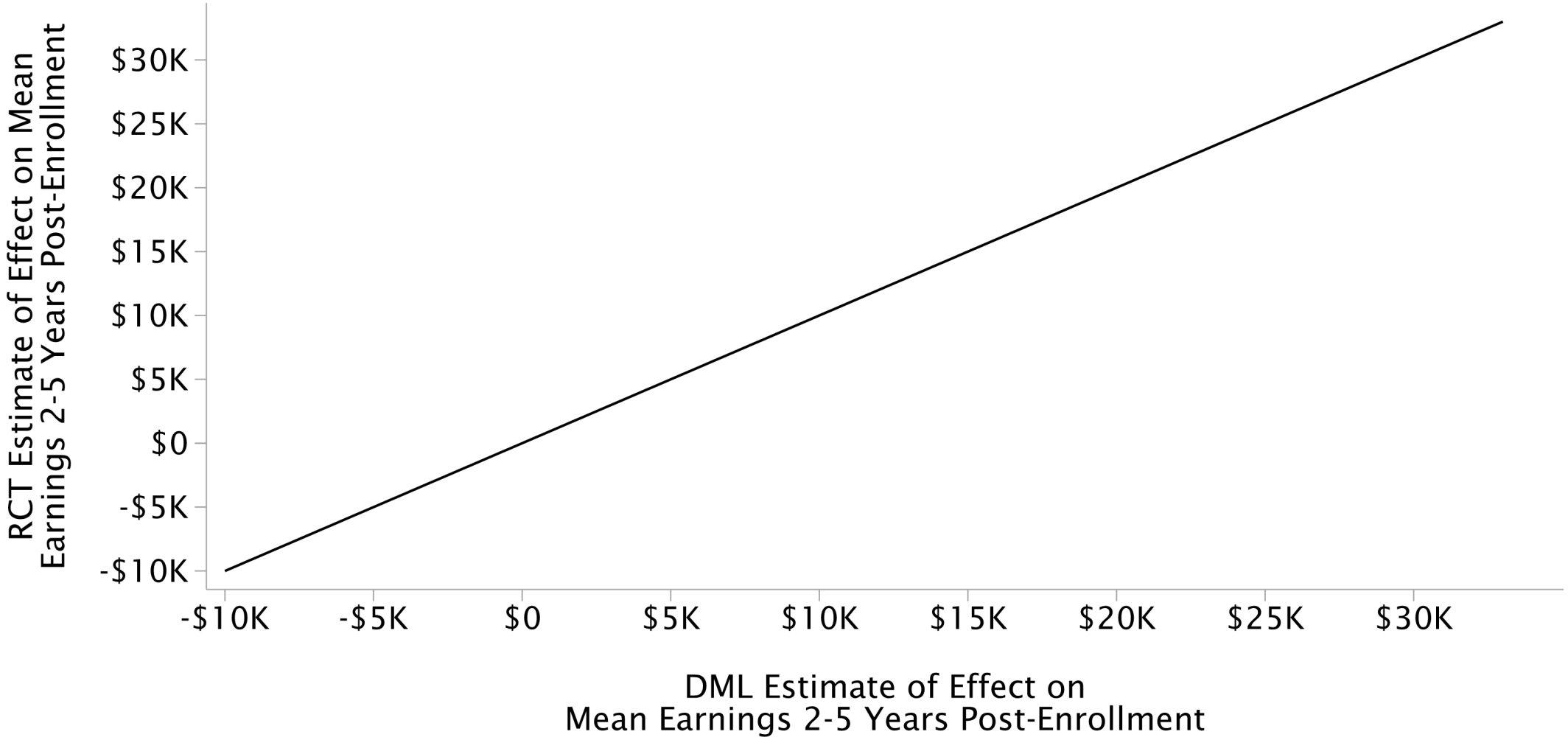


37 Randomized Trials of Job Training Programs Across America



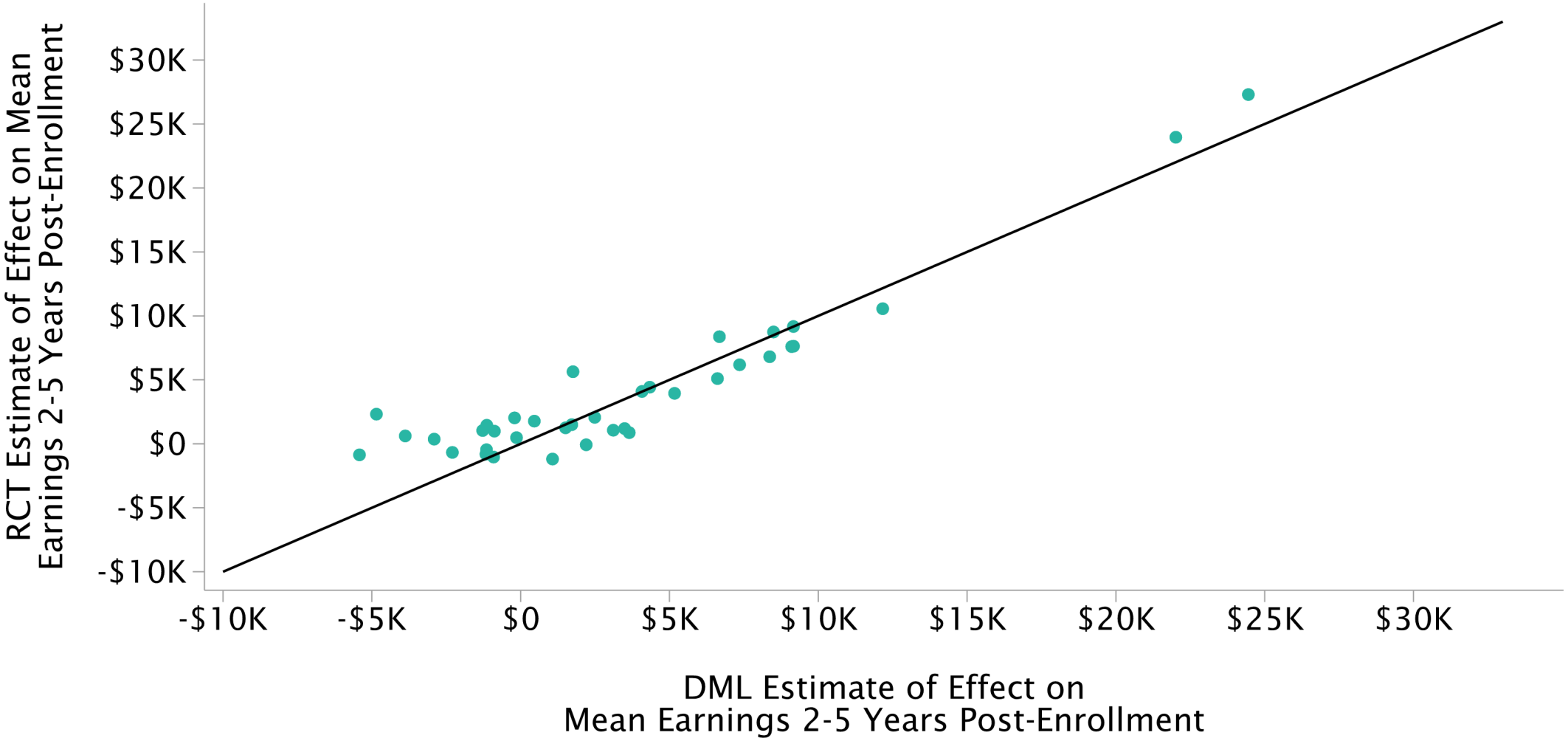
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



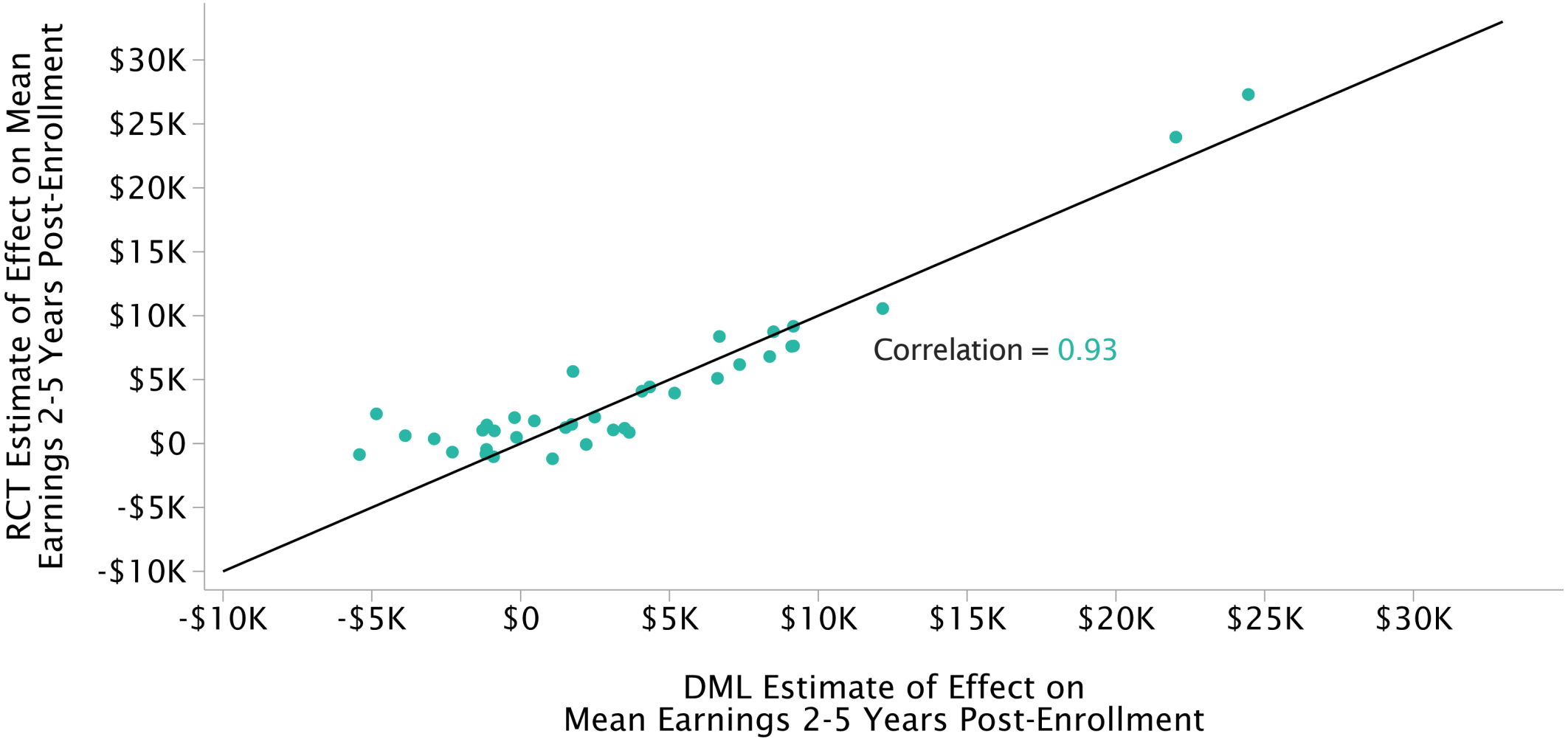
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



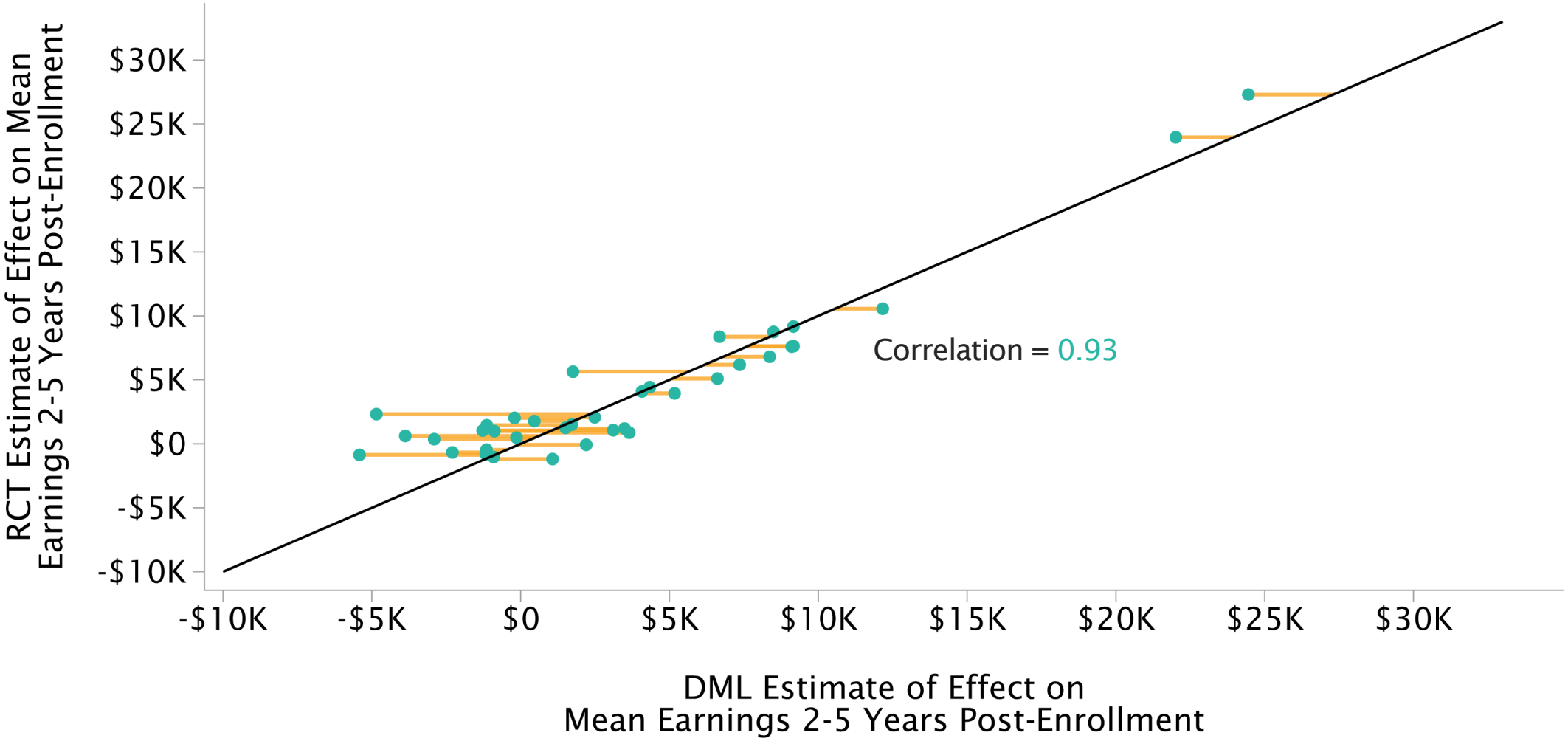
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



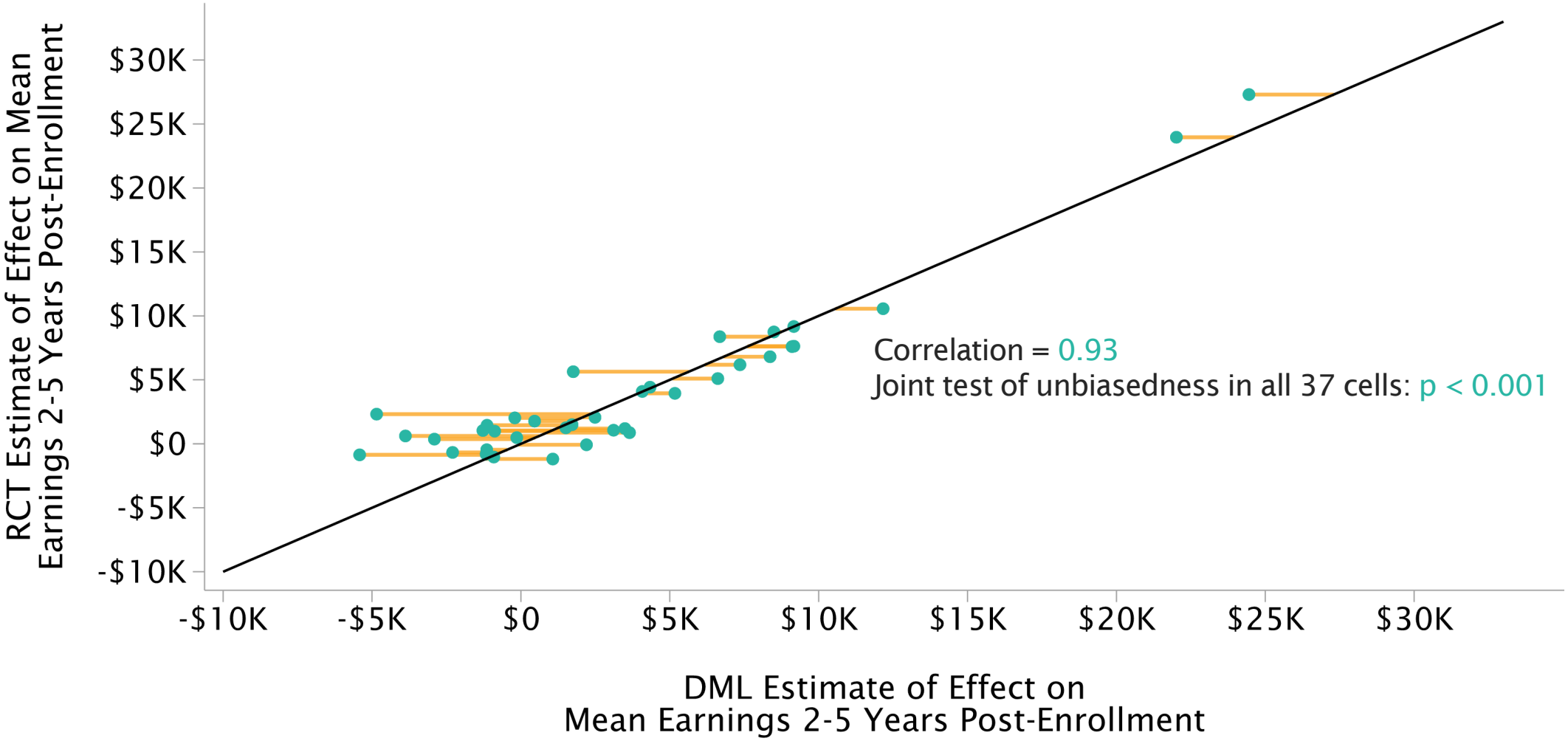
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



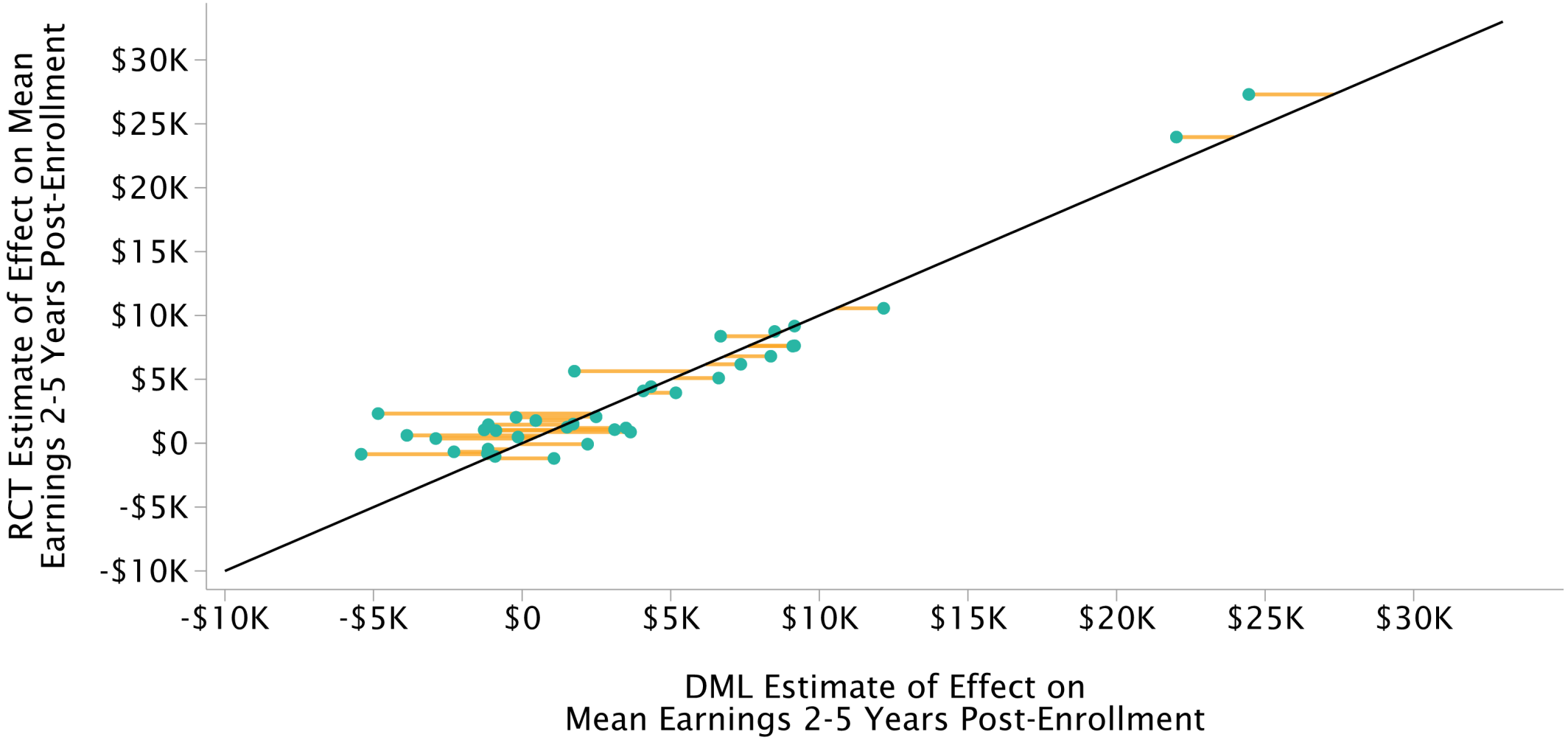
Evaluating Machine Learning Estimates: Comparisons to RCT Estimates

- Selection on observables fails: DML estimates are not unbiased

- Prior work typically focused on testing the null of no bias, but this may not be the policy-relevant question
  - Is the error in the DML estimates small enough that they remain useful for policy decisions?
  - Use difference between DML and RCT estimates to measure the distribution of errors

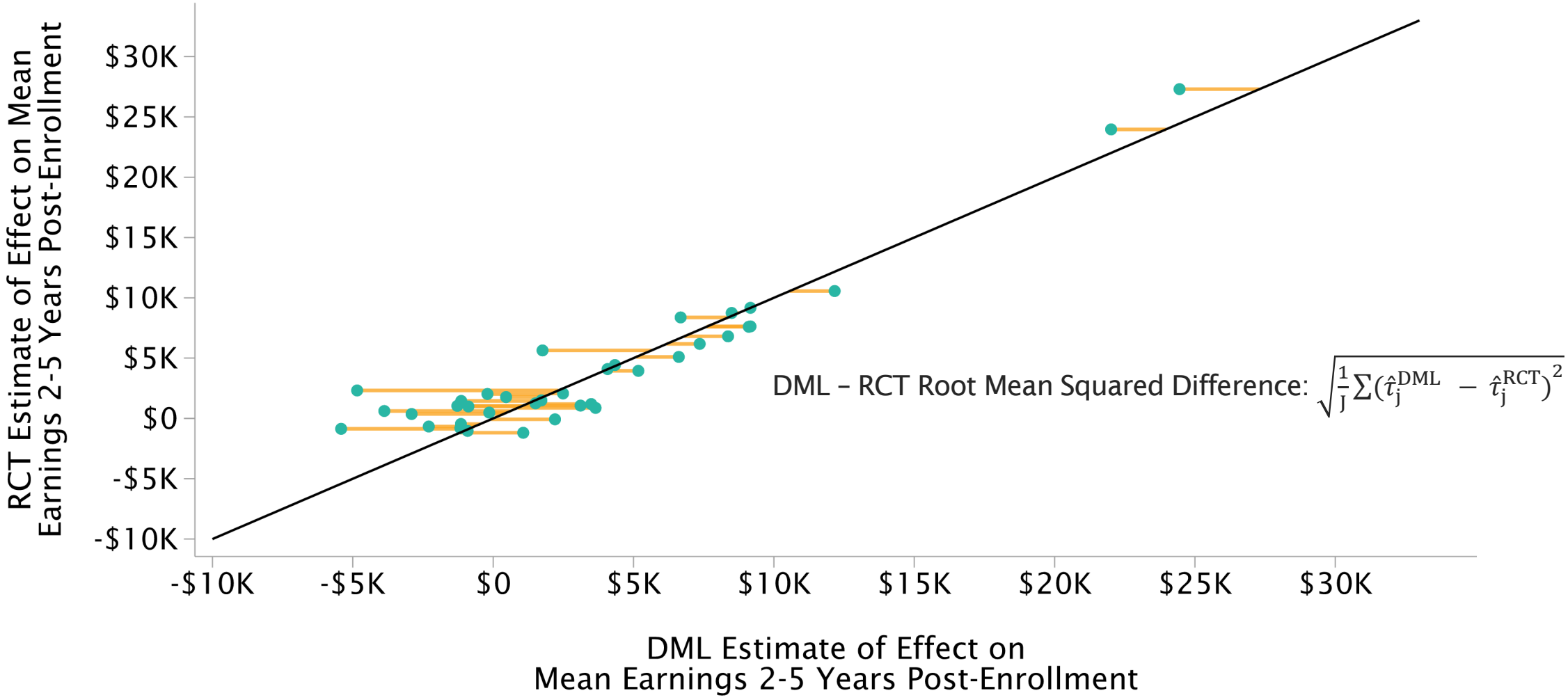
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



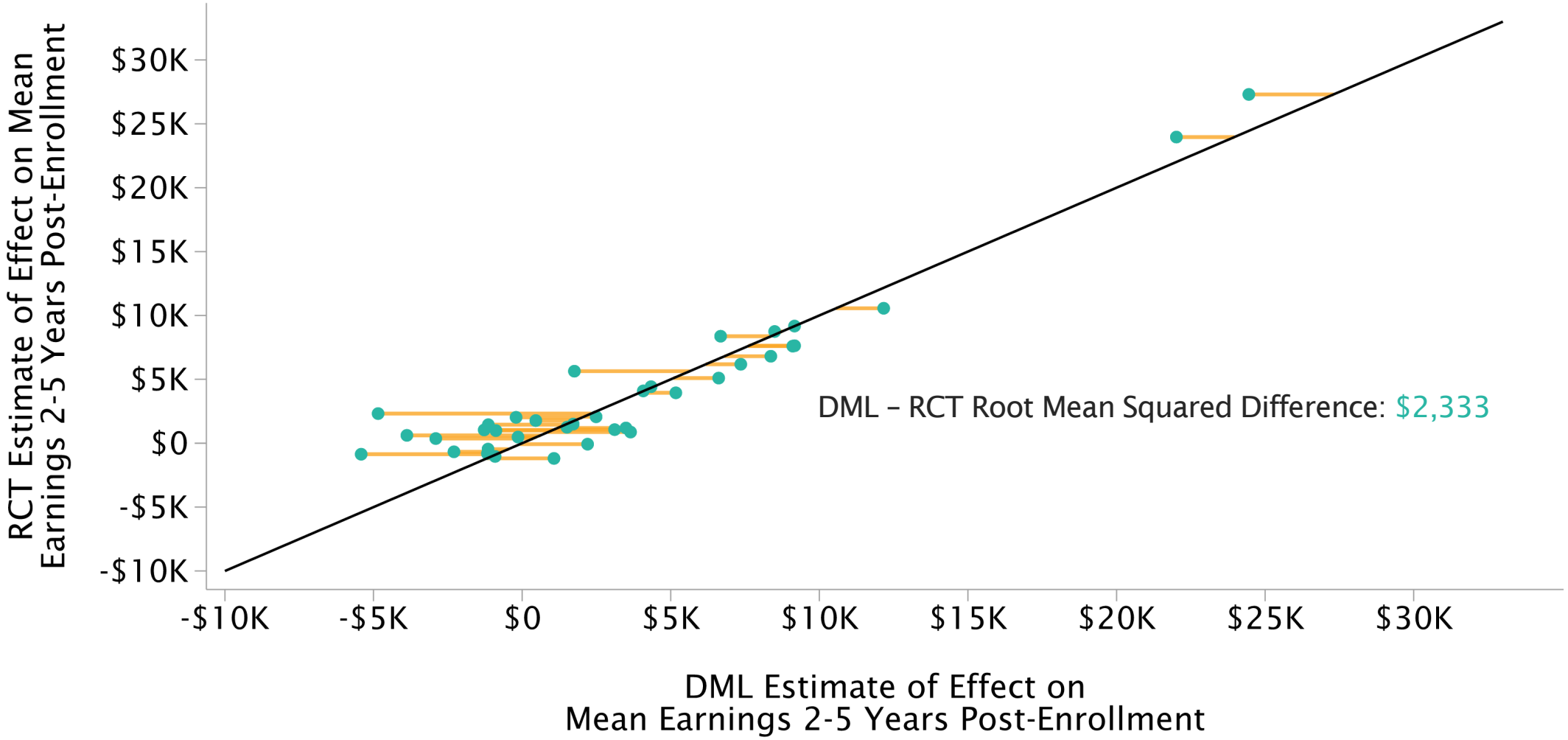
Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



Effects of Workforce Training Programs on Earnings

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



Measuring Error in the Observational Estimates

	Measure	Calculation	Estimate
(1)	Raw RMS DML–RCT difference		$2,333
(2)	Sampling noise in the RCT control mean		$1,366
(3)	Sampling noise in the DML counterfactual		$795

Sampling noise is the square root of mean sampling variance.

Measuring Error in the Observational Estimates

	Measure	Calculation	Estimate
(1)	Raw RMS DML–RCT difference		$2,333
(2)	Sampling noise in the RCT control mean		$1,366
(3)	Sampling noise in the DML counterfactual		$795
(4)	Noise-adjusted RMSE of DML estimator	$\sqrt{(1)^2 - (2)^2 - (3)^2}$	$1,717

Sampling noise is the square root of mean sampling variance.

Measuring Error in the Observational Estimates

	Measure	Calculation	Estimate
(1)	Raw RMS DML–RCT difference		\$2,333
(2)	Sampling noise in the RCT control mean		\$1,366
(3)	Sampling noise in the DML counterfactual		\$795
(4)	Noise-adjusted RMSE of DML estimator	$\sqrt{(1)^2 - (2)^2 - (3)^2}$	\$1,717
(5)	Sampling noise in DML estimates		\$1,465
(6)	Total RMSE at average observational sample size	$\sqrt{(4)^2 + \frac{\bar{N}_1^{\text{RCT}}}{\bar{N}_1^{\text{Obs}}}(5)^2}$	\$2,073

Sampling noise is the square root of mean sampling variance.

Benchmarking Mean Squared Error of DML Estimator

- Natural benchmark for evaluating RMSE of an observational estimator is RMSE of RCT estimates

- If observational estimator has RMSE comparable to RCTs, then it may provide a good substitute; if it is much higher, value in running a trial

Benchmarking Mean Squared Error of DML Estimator

- **Definition: Experimentally Validated Digital Twins Estimator**
  - An observational estimator $\hat{\tau}$ is an *experimentally validated digital twins estimator* if it has expected root mean squared error loss comparable to that of RCT estimates:

$$\text{RMSE}(\hat{\tau}) \lesssim \sigma_{\text{RCT}}$$

$$\$2{,}073 \lesssim \$2{,}126$$

- Intuition: DML has lower RMSE than RCT estimate because it makes up for greater bias with lower sampling error (due to larger samples)

Can the Digital Twins Estimator Detect Economically Meaningful Effects?

- RMSE also provides a basis for statistical inference
- Can be used to assess whether $\hat{\tau}^{\mathrm{DML}}$ can distinguish between more and less effective programs
  - Signal standard deviation of treatment effects across our 37 RCTs: $\hat{\sigma}_{\tau} = \$5{,}699$
  - Digital twins estimator's margin of error: $1.96 \times \mathrm{RMSE}(\hat{\tau}^{\mathrm{DML}}) = \$4{,}063$
- DML estimator's error is small enough that it can identify typical (one SD) differences in programs' treatment effects

Why Does DML Outperform Prior Observational Estimators?

- Replicate LaLonde (1986) and Dehejia and Wahba (1999) on the same 37 RCTs
  - Linear regression with CPS-style covariates, with a CPS-sized comparison sample

- DML estimator differs from LaLonde's in four respects:

1. Comparison sample: size and overlap
2. Covariates
3. Model flexibility
4. Doubly robust estimation



Noise-Adjusted RMSE Comparison to LaLonde (1986)

By Covariates, Model, and Comparison Sample

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde				
Dehejia–Wahba				
Expanded				$1,717

Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.

Noise-Adjusted RMSE Comparison to LaLonde (1986)

LaLonde-style estimates have substantially higher error

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde	$4,485			
Dehejia–Wahba				
Expanded				$1,717

Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.



Noise-Adjusted RMSE Comparison to LaLonde (1986)

A larger, better-matched comparison sample explains some, but not most, of the gain

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde	$4,485	$3,746		
Dehejia–Wahba				
Expanded				$1,717

Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.

Noise-Adjusted RMSE Comparison to LaLonde (1986)

Richer covariates further reduce the error, even with a linear model

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde	$4,485	$3,746		
Dehejia–Wahba				
Expanded	$3,365	$2,698		$1,717

Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.

Noise-Adjusted RMSE Comparison to LaLonde (1986)

Switching to DML alone does not close the gap

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde	$4,485	$3,746	$3,397	
Dehejia–Wahba				
Expanded	$3,365	$2,698		$1,717

Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.

Noise-Adjusted RMSE Comparison to LaLonde (1986)

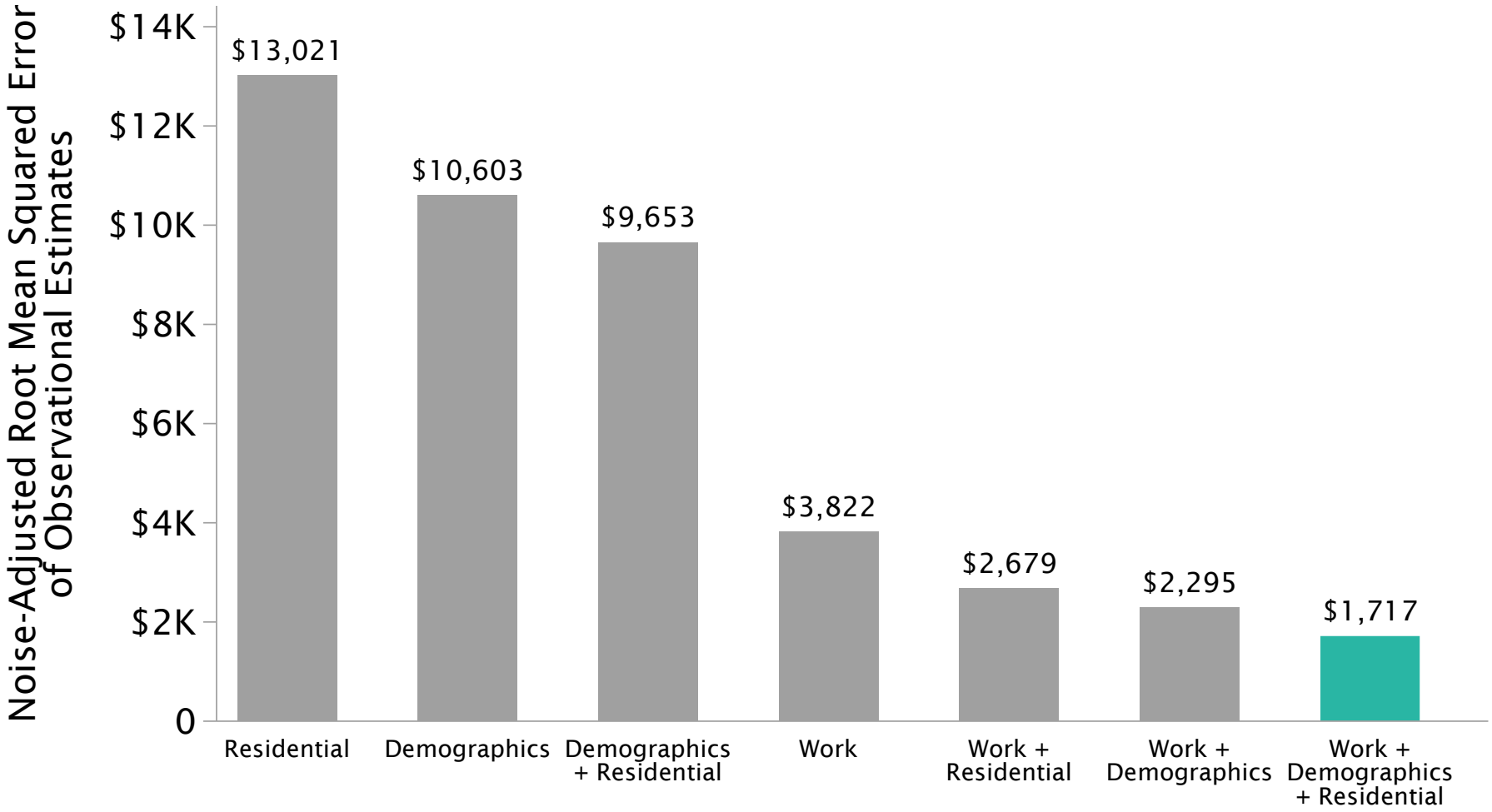
Sample, covariates, and model flexibility are complementary

Model	Parametric (OLS)		Flexible (DML)	
Comparison sample	CPS-sized	Population-scale	CPS-sized	Population-scale
Covariates	LaLonde-style restrictions	Targeted age and geographic restrictions	LaLonde-style restrictions	Targeted age and geographic restrictions
LaLonde	$4,485	$3,746	$3,397	$3,279
Dehejia–Wahba	$3,960	$3,753	$2,971	$2,907
Expanded	$3,365	$2,698	$2,769	$1,717

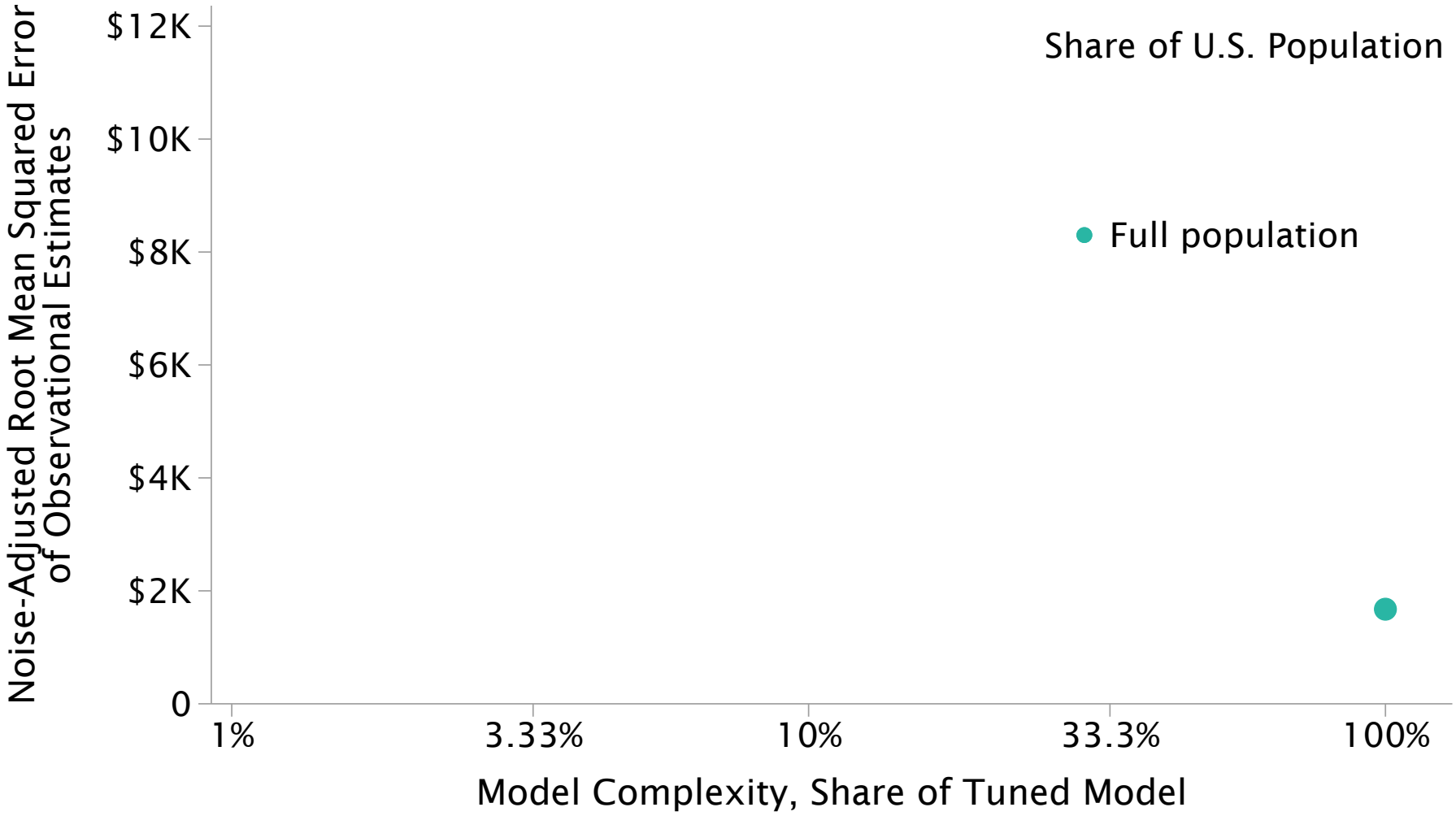
Noise-adjusted RMSE across 37 randomized trials, for annual W-2 earnings 2–5 years after assignment.
CPS-sized estimates average 50 repetitions. Comparison samples differ in size and eligibility restrictions.

Noise-Adjusted RMSE by Set of Covariates

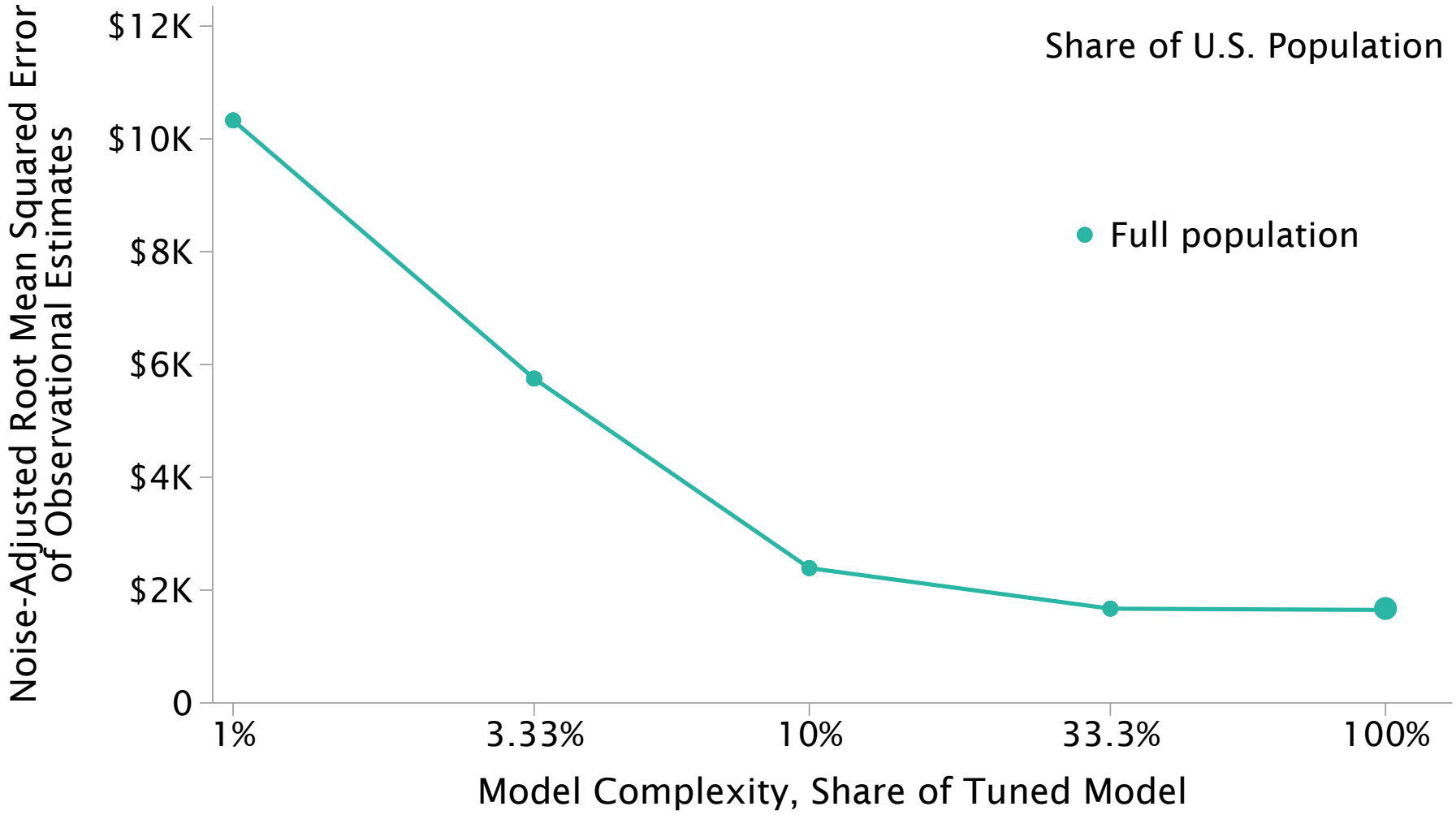
Digital Twins Estimator with Each Combination of Work History, Demographics, and Residential History



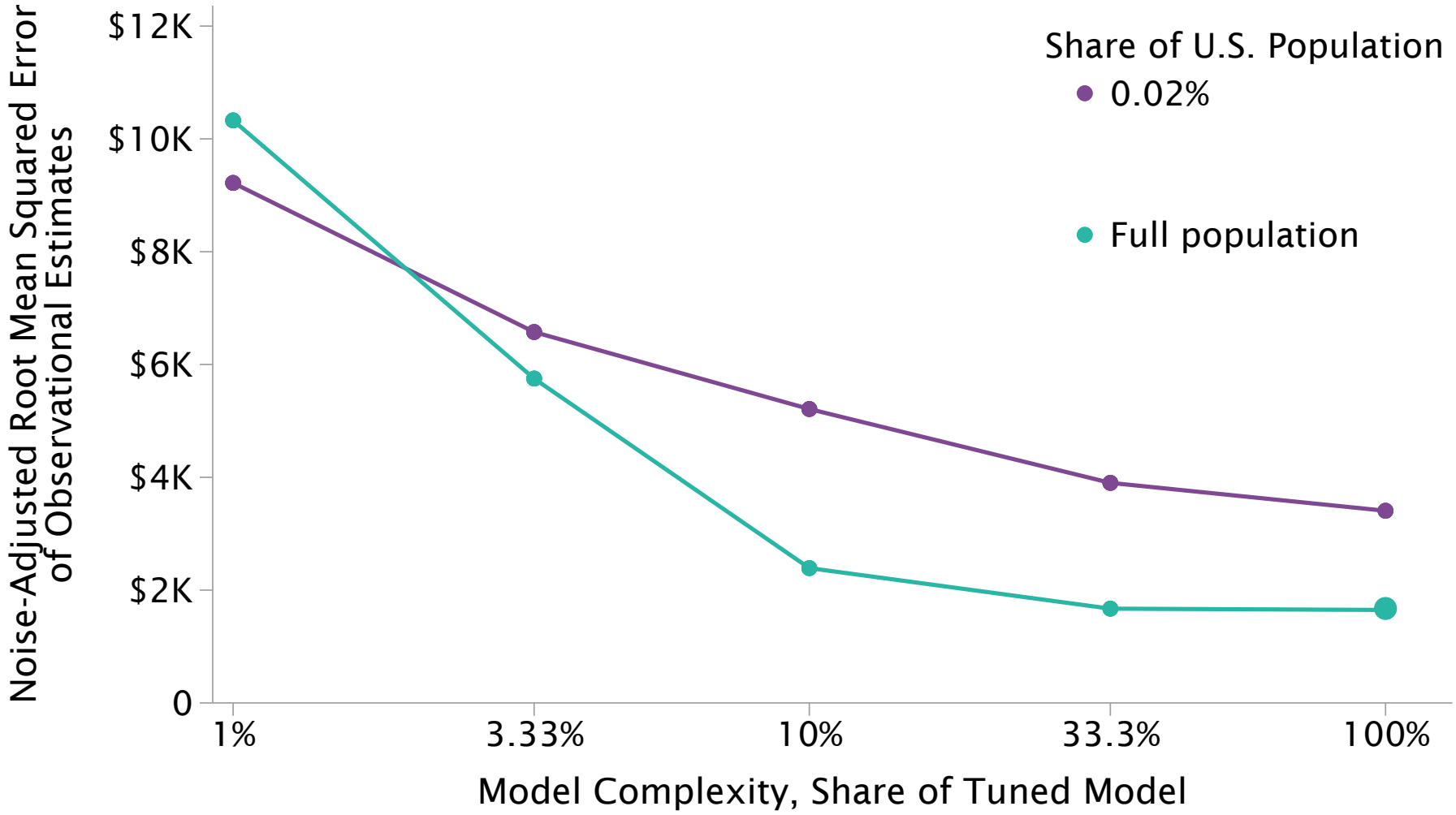
Noise-Adjusted RMSE by Model Complexity and Sample Size



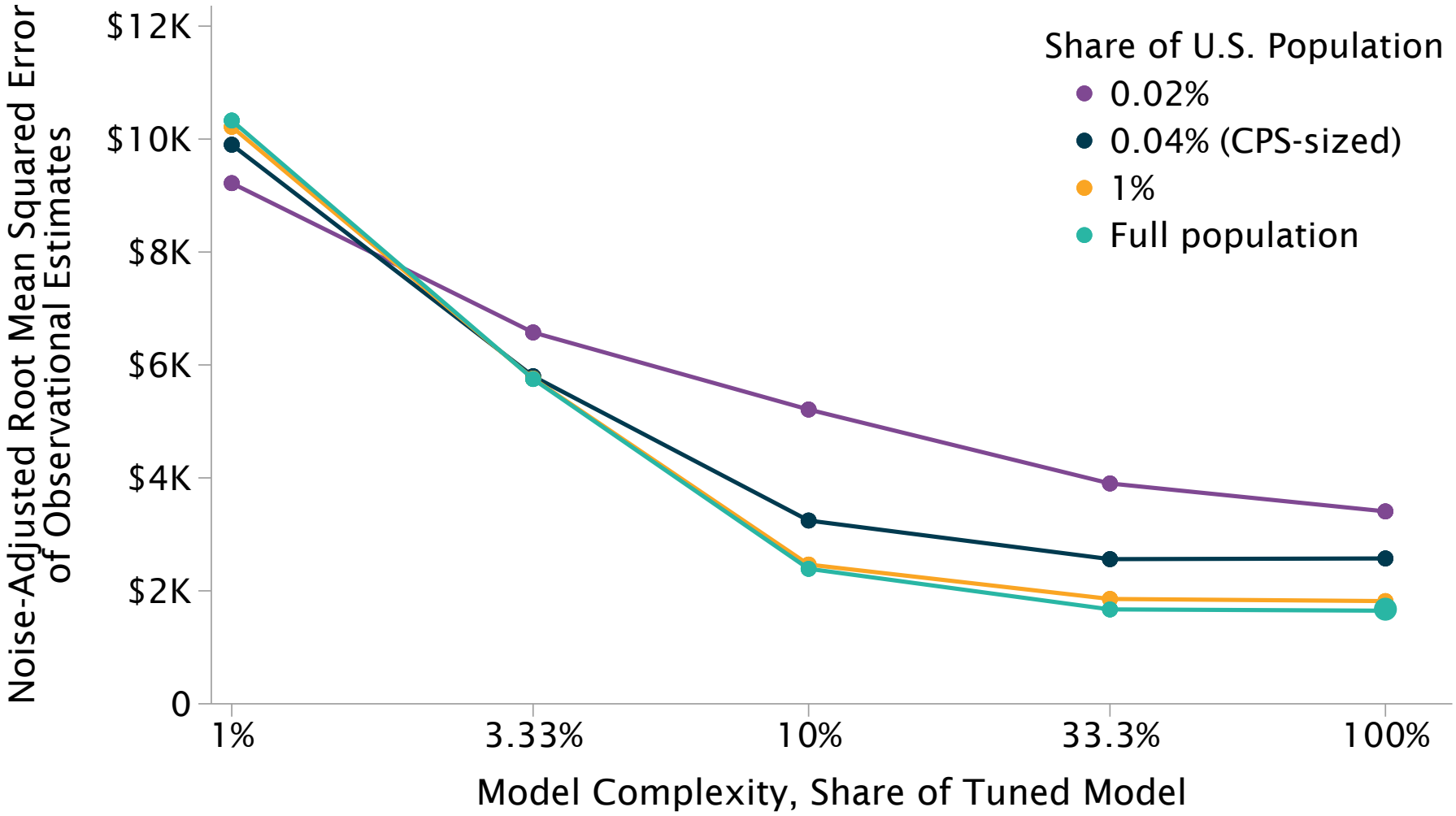
Noise-Adjusted RMSE by Model Complexity and Sample Size



Noise-Adjusted RMSE by Model Complexity and Sample Size



Noise-Adjusted RMSE by Model Complexity and Sample Size



Part 3: Estimating Treatment Effects Beyond the Experimental Sample



Applying the Digital Twins Estimator to Programs Without an RCT

- Goal: estimate treatment effect in a new program × site × enrollment-year cell k outside the experimental validation sample
  - Cannot measure the estimator's bias in this new cell directly

- **Assumption 1 (Common Distribution of Biases):** The target cell bias and validation biases have a common marginal distribution F_B

$$B_j \sim F_B \text{ for every cell } j \text{ in the validation sample and the target cell } k$$

- Credible when selection into participation is similar across cells
  - E.g. career-focused training programs, adults w/ prior work histories

Applying the Digital Twins Estimator to Programs Without an RCT

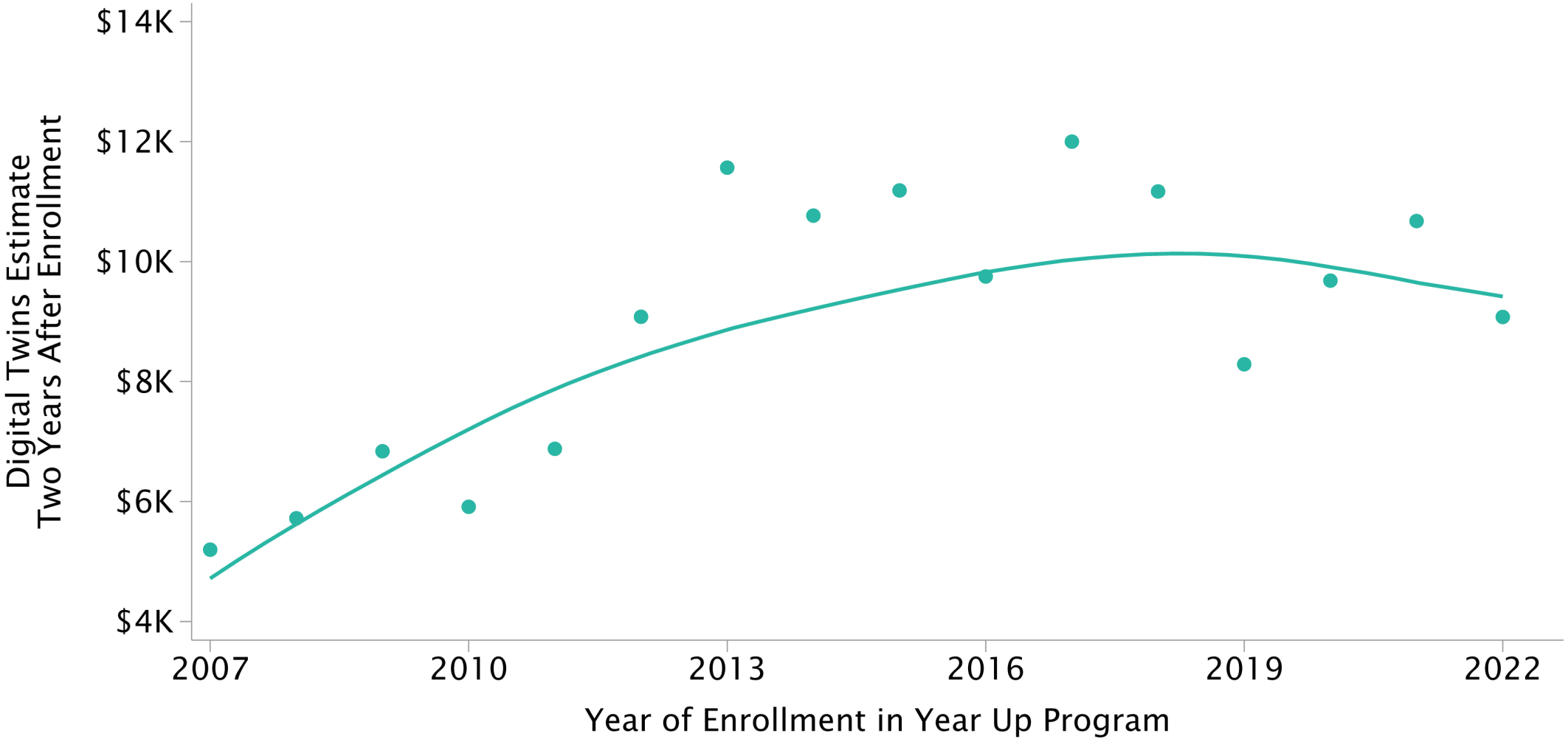
- Under Assumption 1, RMS bias in a new cell *k* is $1,717

- Compute the total RMSE in a new cell by combining this bias with the cell's DML sampling error

$$RMSE\left(\hat{\tau}_k^{DML}\right) = \left(1{,}717^2 + \hat{\sigma}_k^{2,DML}\right)^{1/2}$$

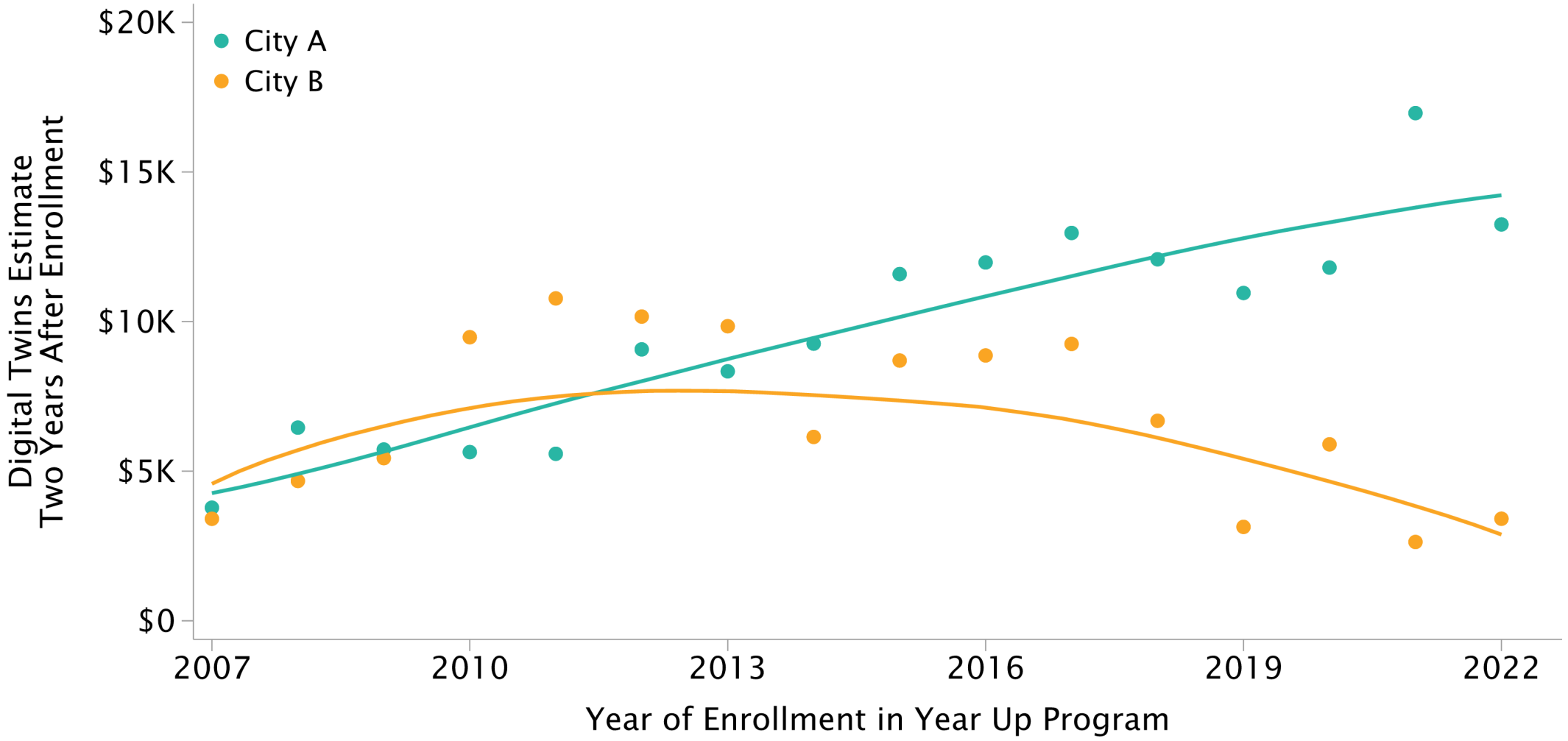
Using Digital Twins for Real-Time Outcome Measurement

Impacts of Year Up Program on Earnings, By Cohort of Program Enrollment



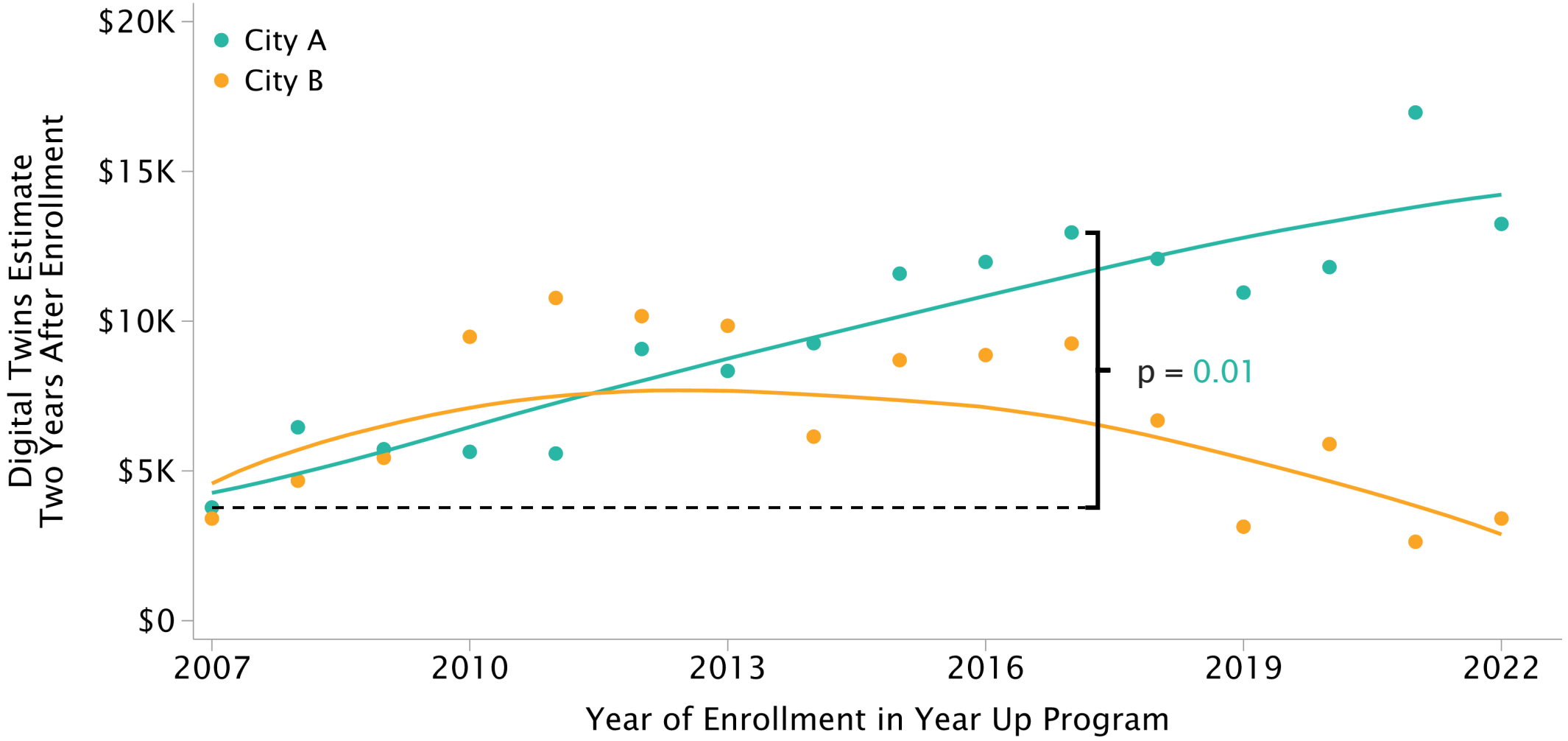
Using Digital Twins for Real-Time Outcome Measurement

Impacts of Year Up Program on Earnings, By Site and Cohort of Program Enrollment



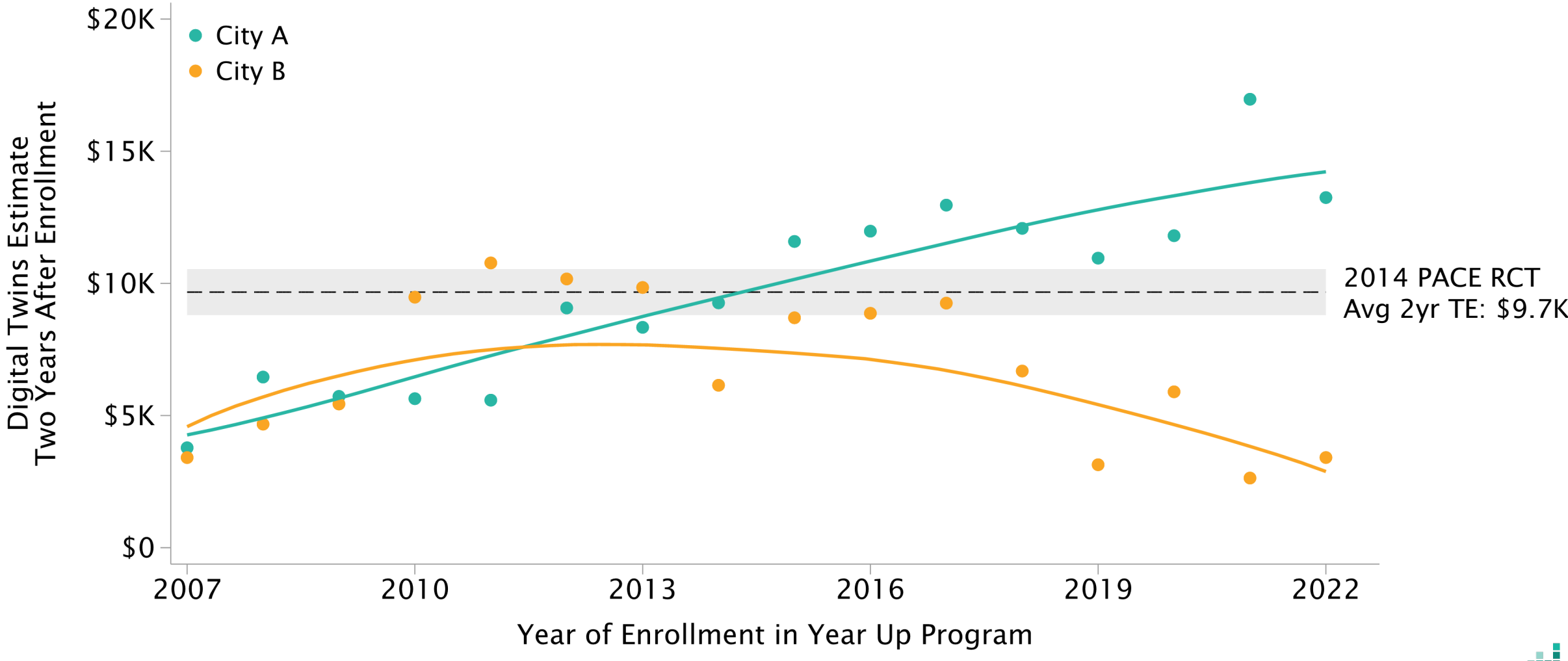
Using Digital Twins for Real-Time Outcome Measurement

Impacts of Year Up Program on Earnings, By Site and Cohort of Program Enrollment



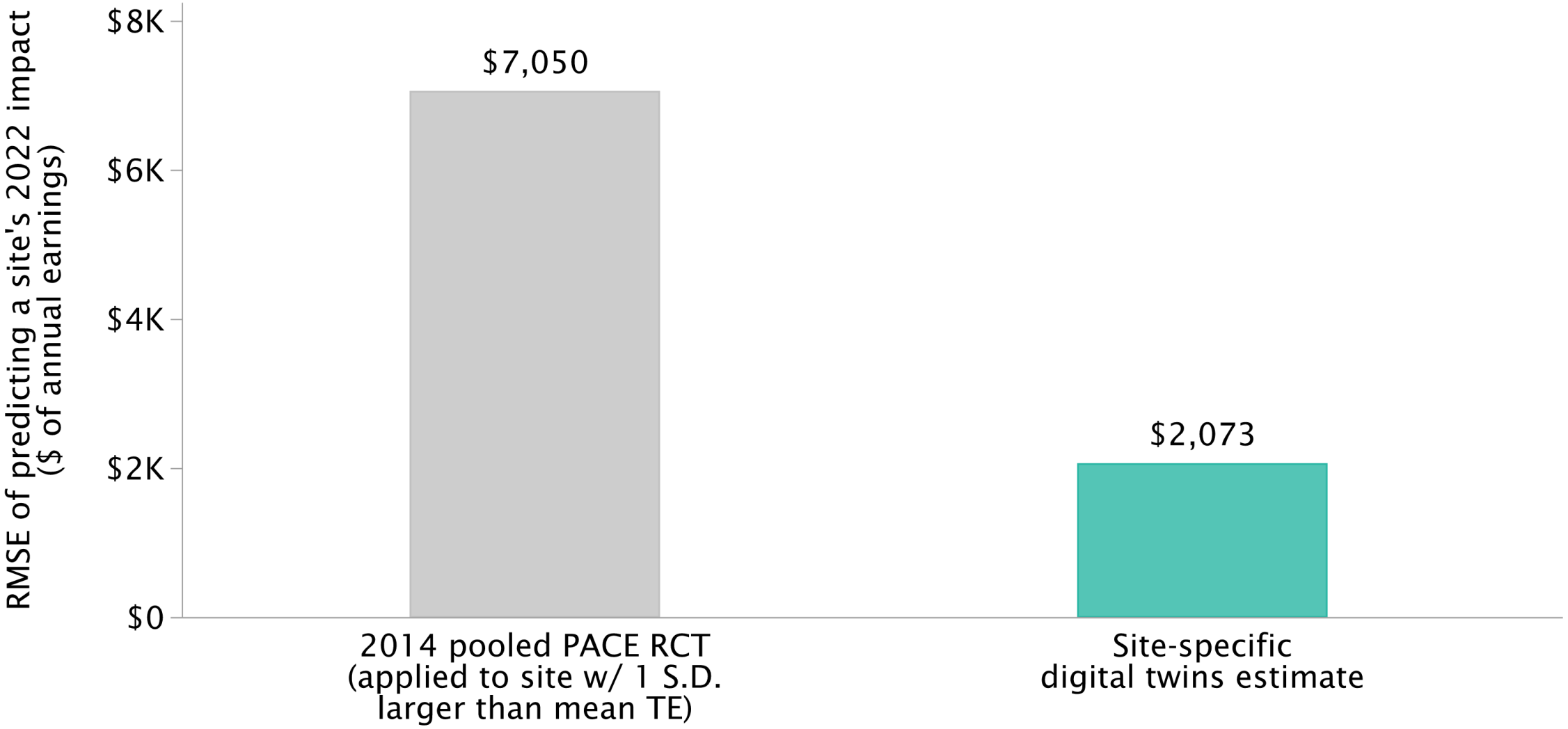
Using Digital Twins for Real-Time Outcome Measurement

Impacts of Year Up Program on Earnings, By Site and Cohort of Program Enrollment



Gains from Cell-Specific DML Estimate vs. Historical Pooled RCT

Predicting Year Up's Impact at a Site with TE 1 S.D. Above Mean

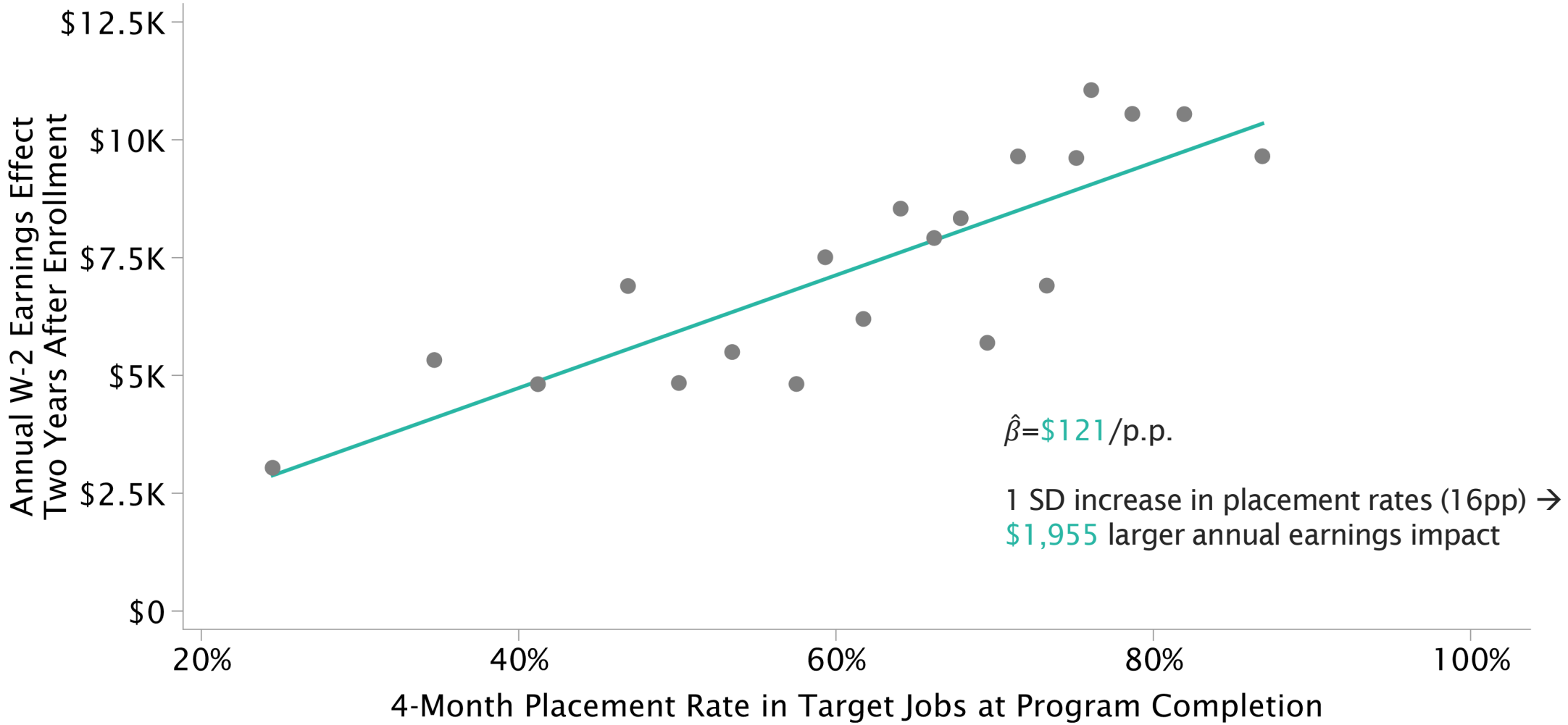


Part 4: Mechanisms Underlying Effective Training Programs



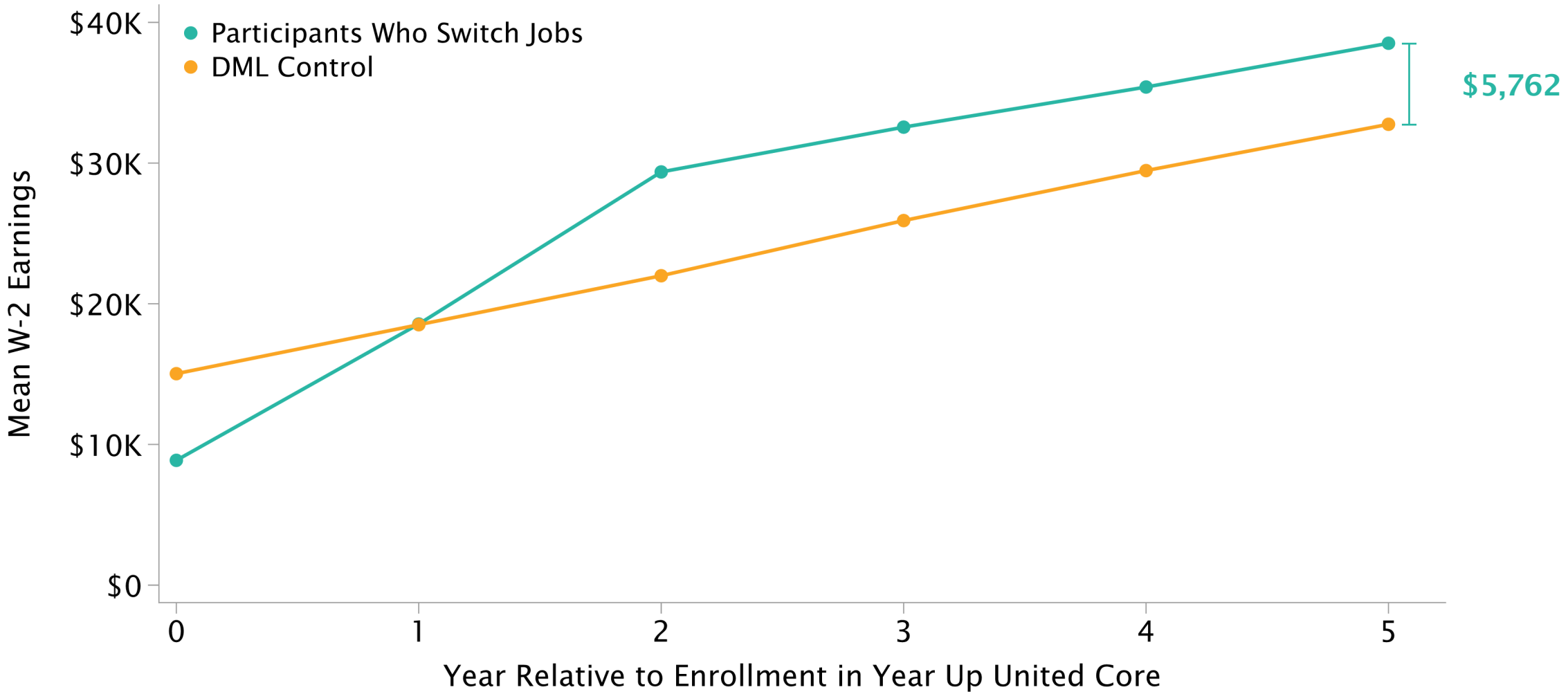
Impacts of Year Up on Earnings vs. Placement Rates in Target Jobs

By Site and Year of Enrollment: 271 Site-by-Enrollment-Year Cells, 2007–2023; Earnings Two Years After Enrollment



Earnings of Job Switchers vs. Their Digital Twins

Year Up Participants Who Changed Employers Within Five Years of Enrollment (90% of Participants)



Conclusions

- Methodology: Rich administrative data + flexible machine learning models can yield observational estimates with error comparable to RCTs

- Workforce training: program impacts vary sharply across sites and over time

  - The most effective programs pair human capital (skills) with social capital (internships and job placements)

- Broader lesson: in the age of big data, RCTs may be most valuable to validate observational models rather than identify treatment effects themselves

Next Steps and Future Work

1. Apply the method to evaluate a broader range of training and career advancement programs
    - Ex: Western Governors University, Dallas College, Texas State Technical College, Massachusetts Dept of Higher Education programs

2. Study if using richer models (e.g., LLMs) to construct digital twins can further reduce RMSE
    - Use richer information about employers and neighborhoods to distinguish otherwise similar workers

3. Develop experimentally validated digital twins models in other domains (e.g., health, education, criminal justice)

Appendix

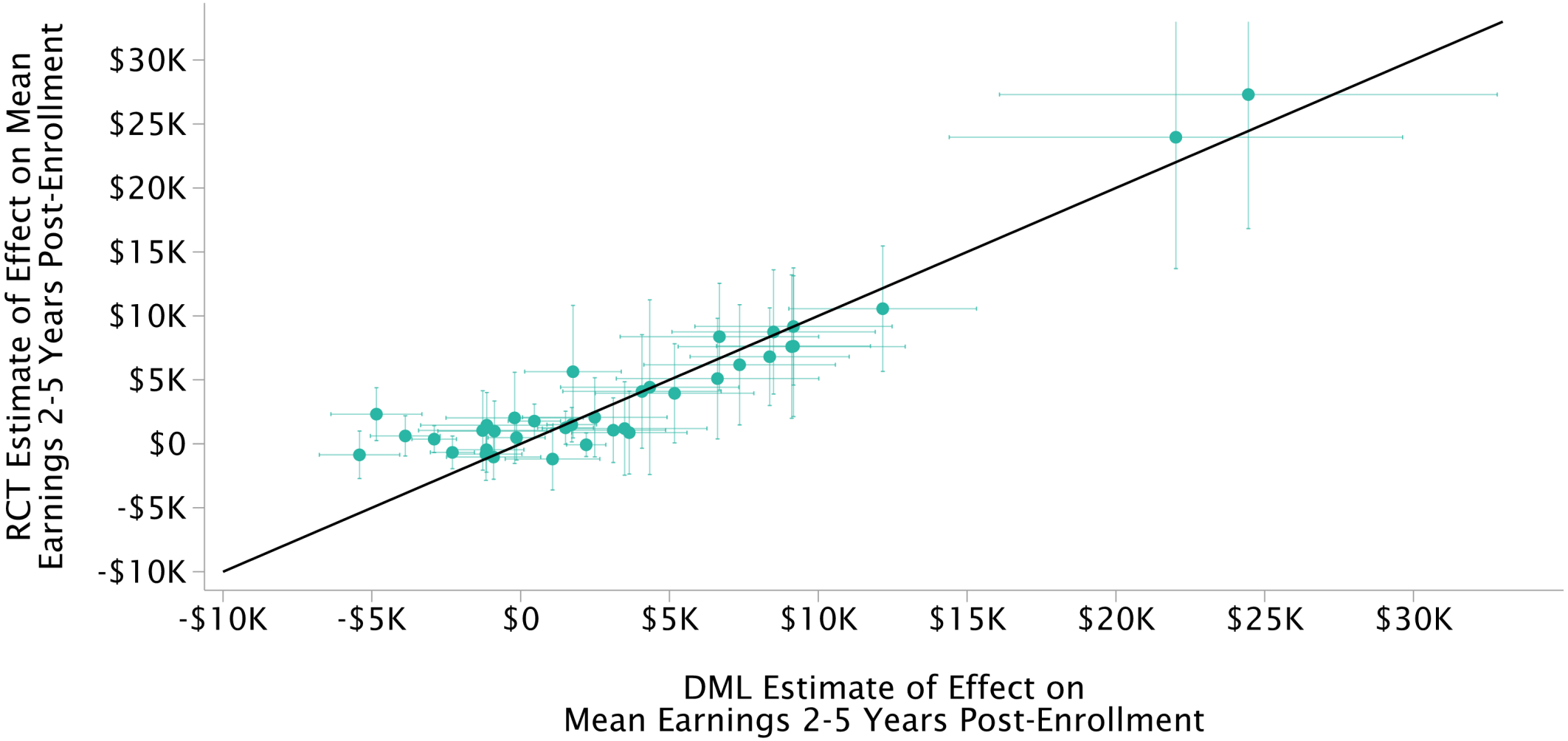
List of Training Programs: 37 RCTs Across 24 Programs

- WorkAdvance trial with four sectoral job training programs in 2011-13
  - Per Scholas: Information Technology jobs in Bronx, NY, 15 weeks
  - Madison Strategies: Transportation, manufacturing jobs in Tulsa, OK, 2-32 weeks
  - Towards Employment: Health care, manufacturing jobs in Cleveland, OH, 2-32 weeks
  - St. Nick's Alliance: Environmental Remediation jobs in Brooklyn, NY, 5-12 weeks
- Year Up PACE Trial in nine cities in 2013-14
- Project QUEST: Health care training in San Antonio
- SEIS Trial (Sectoral Employment Impact Study): Per Scholas, JVS Boston, and WRTP
- ERA: Employment Retention and Advancement demonstration
- STED: Subsidized and Transitional Employment Demonstration



Effects of Workforce Training Programs: 95% Confidence Intervals

Observational vs. Randomized Trial Estimates for Programs Implemented in 37 Sites



Standard Errors of RMSE Estimates

Table 1 Estimates and Program-Level Bootstrap Standard Errors

Measure	Estimate	Bootstrap s.e.
Raw RMS difference, DML vs. RCT	$2,333	(403)
Noise-adjusted RMS bias	$1,717	(730)
Sampling noise of DML estimates at RCT sample sizes	$1,465	(328)
Total RMSE of DML estimates at RCT sample sizes	$2,257	(405)
Total RMSE at average observational sample sizes	$2,073	(462)
RMSE of RCT estimates	$2,126	(396)
Noise-adjusted SD of true treatment effects	$5,699	(2,230)

Standard errors from 10,000 program-level cluster bootstrap replications over the 24 programs (Table 1). The total-RMSE comparison is 'comparable', not 'equal'.

