Using Prior Scores to Evaluate Bias in Value-Added Models

Raj Chetty, Stanford University and NBER
John N. Friedman, Brown University and NBER
Jonah Rockoff, Columbia University and NBER

January 2016

Introduction: Bias in Value-Added Models

- Outcome-based value added (VA) models increasingly used to measure the productivity of many agents
    - Teachers, schools, neighborhoods, doctors, CEOs…

- Central question in determining whether VA measures are useful for policy: to what extent are VA estimates biased by selection?
  [e.g., Rothstein 2009; Kane and Staiger 2008; Chetty, Friedman, Rockoff 2014]
    - Ex: do differences in latent abilities of students assigned to teachers bias estimates of teacher VA?

Evaluating Bias Using Lagged Outcomes

- One intuitive approach to assessing degree of bias in VA models: test for balance in lagged values of the outcome
    - Simple to implement: regress prior scores on current teacher VA [Chetty, Friedman, Rockoff 2014; Rothstein 2015]
    - Intuition: current teachers cannot have causal effects on prior scores
    - Analogous to standard analysis of pre-trends in outcomes used to evaluate bias in program evaluation literature

Overview

- We show that balance tests using lagged values of the outcome are sensitive to model specification in VA models
  - Prior scores will typically be correlated with VA estimates even when VA estimates are unbiased
  - More generally, tests using prior scores are uninformative about degree of forecast bias when VA model is misspecified
  - Intuition: Correlated shocks enter *both* current VA estimate and lagged outcome in ways that are sensitive to model specification

Overview

- Why are lagged outcome tests of balance more robust in conventional treatment effect settings (e.g., class size)?

- Two key differences in VA models:

  1. Treatment itself is estimated, rather than exogenously observed

  2. Estimation error does not vanish in large datasets because sample size per teacher remains small asymptotically

- With exogenous treatments, noise in lagged outcomes uncorrelated with treatment and estimation error vanishes asymptotically

- Experimental/quasi-experimental methods provide a less model-dependent way to assess bias in VA models

Outline

1. Specification of Value-Added Model

2. Monte Carlo Simulation Results

3. Other Approaches to Evaluating Bias

Model Setup: Students and Teachers

- We consider estimation of teacher effects for concreteness, but results translate directly to other applications

- Data on students' test scores and classroom assignments in years t = 1, 2 used to predict teacher quality in years t > 2

- Student i is assigned in year t to classroom $c(i,t)$ and teacher $j(c(i,t)) = j(i,t)$
  - Each teacher *j* teaches *C* classrooms per year in a single grade
  - Each classroom *c* has *I* students

Model Setup: Tracks and Correlated Shocks

- Key new element used to assess sensitivity to model specification: classrooms grouped into tracks (s)
  - Ex: regular vs. honors classes
  - Students and teachers assigned to a given track $s(i)$ in all years
  - Classroom shocks within tracks are correlated, both within and across grades
    - For instance, curriculum in a given track may line up particularly well with tests in certain years

Data Generating Process for Test Scores

- Student’s test score in year t is given by

$$A_{it} = \underset{\substack{\nearrow \\ \text{fixed} \\ \text{ability}}}{\delta_i} + \underset{\substack{\nearrow \\ \text{ability} \\ \text{trend}}}{\alpha_i t} + \underset{\substack{\nearrow \\ \text{teacher} \\ \text{effect}}}{\mu_{j(i,t)}} + \underset{\substack{\nearrow \\ \text{class} \\ \text{shock}}}{\theta_{c(i,t),t}} + \underset{\substack{\nearrow \\ \text{track} \\ \text{shock}}}{\psi_{s(i,t),t}} + \underset{\substack{\nearrow \\ \text{student} \\ \text{shock}}}{\varepsilon_{it}}$$

- Assume that teacher value-added (μ_j) does not vary over time

- Student assignment to teachers may be correlated with ability
  [Rothstein 2010]
  - Static tracking: μ_j correlated with δ_i (fixed ability)
  - Dynamic tracking: μ_j correlated with α_i (ability trends)

Estimator for Value-Added

- Teacher VA estimated using a standard gains specification
  - Average change in students' end-of-year test scores, adjusting for noise using a standard shrinkage factor
- Let $\Delta A_{it} = A_{it} - A_{i,t-1}$ denote student i's test score gain in year t
- Estimator for VA of teacher *j* using test score data from years 1 and 2:

$$\hat{\mu}_j = \lambda \Delta \overline{A}_{j,t=2}$$

$$\text{where } \lambda = \frac{\sigma_{\mu}^{2}}{\sigma_{\mu}^{2}+\sigma_{\Delta\psi}^{2}+\sigma_{\theta}^{2}/C+\sigma_{\Delta\varepsilon}^{2}/CI}$$

- This estimator minimizes MSE of out-of-sample forecasts of test scores and is posterior expectation of VA with Normal distributions

Forecast Bias: Definition

- Consider running an experiment where students are randomly assigned to teachers and estimating the regression:

$$\Delta A_{it} = a + b\hat{\mu}_{j(i,t)} + \zeta_{it}$$

- Prediction coefficient in this regression identifies degree of forecast bias $(1 - \mathrm{b})$ [Kane and Staiger 2008; Chetty, Friedman, Rockoff 2014]

  - If VA estimates are forecast unbiased $(b = 1)$, assigning a student to a teacher with one unit higher estimated VA will increase his score by one unit

Estimating Forecast Bias

- Gains model yields forecast unbiased estimates when there is static tracking (sorting on δ_i) but not with dynamic tracking (sorting on α_i)

- How can we distinguish these two cases and, more generally, estimate degree of forecast bias?

  - Is correlation of VA estimates with prior scores informative?

- Use a set of Monte Carlo simulations to answer this question

Baseline Parameters for Monte Carlo Simulations

Governing Student, Classroom, Year, and Track Effects

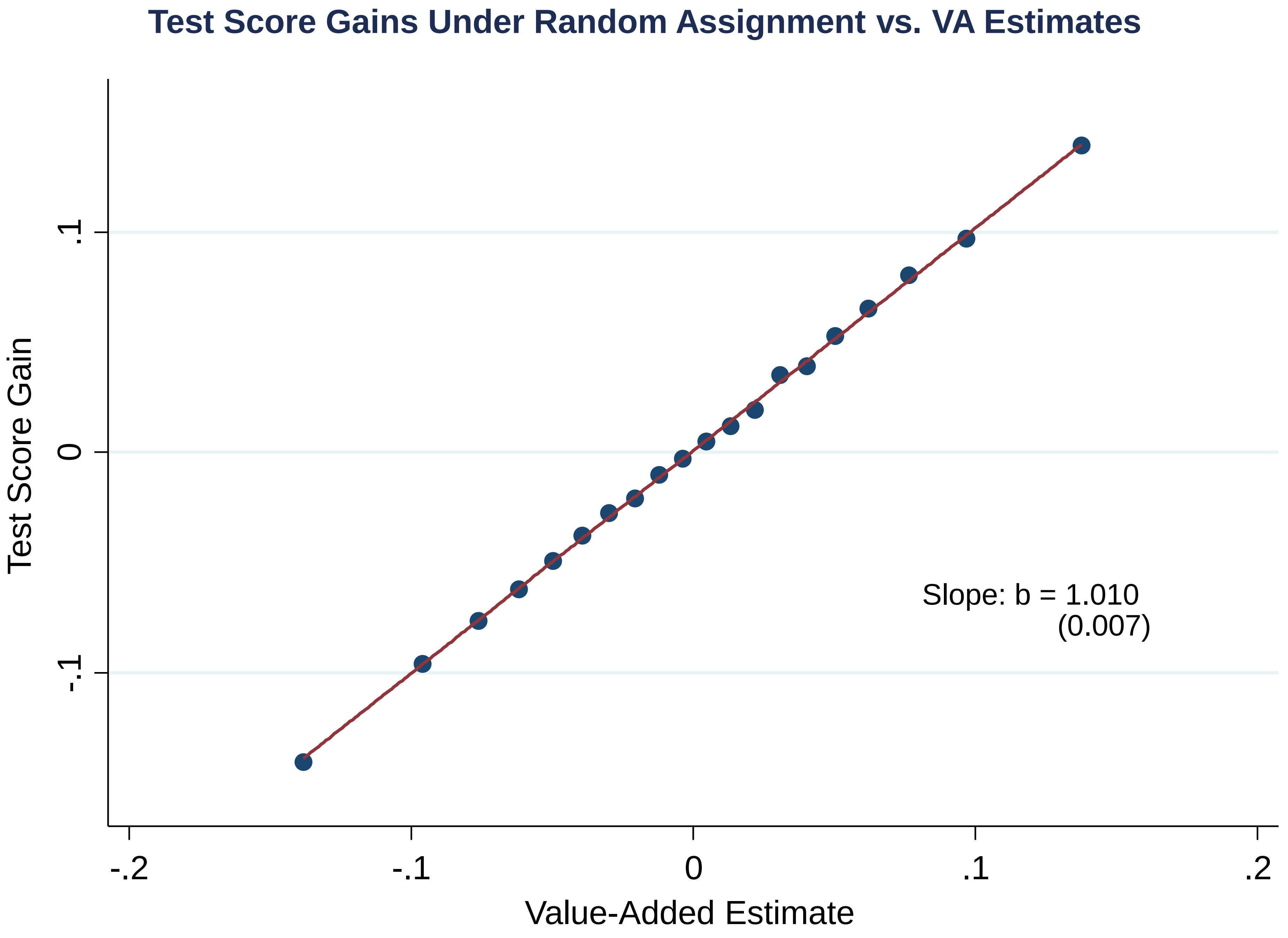
Parameter	Value
Number of Schools	2000
Number of Tracks per School	5
Number of Teachers per Track	4
Number of Classrooms per Teacher (C)	4
Number of Students per Classroom (I)	25
SD of Student Ability (σ_δ)	0.88
SD Of Trend Differences Across Students (σ_α)	0.15
SD Of Teacher Value-Added (σ_μ)	0.10
SD of Classroom Shocks (σ_θ)	0.08
SD of Track-Year Shock (σ_ψ)	0.06
Degree of Sorting (Level)	0.25
Degree of Sorting (Trend)	0.00

Simulation Results

- Begin by considering case with only static tracking, so there is no bias in VA estimates

- First examine relationship between test score gains under random assignment and VA estimates based on observational data

    - As expected, prediction coefficient is 1 in this experiment (no forecast bias)

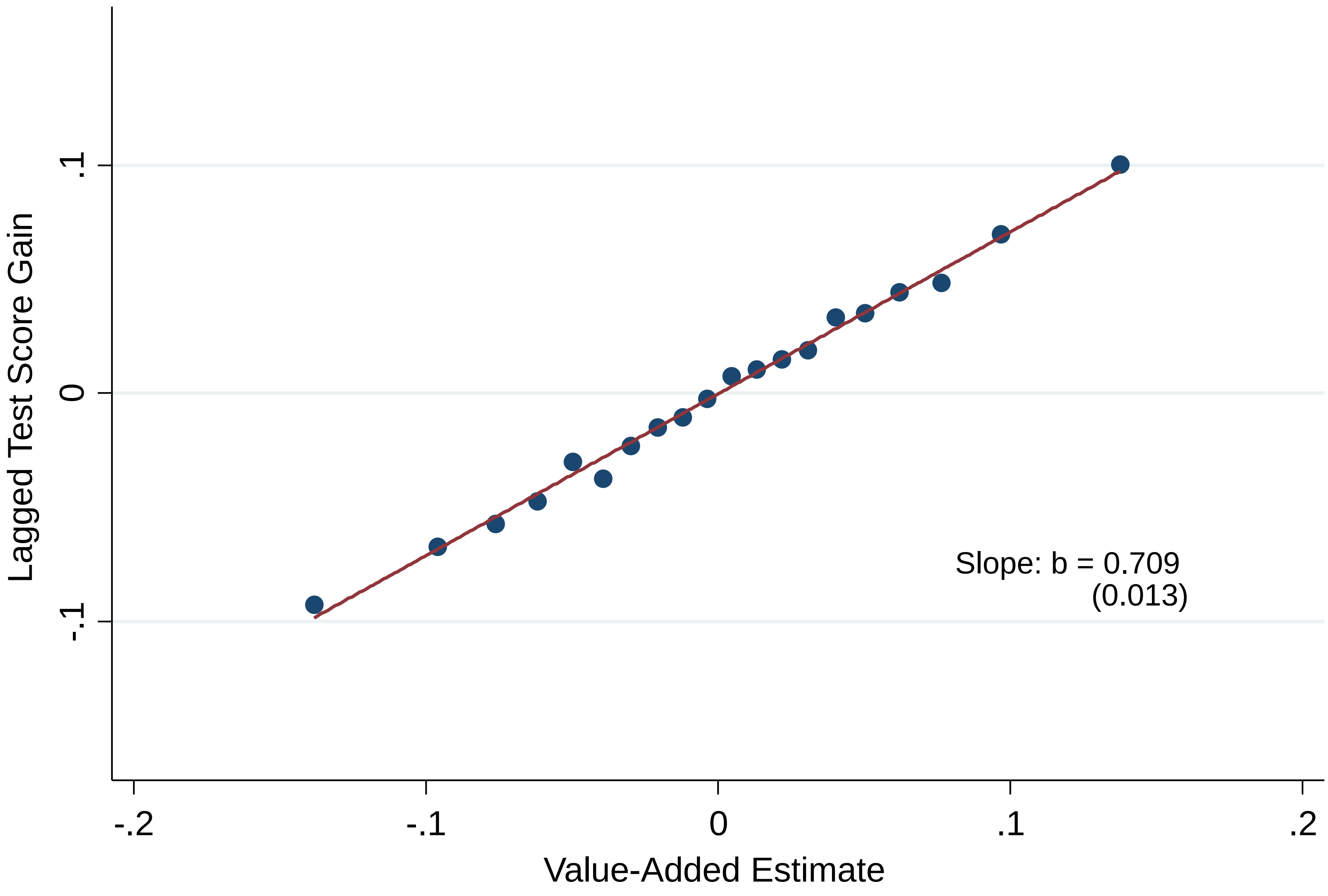

Test Score Gains Under Random Assignment vs. VA Estimates
Test Score Gain
.1
0
-.1
-.2
-.1
0
.1
.2
Value-Added Estimate
Slope: b = 1.010
(0.007)

Correlation with Prior Scores

- Now regress lagged gains $\Delta A_{i,t-1}$ on current teacher's VA estimate

$$\Delta A_{i,t-1} = a + b\hat{\mu}_{j(i,t)} + v_{it}$$

Lagged Test Score Gains vs. Current VA Estimates



Correlation with Prior Scores

- Why does current teacher's VA predict lagged test score gain even though there is no bias in this model?

- Track-specific shock ψ_{st} enters both VA estimate and lagged gains because ψ_{st} affects students in all grades in a given track

- Ex.: Suppose VA estimated for 6th grade from 1995 gains
  - Positive track shock in 1995 artificially increases gains, VA estimates
  - Lagged gains for 6th graders in 1996 also affected by the same track shock
  - Therefore VA estimates and lagged gains are correlated

Correlation with Prior Scores

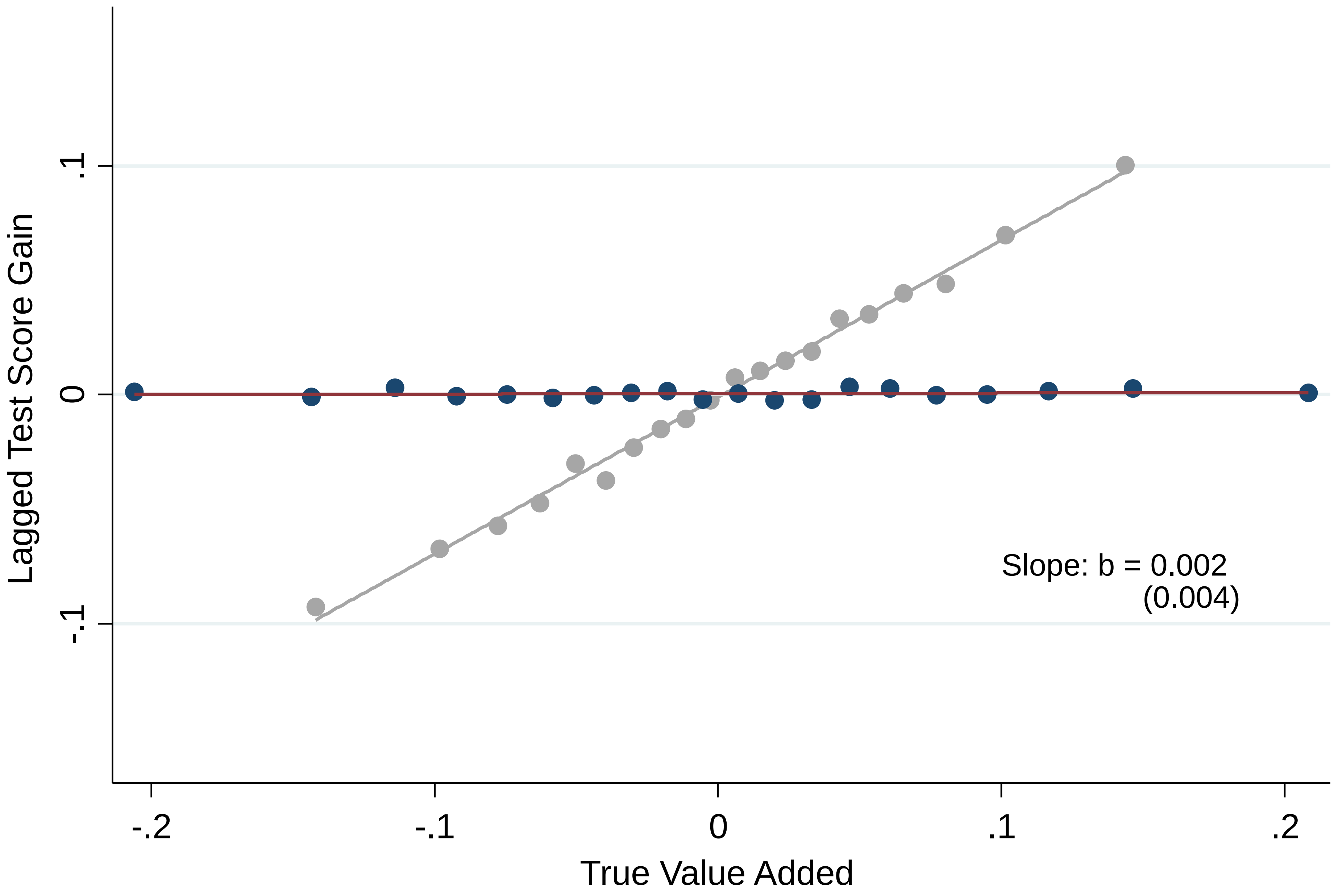
- More generally, relationship between current VA and lagged gains is governed by variance of track-specific shocks:

$$Cov\left(\hat{\mu}_{j(i,t)}, \Delta A_{i,t-1}\right) = \lambda \sigma^2_{\Delta\psi} > 0$$

  - In a model with no track shocks, lagged outcome balance test correctly diagnoses bias

- Root of problem: estimation error in VA
  - If one observed true VA directly (or is studying an exogenous treatment like class size), no correlation with lagged gains

Lagged Test Score Gains vs. True Teacher VA



Slope: b = 0.002
(0.004)

Variants of Lagged Outcome Balance Test

- Common variants of lagged outcome balance test suffer from the same problem
  - For instance, testing whether controlling for lagged gain affects forecast coefficient on VA estimate

Analysis of Variance: Teacher-Level Bias

- We have focused thus far on forecast bias (average prediction)
  [Kane and Staiger 2008, Chetty, Friedman, Rockoff 2014]

- Alternative, more stringent measure: teacher-level bias
  [Rothstein 2010]
  - Is there excess variance across teachers in lagged gains?
  - Typically implemented using an F test in a regression of lagged gains on teacher fixed effects

Effects of Teacher VA on Current and Lagged Test Score Gains

Results from Monte Carlo Simulations

	Randomized experiment	Lagged scores	Lagged scores versus true VA	Observational out-of-sample forecast	Observational, controlling for lagged gain
	(1)	(2)	(3)	(4)	(5)
Dependent Variable:	Current gain	Lagged gain	Lagged gain	Current gain	Current gain
VA estimate	1.010	0.709		0.991	0.833
	(0.007)	(0.013)		(0.017)	(0.016)
True VA			0.002		
			(0.004)		
Control for lagged gain					X
Naïve F-test for teacher effects		F = 2.238 p<0.001			

Notes: *Standard errors clustered by track.*

Analysis of Variance: Teacher-Level Bias

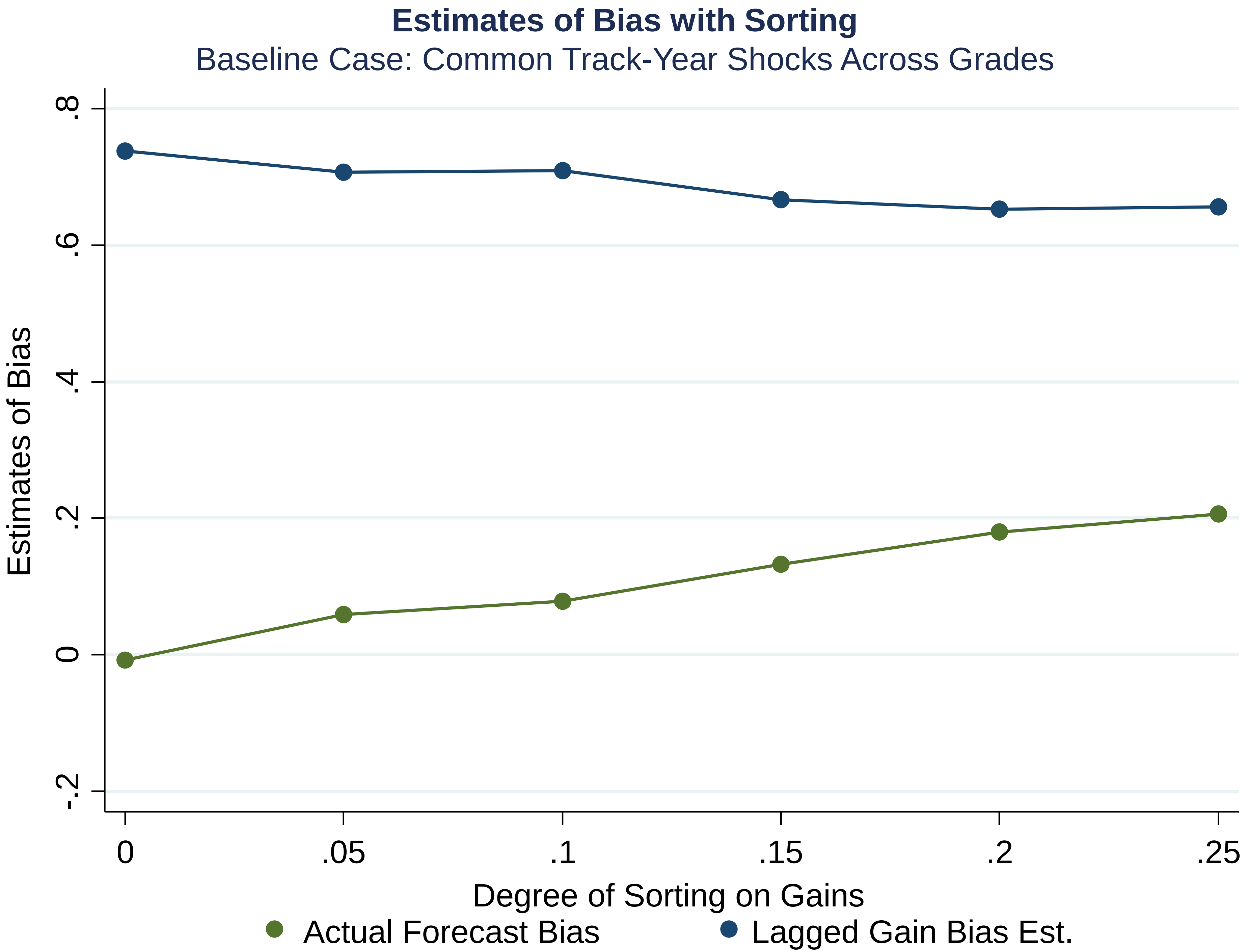
- F test rejects despite fact that lagged gains are not grouped by teacher because variance structure is incorrectly specified
    - Does not account for correlated errors within tracks
- Accounting for this error structure would fix the problem, but again illustrates sensitivity of test to model specification
- Specification matters more in VA models because estimation error does not vanish in large samples
    - In conventional treatment effect settings, misspecification of error structure does not matter for inference in large datasets
    - Sample size per treatment group grows asymptotically
    - In VA models, misspecification matters even in large samples because sample size per teacher does not grow asymptotically

Sorting on Gains

- Now turn to case where VA estimates are in fact biased due to sorting on gains

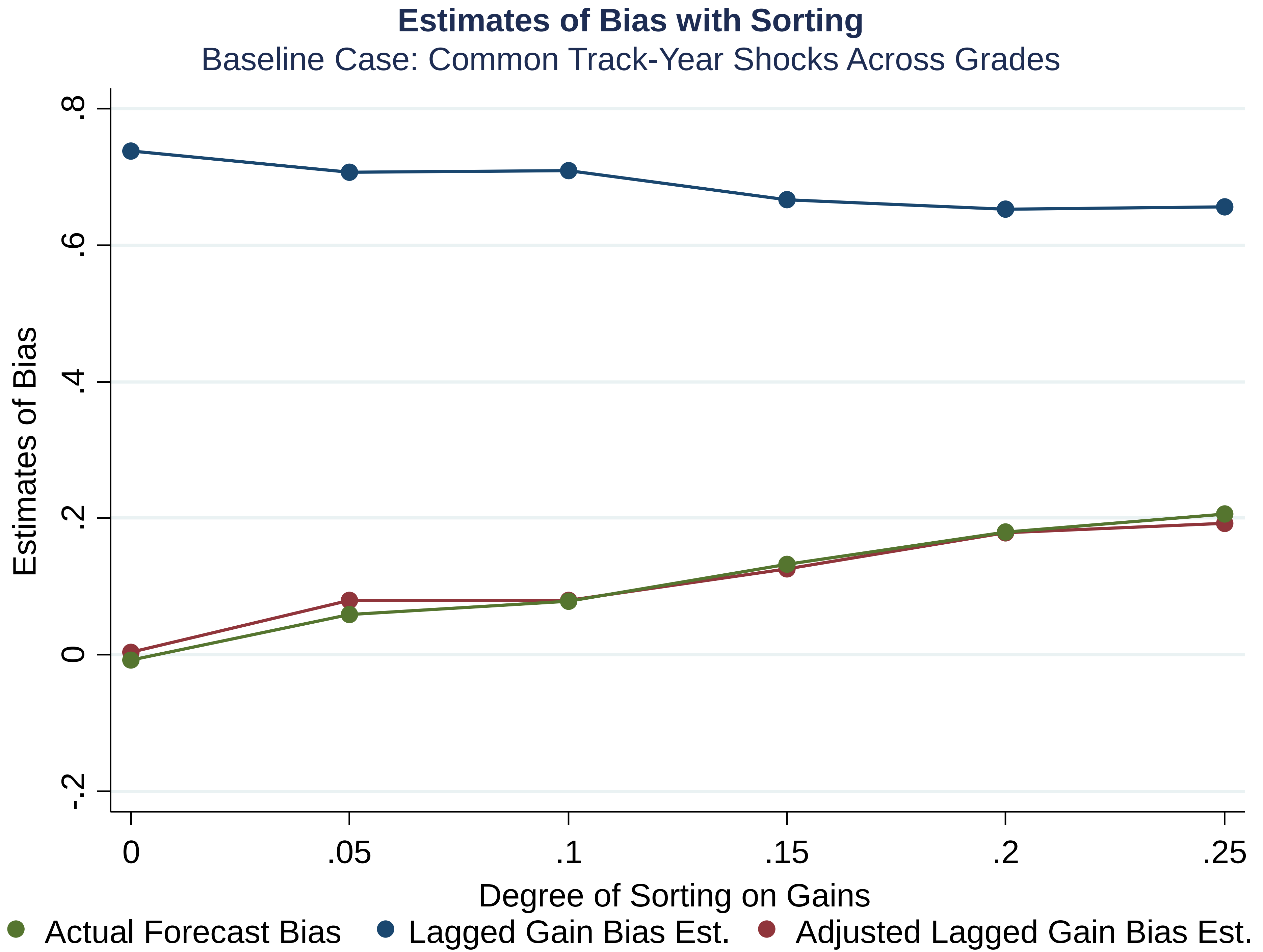
- In model without track shocks, straightforward to show that coefficient from regression of lagged gains on VA exactly matches forecast bias

- No longer true once we allow for correlated shocks within tracks


Estimates of Bias with Sorting
Baseline Case: Common Track-Year Shocks Across Grades
Estimates of Bias
-.2
0
.2
.4
.6
.8
0
.05
.1
.15
.2
.25
Degree of Sorting on Gains
● Actual Forecast Bias
● Lagged Gain Bias Est.

Model-Based Correction to Lagged Outcome Test

- Results above consider naïve implementation of lagged outcome test that does not respect error structure used to estimate VA model
  - Unfair comparison: information used to estimate VA model not used when implementing lagged score test
- Potential solution: adjust lagged outcome test to account for mechanical correlation due to common track shocks
  - Subtract out variance due to common track shocks to form an adjusted estimate
  - Resolves problem when VA model is correctly specified


Estimates of Bias with Sorting
Baseline Case: Common Track-Year Shocks Across Grades
Estimates of Bias
-.2
0
.2
.4
.6
.8
0
.05
.1
.15
.2
.25
Degree of Sorting on Gains
● Actual Forecast Bias
● Lagged Gain Bias Est.
● Adjusted Lagged Gain Bias Est.

Bias Tests Under Misspecification

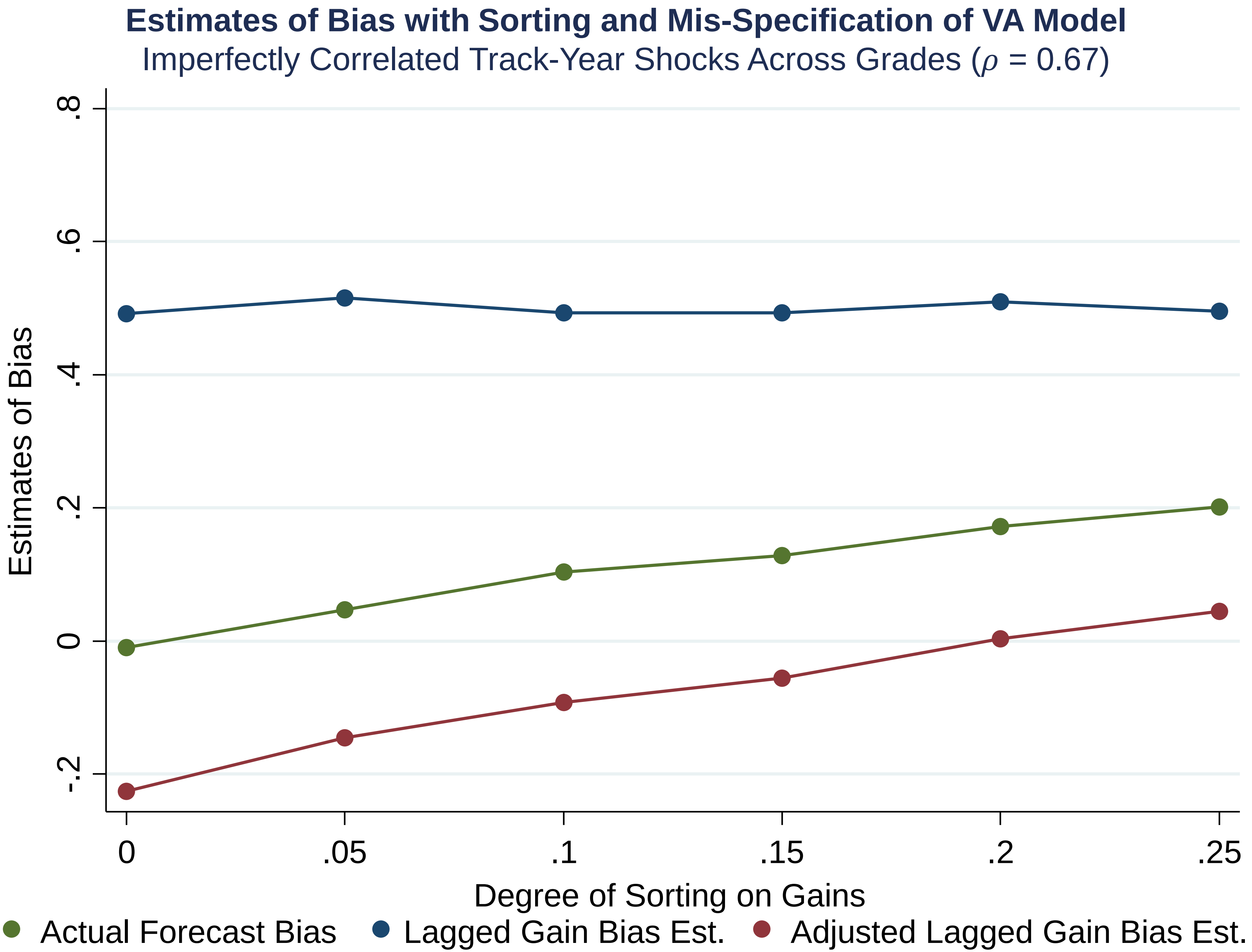
- Deeper problem: such parametric corrections rely heavily on model specification

- More plausible case: model used to estimate VA itself mis-specified

- For example, suppose track-year shocks are in fact not perfectly correlated across grades

    - But econometrician continues to assume they are *both* when estimating VA and when implementing lagged outcome test

    - Now parametric correction to lagged outcome test under assumed model no longer works


Estimates of Bias with Sorting and Mis-Specification of VA Model
Imperfectly Correlated Track-Year Shocks Across Grades (ρ = 0.67)
Estimates of Bias
.8
.6
.4
.2
0
-.2
0
.05
.1
.15
.2
.25
Degree of Sorting on Gains
● Actual Forecast Bias
● Lagged Gain Bias Est.
● Adjusted Lagged Gain Bias Est.

Bias Tests Under Misspecification

- Another potential correction: use leave three years out when estimating VA
  - With iid track shocks, eliminates link between lagged gains and current VA estimates
  - But this method fails if track shocks are serially correlated

Sensitivity of Lagged Outcome Balance Tests

- General lesson: results of lagged outcome tests in VA models are sensitive to model specification
  - Given a VA model, one can always devise a test using lagged outcomes that will yield consistent estimates of bias
  - But proper specification of test depends heavily on model
- Of course, misspecification will create biased VA estimates too
  - Key point: lagged outcome test does not provide a robust guide to the degree of bias is such situations

Other Approaches to Evaluating Bias

- Given sensitivity of lagged outcome tests to model specification, what alternative methods can be used to assess bias in VA models?

- Conceptually, need methods that use data guaranteed to be unrelated to estimation error in VA

- Two existing approaches

Other Approaches to Evaluating Bias

1. Use pre-determined, exogenous covariates (e.g., race or parental income) to evaluate balance

    - Advantage: Noise in outcomes does not directly enter such variables, making such tests less fragile

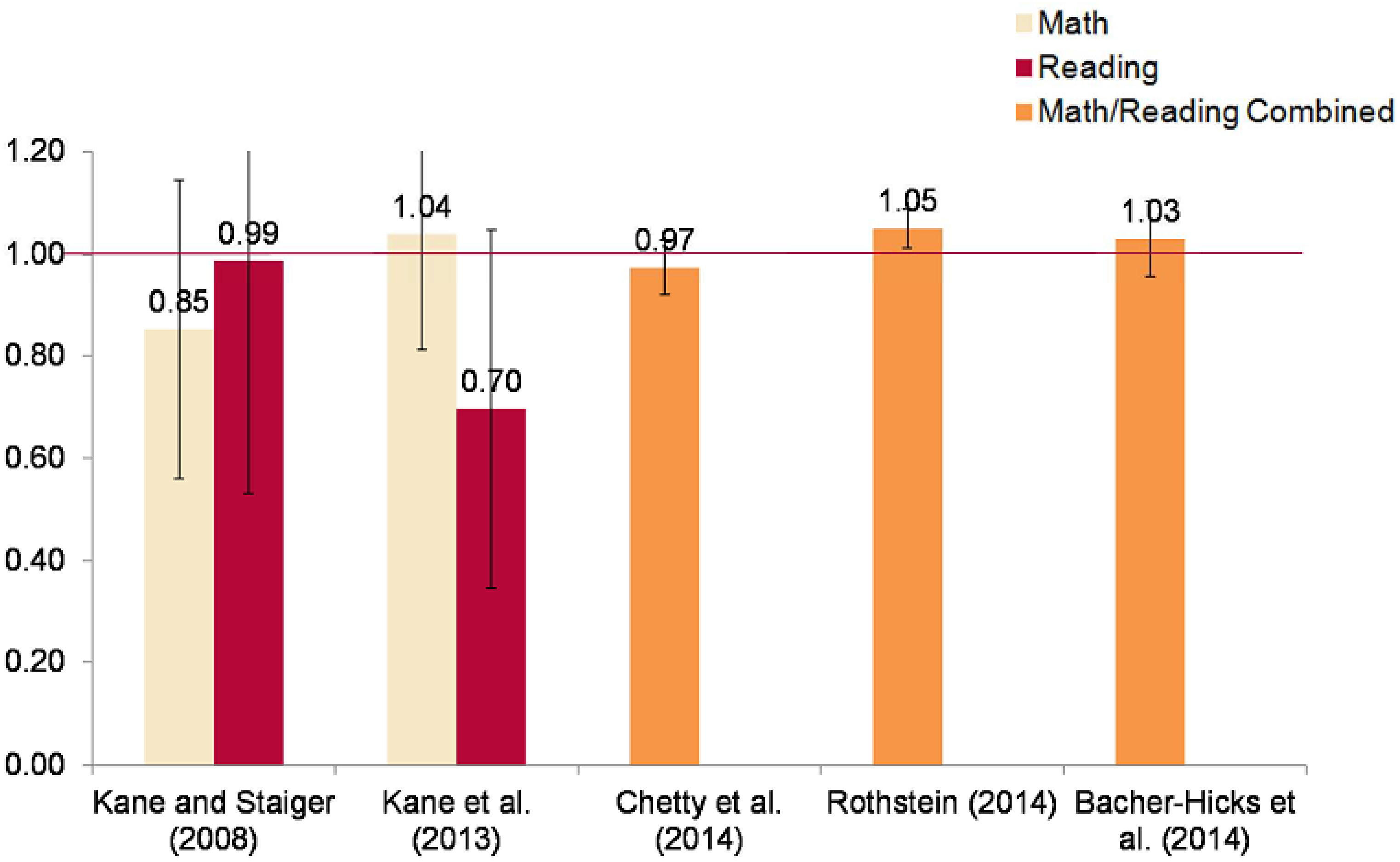
    - Drawback: does not necessary account for dynamic selection effects

Other Approaches to Evaluating Bias

2. Out-of-sample experiments/quasi-experiments
   [Kane and Staiger 2008]

   - Randomly assign new students to teachers and test whether prediction coefficient on VA estimates is 1
   - More difficult to implement than tests for balance and typically yields less precise estimates
   - But several studies have now estimated forecast bias in VA models in education using this approach
   - Glazerman and Protnik (2014) present a summary of estimates for teacher VA models

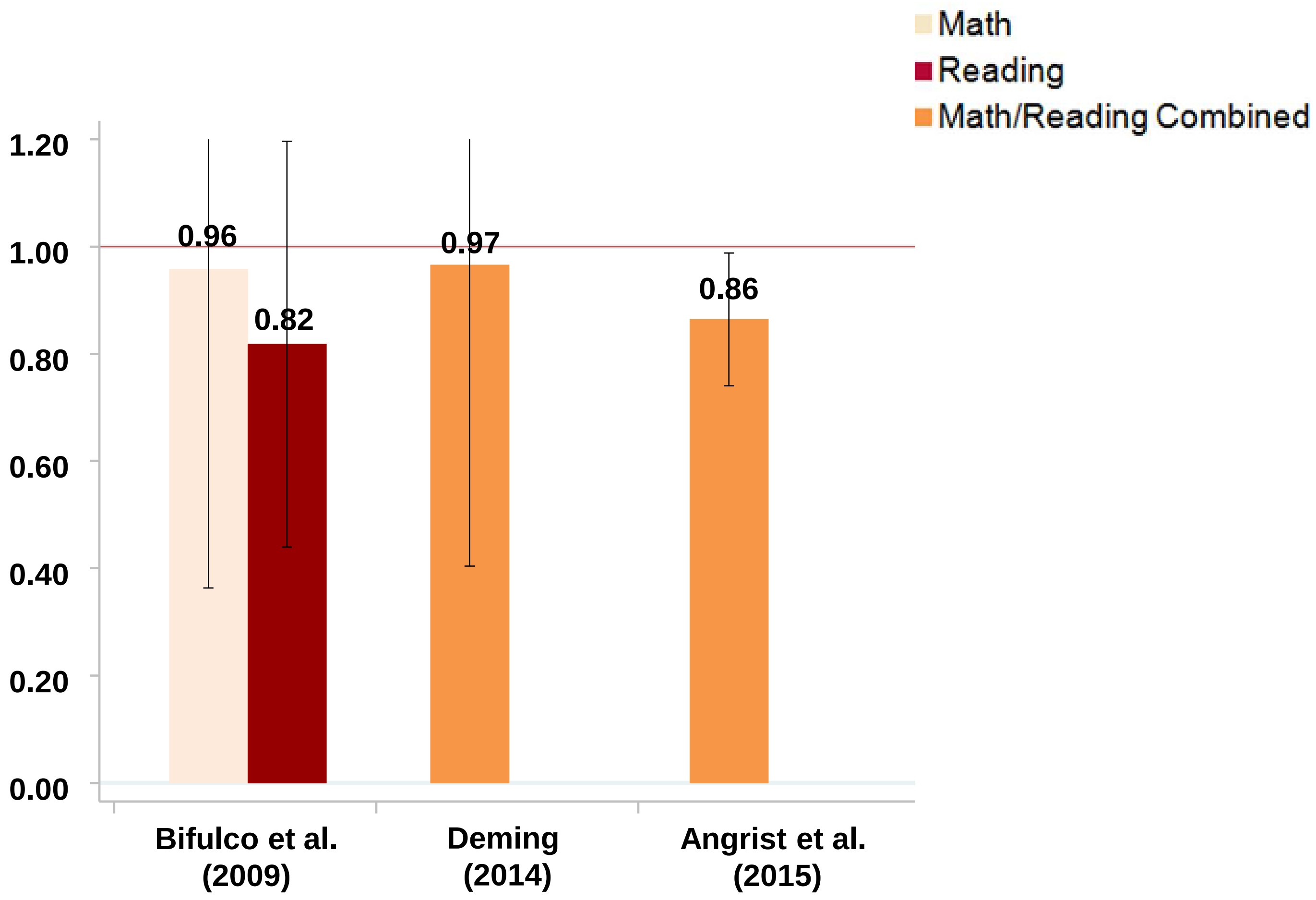
Experimental/Quasi-Experimental Estimates of Forecast Bias in Teacher VA



Note: error bars represent 90% confidence intervals

Source: Glazerman and Protik (2014). "Validating Value-added Measures of Teacher Performance"

Experimental/Quasi-Experimental Estimates of Forecast Bias in School VA



Note: error bars represent 90% confidence intervals

Sources: Bifuclo, Cobb, and Bell (2009, Table 6, Cols 1 and 3);
Deming (2014, Table 1, Col 6); Angrist et al. (2015, Table 3, Col 3)

Conclusion

- Estimation of VA creates a complex error structure for the treatment that is correlated with prior outcomes in non-transparent ways
  - Makes tests for bias using lagged outcomes more sensitive to model specification than when treatment is directly observed
- Experimental/quasi-experimental methods provide an approach to assessing bias that is less sensitive to model specification
- Potential directions for future work:
  - Compare alternative VA estimators when model is misspecified
  - In addition to measuring bias, gauge welfare gain from using biased estimates [e.g., Angrist, Hull, Pathak, Walters 2015]