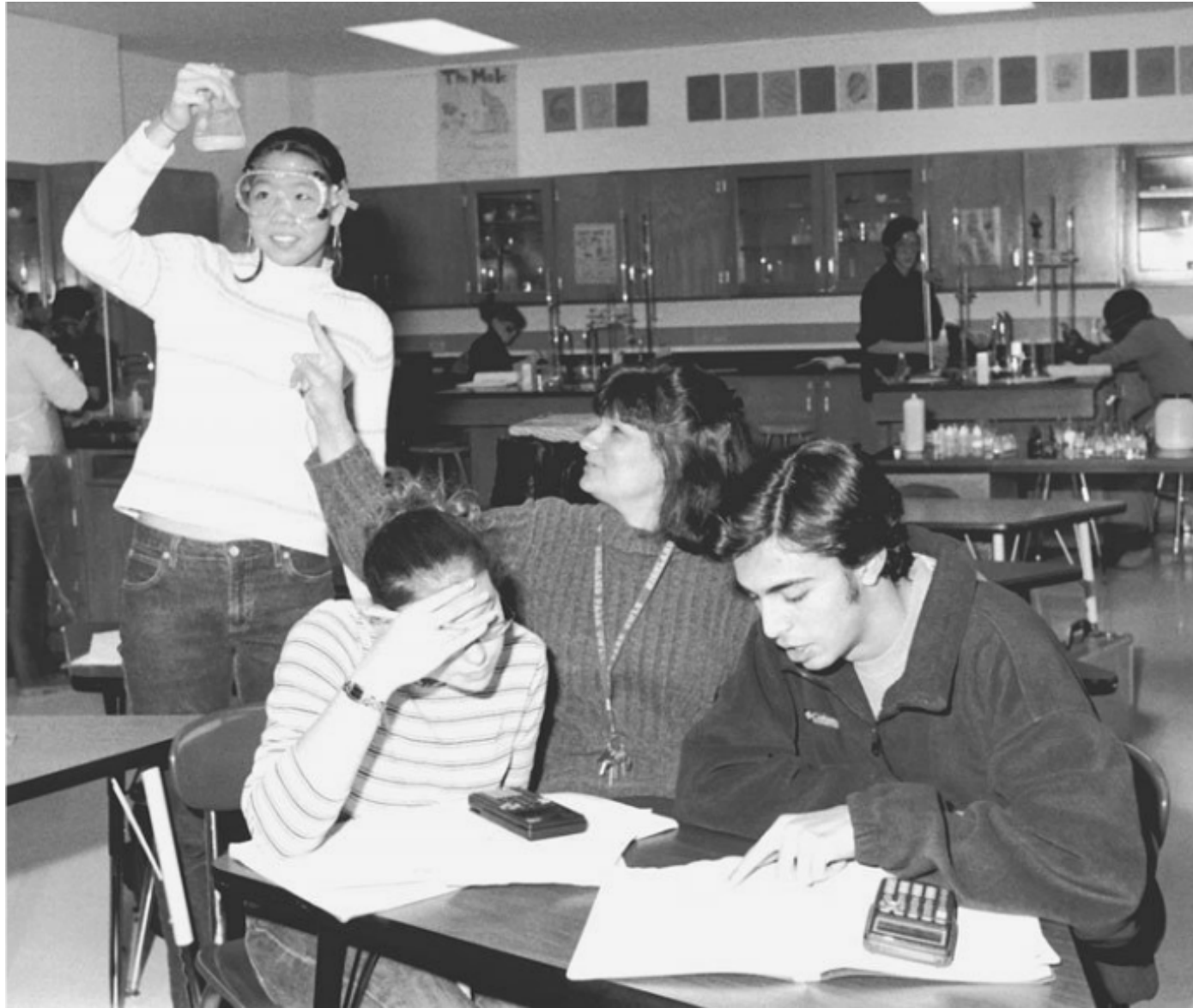


Evaluating Teachers

Jonah Rockoff
Columbia Business School

Is This Good Teaching?



Evaluating Teachers is Not Simple

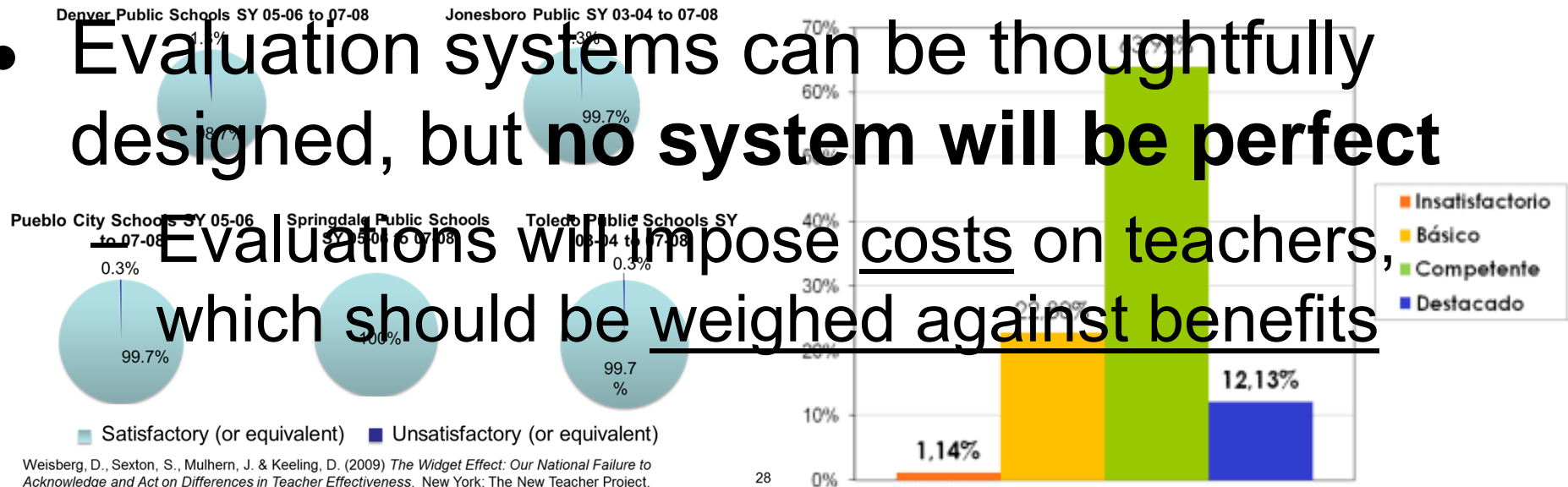
- Measuring teacher performance is a thorny problem for researchers and policymakers
 - Teaching is a highly complex job
 - Education has multiple (many?!) goals
 - Outcome measurement is infrequent/incomplete
 - Scaling and weighting of measures is non-trivial
 - Heterogeneity of other “inputs” into the educational process (e.g. home environment)
 - Highly regulated sector, often unionized labor

Evaluating Teachers is Important

- Performance varies, as in any professional occupation, and **evaluation is worthwhile**
- Most evaluation systems are not rigorous or provide little differentiation, and do a dis-service to teachers and students

- Evaluation systems can be thoughtfully designed, but **no system will be perfect**

Evaluations will impose costs on teachers,
which should be weighed against benefits



Measurement Approaches

1. “Quality” of inputs

- Credentials (e.g. masters degree, certification)
- Classroom observation ratings
 - Often perfunctory and meaningless, but recent movement to “2.0” observation processes

2. “Effectiveness” on outcomes

- Measurement of student learning growth

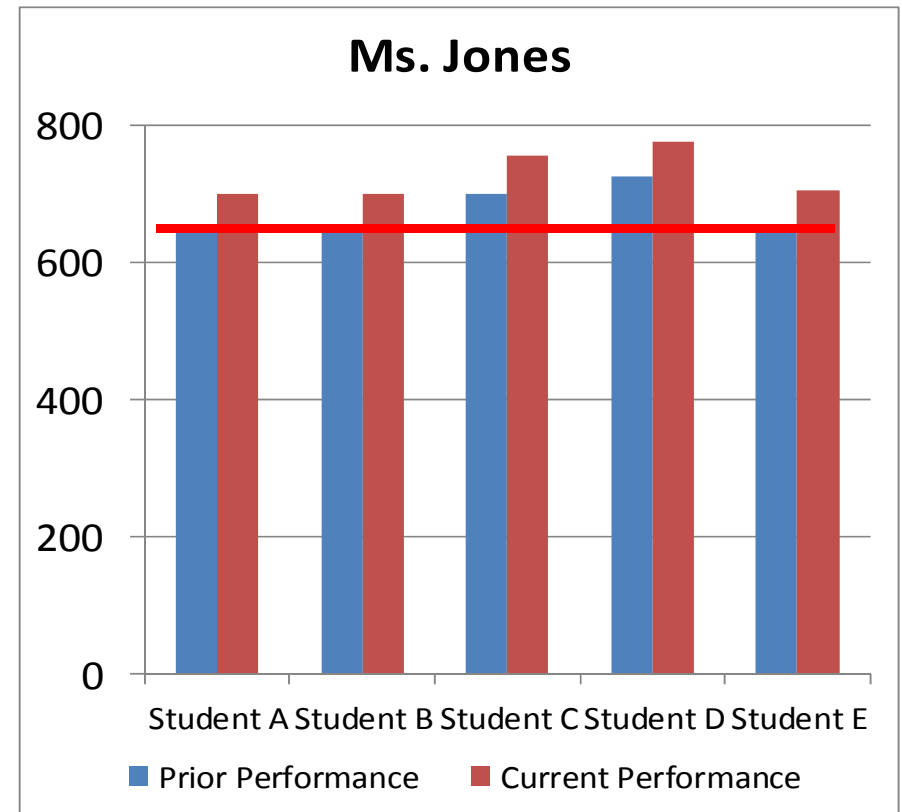
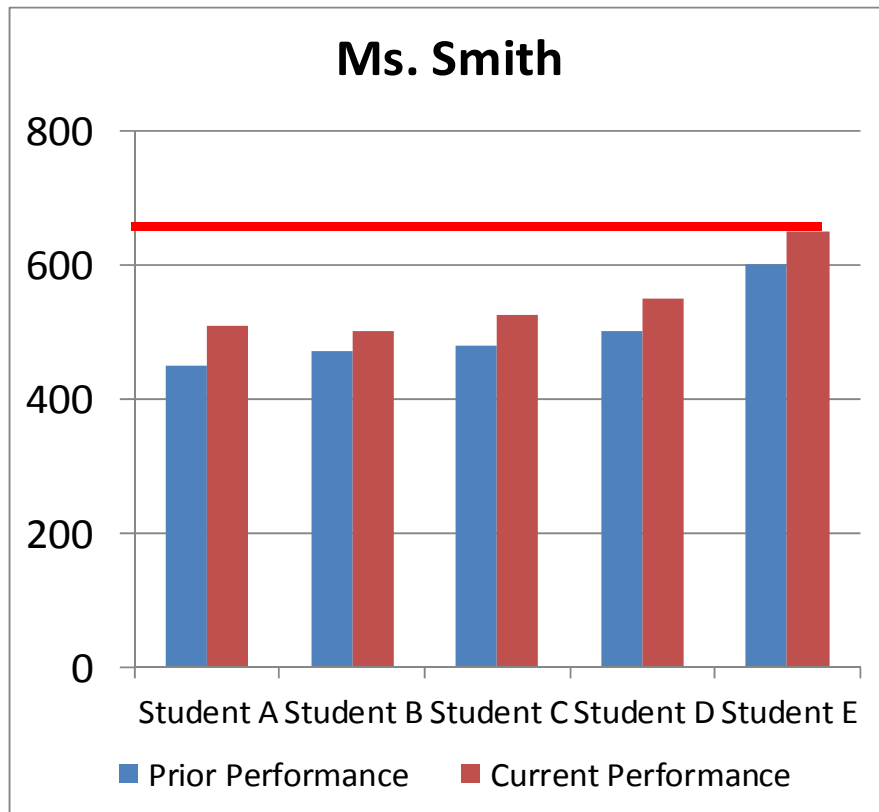
Measuring Teachers w/ Outcomes

- If we really care about student learning, why not measure it directly rather than infer it?
- Main hurdle: classrooms are not identical



The Measurement Problem (Part 2)

- Now which teacher is more effective??



— Proficiency

Ms. Smith avg growth = 47 points

Ms. Jones avg growth = 50 points

Basics of Value Added Analysis

- Value added: comparing actual student growth to student-specific benchmark
 - Accurate benchmarking of student growth is the big challenge in this type of analysis
- Generate benchmarks with (lots of) data
 - Regression analysis: examine growth for students who are similar on many dimensions including (and especially) prior achievement

Basic Findings from VA Research

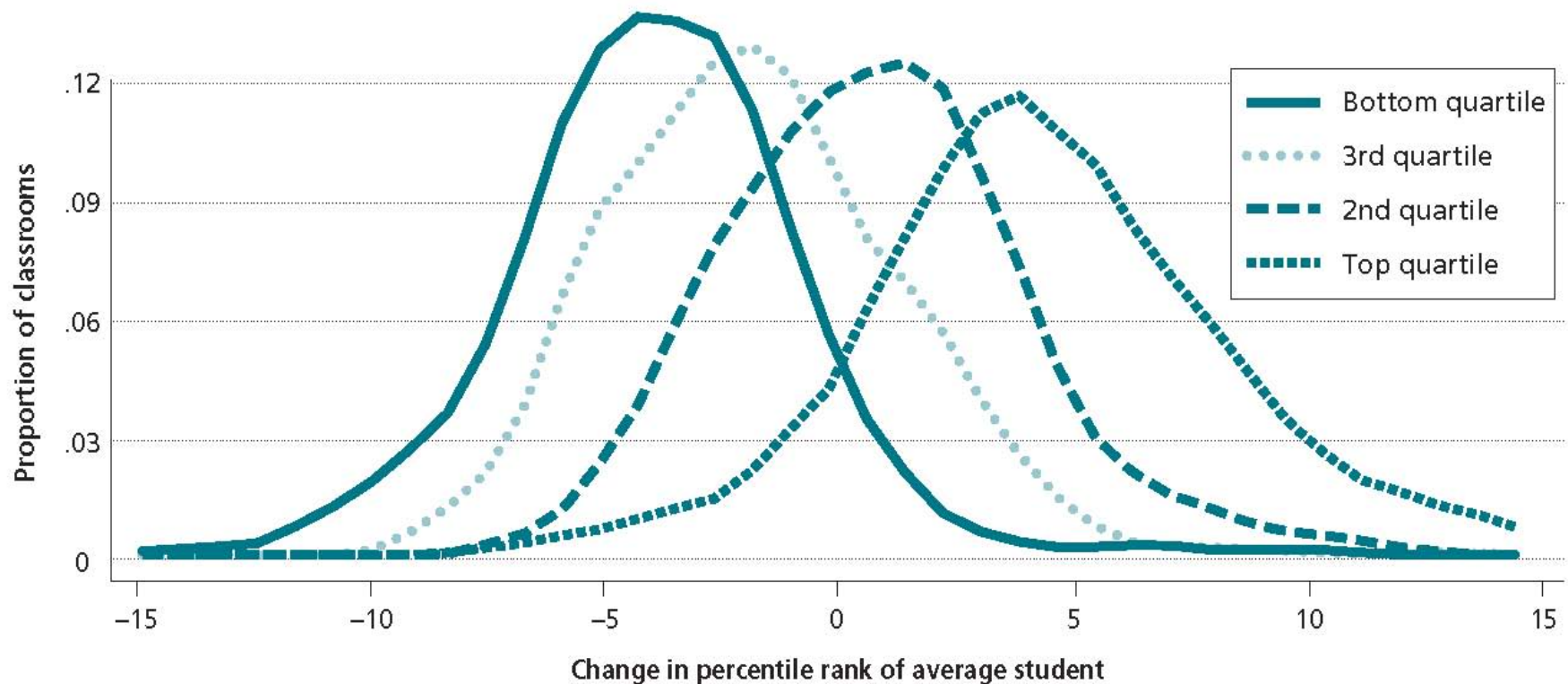
- Substantial variation in VA across teachers
 - Difference between teachers at 75th and 25th percentiles is $\sim 1/5^{\text{th}}$ of the achievement gap between poor & non-poor students in the U.S.
 - Much of the variation is within schools
- VA estimates predict future teacher with sufficient stability to be useful for policy
 - Year to year reliability in the 0.3 - 0.5 range
 - “Year to career” reliability of 0.5 – 0.7

Imperfect and Useful Performance Data



How Predictive is Value Added?

- Teacher performance in third year, separated by two-year quartile ranking



Gordon, R., Kane, T., Staiger, D. (2006). Identifying effective teachers using performance on the job. *Brookings Institution Hamilton Project Paper*.

But Raising Scores Is Not The Goal

- Do high VA teachers cause students to have better outcomes later in life?
- In addition to answering this question, we also present two new pieces of evidence on potential biases in value-added
- Test #1: VA is uncorrelated with parental characteristics omitted from our model
- Test #2: VA predicts grade-level scores when teacher assignments change



The Impact of Individual Teachers on Student Achievement:
Evidence from Panel Data

By RONAH B. ROCKOFF

School administrators, parents, and students themselves widely support the notion that teacher quality is vital to student achievement, despite inconsistent evidence linking achievement to observable teacher characteristics. Eric Hanushek (1986) has had many observers to conclude that, while teacher quality may be important, variation in teacher quality is driven by characteristics that are difficult or impossible to measure. Research that has therefore come to focus on using matched student-teacher data to separate student achievement into a series of “fixed effects,” and assigning importance to individuals, teachers, schools, and so on.

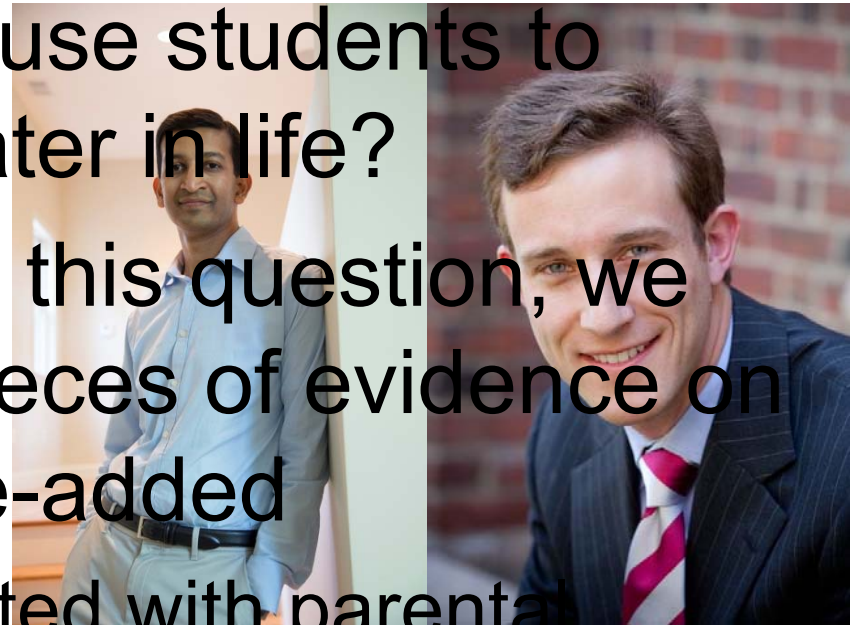
Credible identification of teacher fixed effects requires matched student-teacher data wherein both student achievement and teachers are observed in multiple years. This type of data is not readily available to researchers, in large part because school districts do not use panel data for evaluation purposes. In previous studies, researchers have either collected information directly from school districts (Hanushek, 1971; Richard Murnane, 1975; David Armor et al., 1976; Albert Park and Emily Hanum, 1996) or used a random sample of schools and collected data on teachers’ attributes (Dennis Aaronson et al., 2003; Steven G. Rivkin et al.,

2003). Almost all of the empirical difficulties in these studies are related to data quality. For instance, teacher effects cannot be separated from other classroom-specific factors in several of these studies because teachers are only observed with one class of students.

I use a rich set of panel data on student test scores and teacher assignment to estimate more accurately how much teachers affect student achievement. Panel data on students’ test scores allows for a focus on differences in the performance of the same student with different teachers, and thus to distinguish variation in teacher quality from variation in students’ cognitive abilities and other characteristics. Observing the same teacher with multiple classrooms allows me to differentiate teacher quality from factors such as class size. In addition, by focusing on variation in student achievement within particular schools and years, I separate variation in teacher quality from variation in school-level educational inputs (despite possible quality and time-varying inputs that affect test performance at the school level).

This analysis extends research on teacher quality in two additional ways. First, I use a random effects model and a fixed effects model to take explicit account of estimation error. Since estimation error will bias upward the variance of the distribution of teacher fixed effects, the corrected measure provides a more accurate portrayal of the within-school variation in teacher quality. Second, I examine the relation between student achievement and teaching experience using variation across years for individual teachers. This strategy will not confound the causal effect of teaching experience with other random selection based on teacher quality or differences in teacher quality across cohorts.

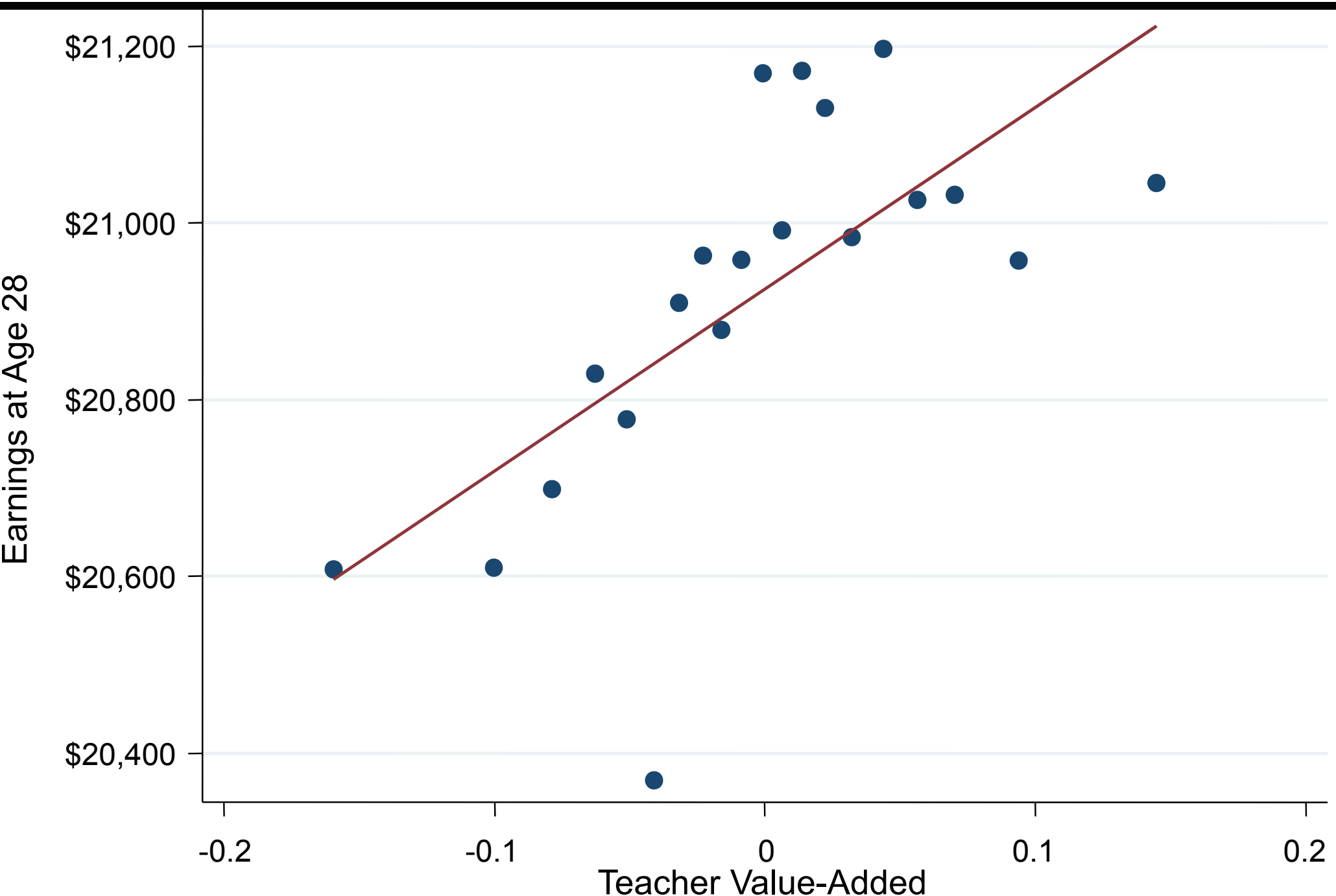
My empirical results indicate large differences in quality among teachers within schools. A one-standard-deviation increase in teacher quality raises test scores by approximately 0.1 standard deviations in reading and math on



* Department of Economics, Harvard University. I am grateful to David LaParo and D. R. L. I thank Gary Chamberlain, John Hoxby, Rick Kane, Caroline Hoxby, Bryan Jacob, Christopher Jencks, Larry Katz, Kevin Lang, Richard Murnane, and Steve Rivkin for extensive comments, as well as Raj Chetty, Bryan Graham, Adam Looney, Sarah Reber, Tara Watson, and seminar participants at Harvard University. I also thank the following for their helpful suggestions on this study: Amy Smith, Pauline Loe, and the local education officials who worked with me to make this project possible. This work was supported by a grant from the Inequality and Social Policy Program at Harvard’s Kennedy School of Government.

¹ The Tennessee Value Added Assessment System, where districts, schools, and teachers are compared based on test-score gains averaged over a number of years, is a noteworthy exception.

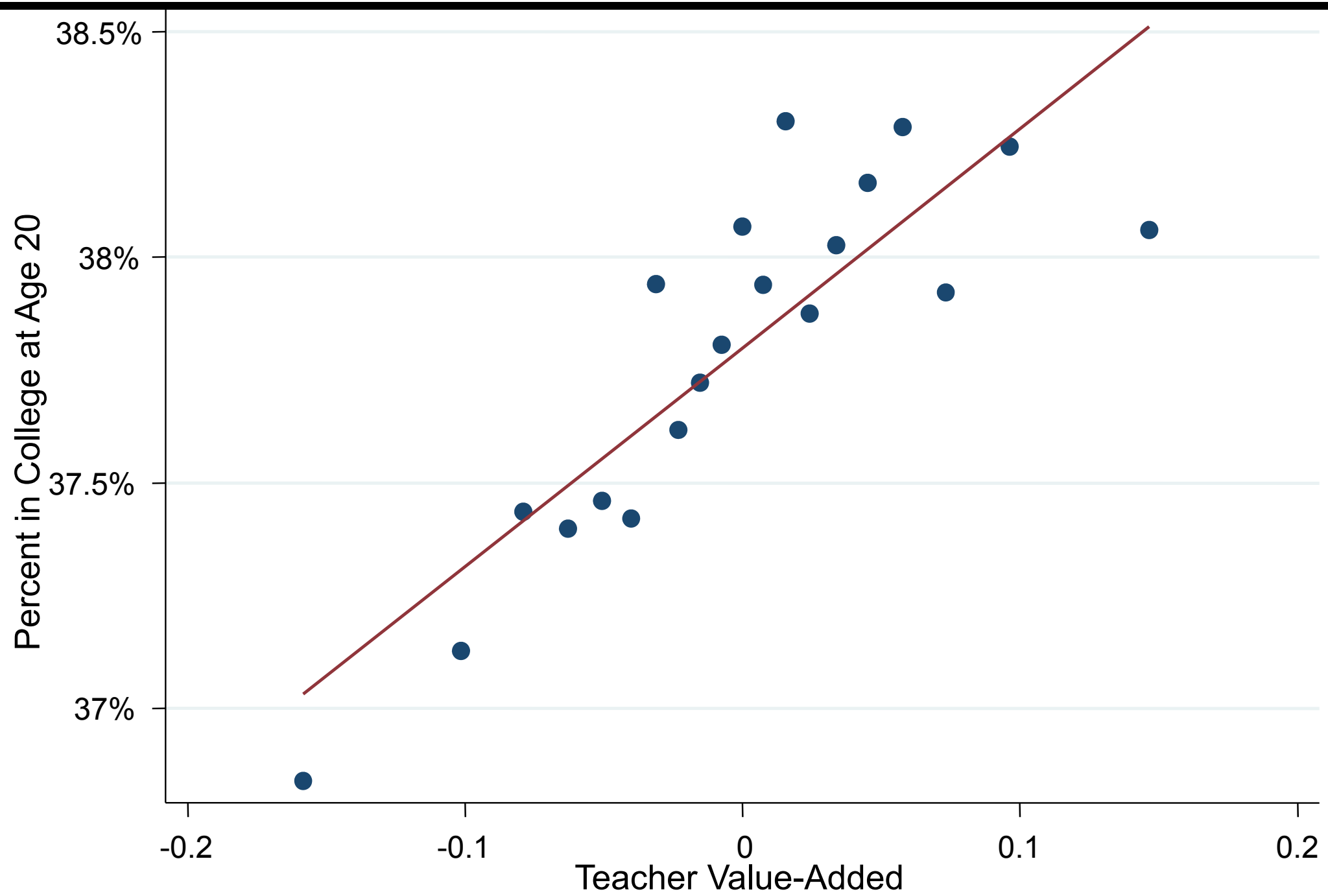
Lasting Gains from High VA Teachers



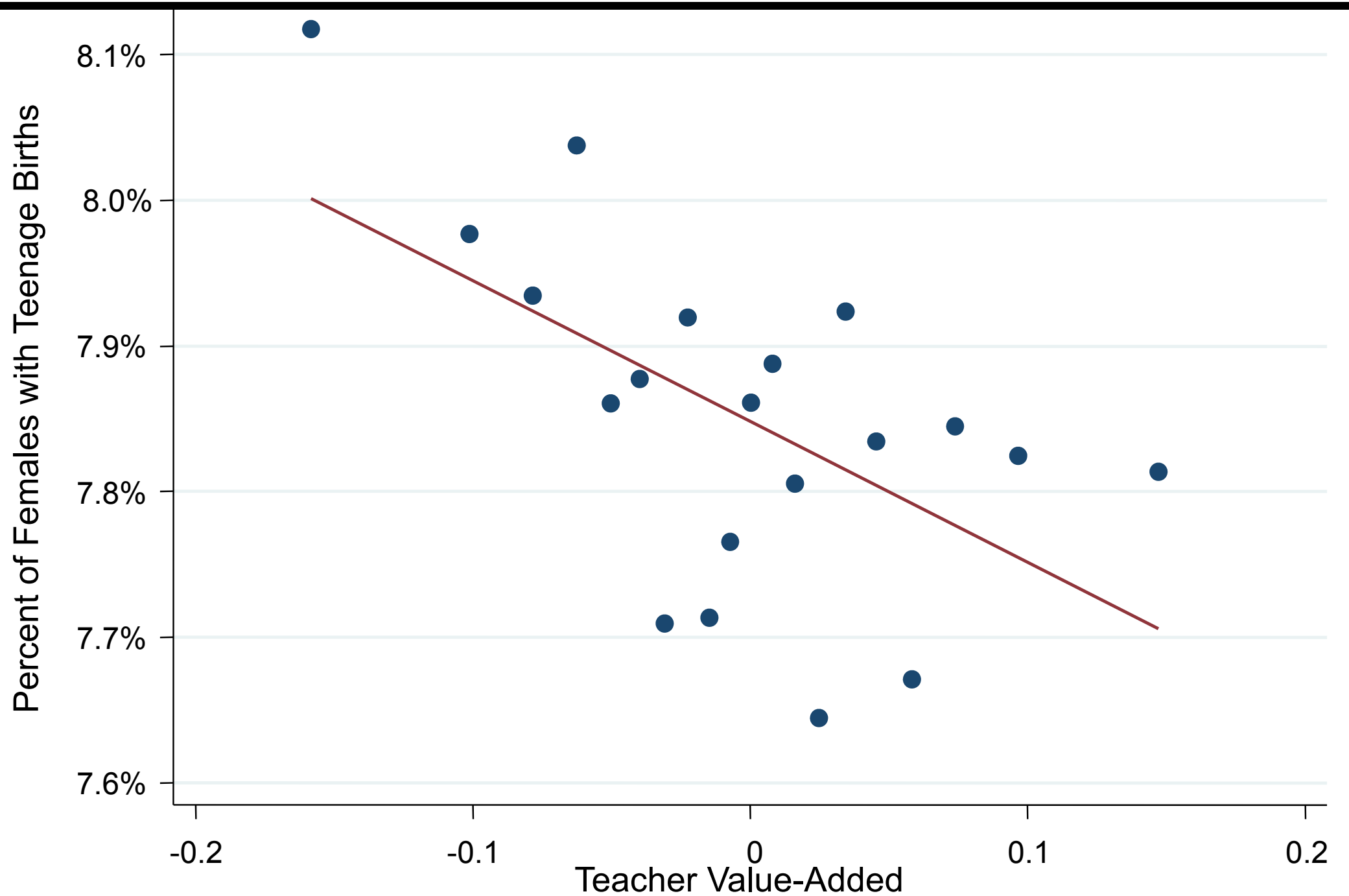
What is a Great Teacher Worth?

- What is the impact of having a top 5% teacher on present value of lifetime earnings for a class of average size (28 students)?
 - Relative to having an average teacher
- Result: \$266,000 per classroom
- In addition, we find impacts of teachers on college attendance, retirement savings, teenage pregnancy, neighborhood quality,...

Lasting Gains from High VA Teachers



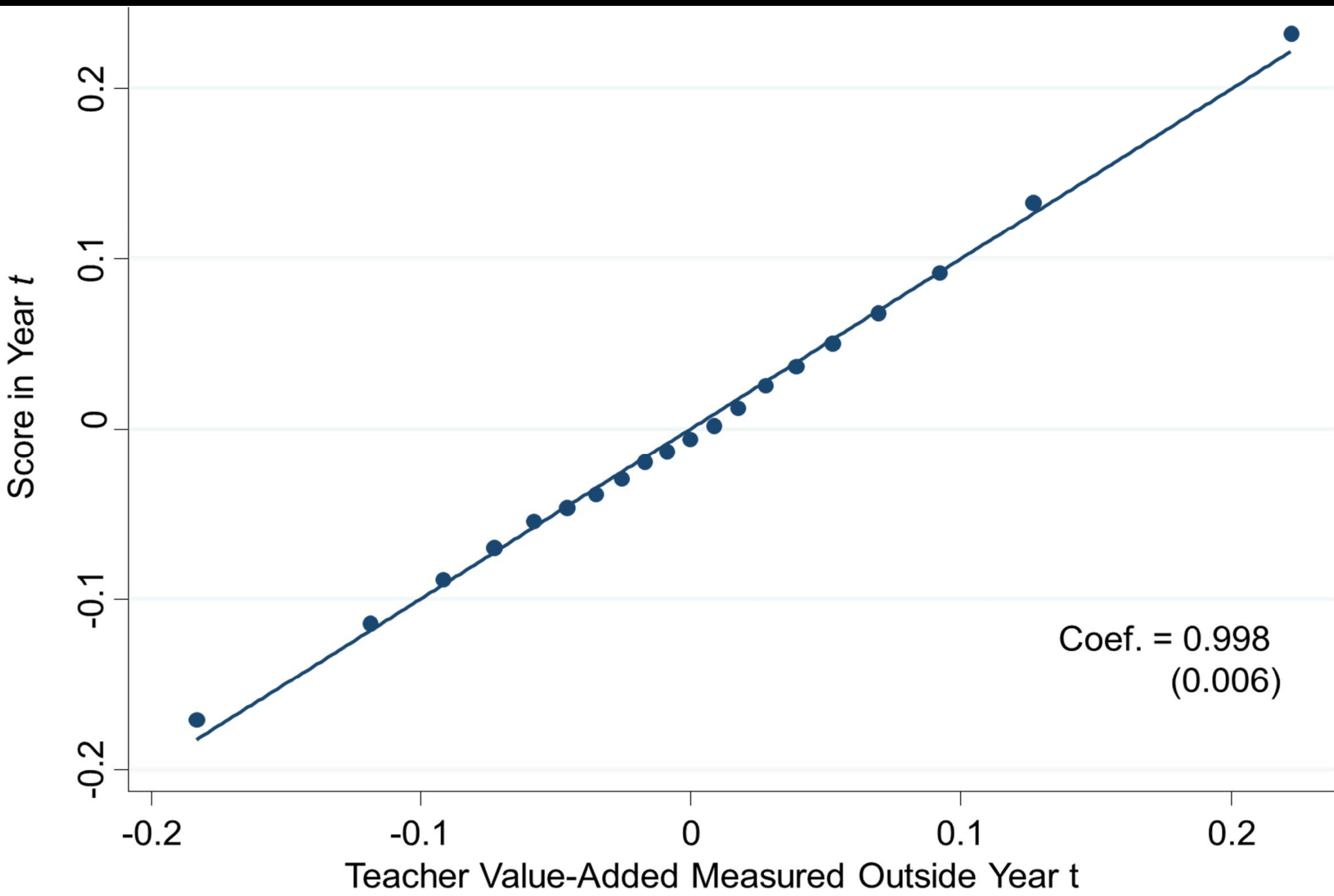
Lasting Gains from High VA Teachers



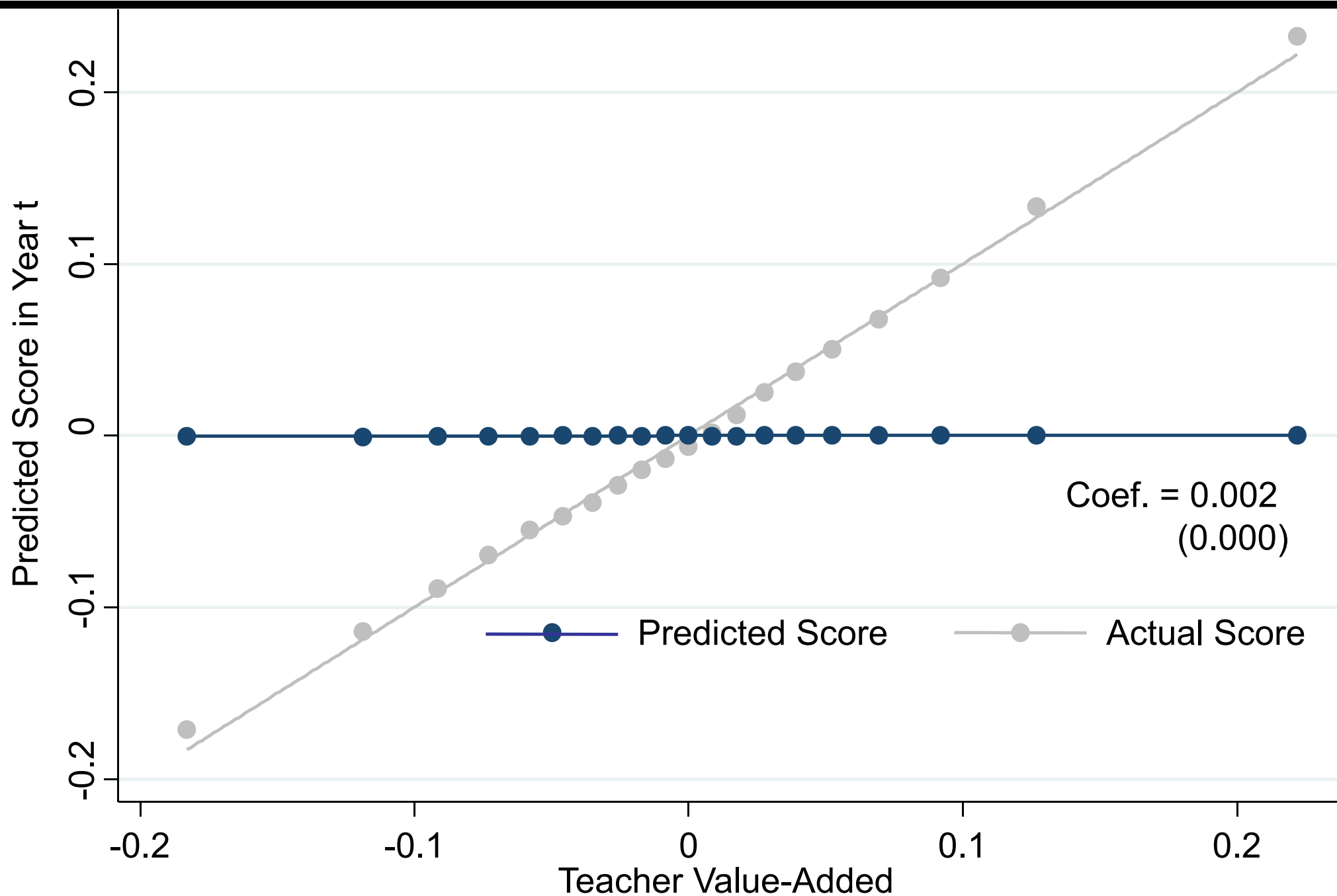
New Tests for Bias

- Test #1: Performance should be unrelated to omitted parental characteristics
 - Conditional on controls for prior tests, etc.

Predicting Performance Across Years



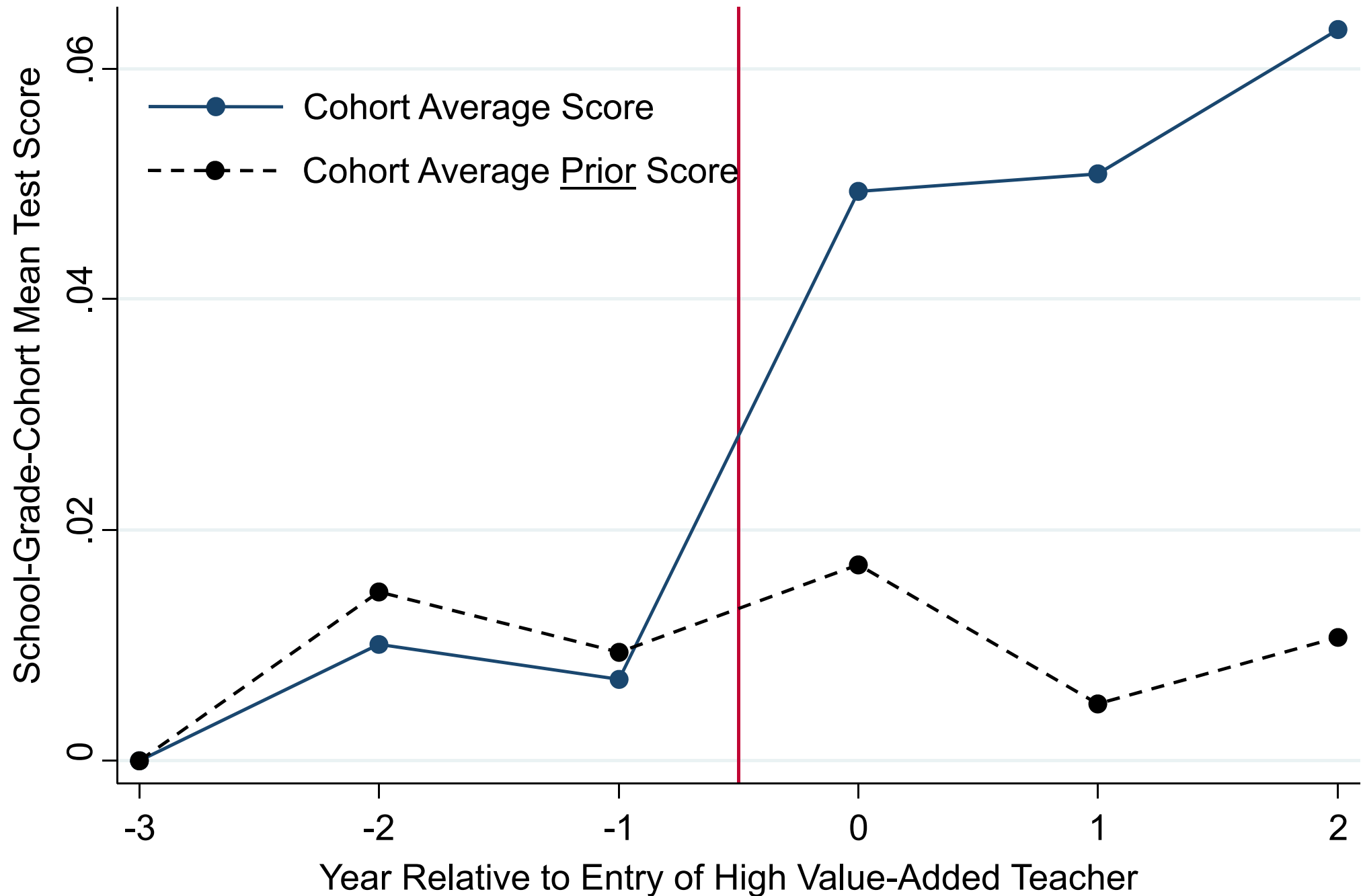
Test #1: Predicted Score Based on Parents



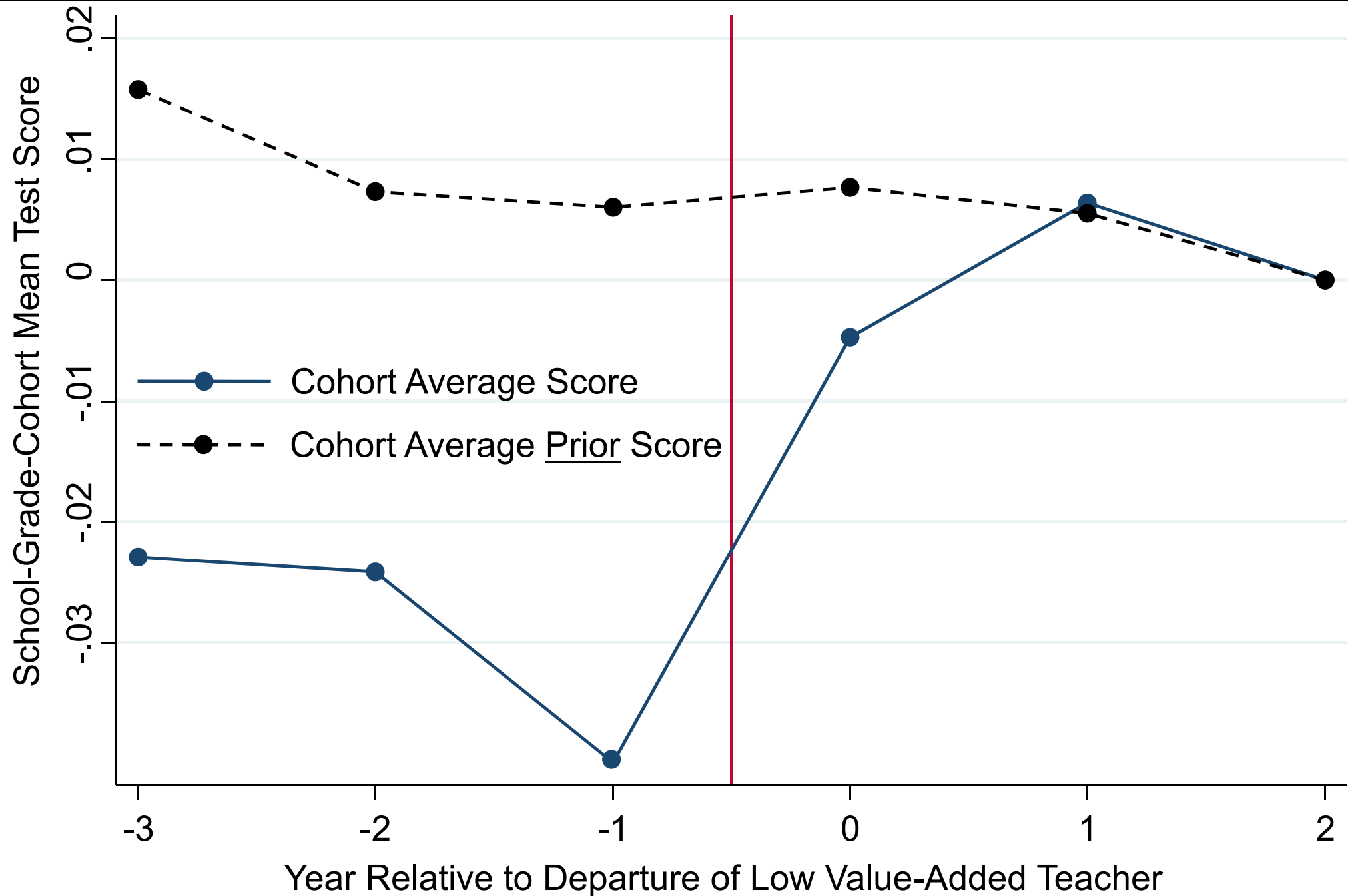
New Tests for Bias

- Test #1: Performance should be unrelated to omitted parental characteristics
- Test #2: Performance should be able to predict grade-level changes in scores using changes in teacher assignments

Test #2: Highly Effective Teacher Entry



Test #2b: Highly Ineffective Teacher Exit



Caveats/Limitations

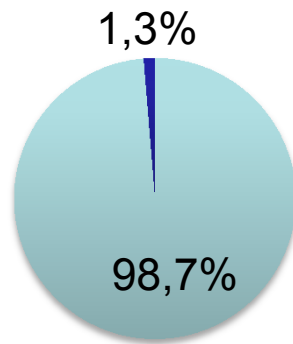
- Performance may be context specific
 - Though evidence suggests performance persists across grades/schools within subject
 - USDOE Teacher Transfer Initiative
- Availability limited to tested grades/subjects
 - Requires a measure of student learning growth
- A relative performance metric
 - Half the teachers will always be below average
- Essentially summative, not formative
 - Batting average analogy is helpful

Can We “Know It When We See It”?

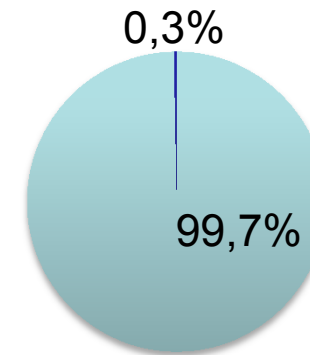
- “Any real educator can know within five minutes of walking into a classroom if a teacher is effective.”
 - Randi Weingarten, United Federation of Teachers
- Can principals and/or peer teachers observe effective teaching in action?
 - Not nearly as easy as the quotation suggests

Current Observation-Based Evaluations Do Not Differentiate Among Teachers in U.S.

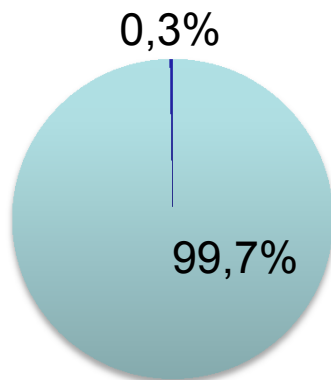
Denver Public Schools SY 05-06 to 07-08



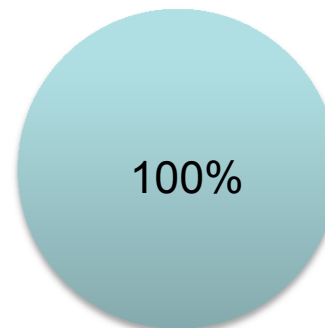
Jonesboro Public SY 03-04 to 07-08



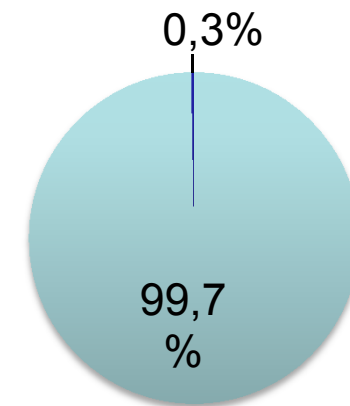
Pueblo City Schools SY 05-06 to 07-08



Springdale Public Schools SY 05-06 to 07-08



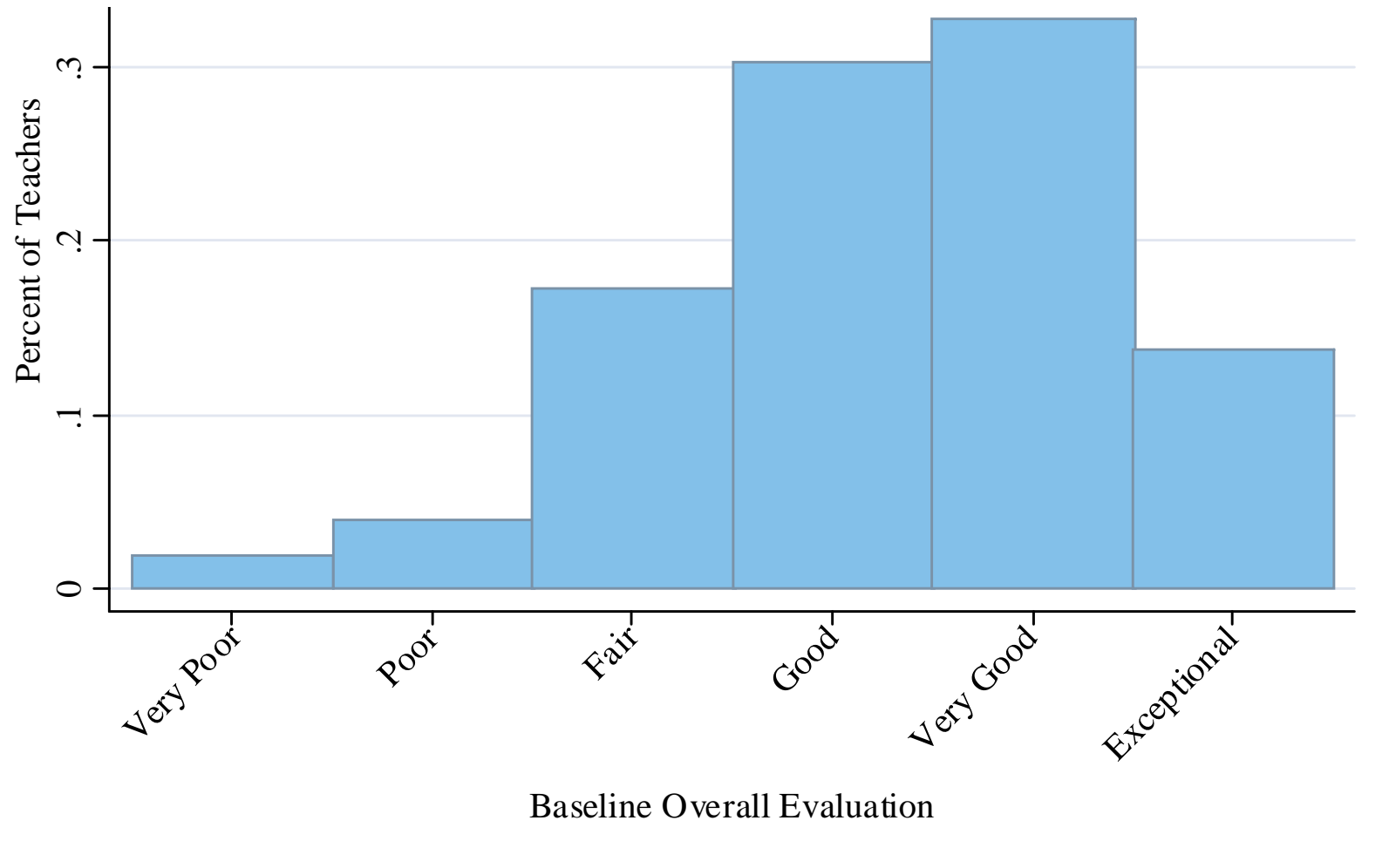
Toledo Public Schools SY 03-04 to 07-08



■ Satisfactory (or equivalent) ■ Unsatisfactory (or equivalent)

But They *Could* Differentiate

- Substantial variation in principals' opinions in low-stakes survey evaluations

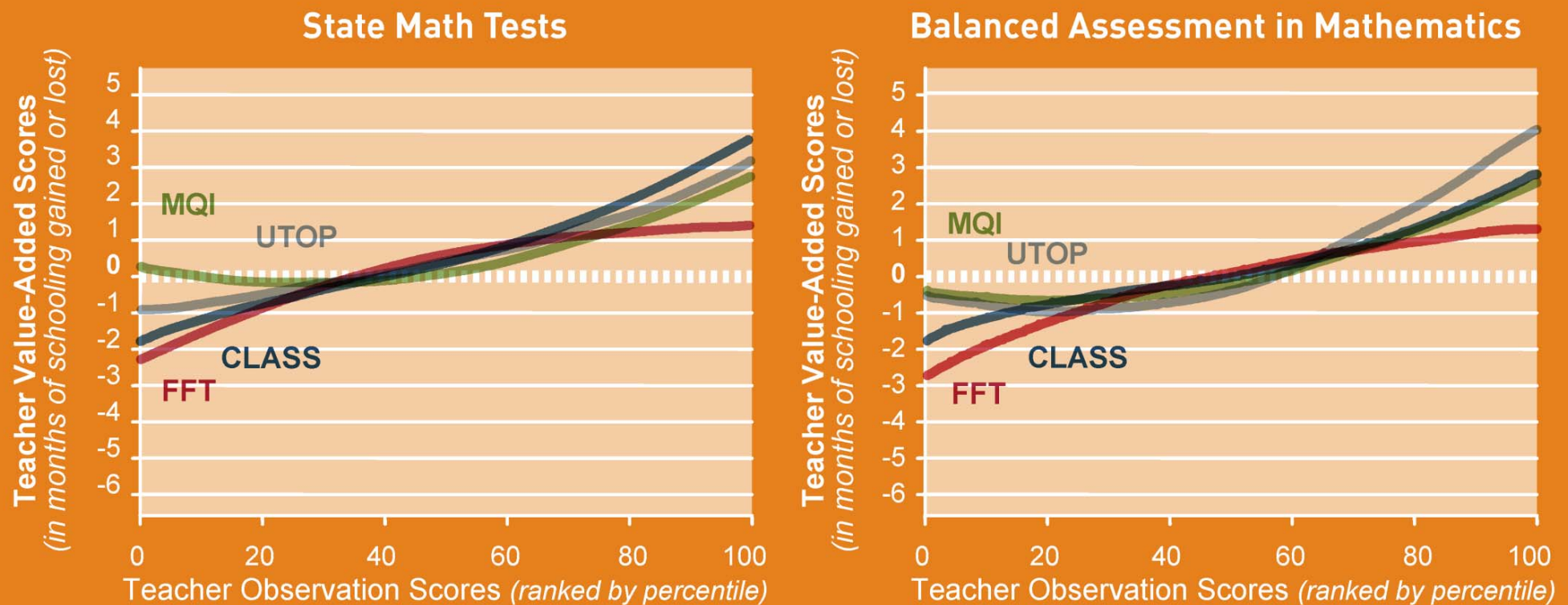


The MET Project (Gates Foundation)

- A rigorous effort to understand some basic facts regarding classroom observations
 - Large diverse sample from various states
 - Multiple observations of practice, scored on established rubrics by trained evaluators
 - Also look at student opinions and value-added
- Bottom line: measuring effective classroom teaching is possible, but reliability is difficult

Evaluations Do Correlate with Growth

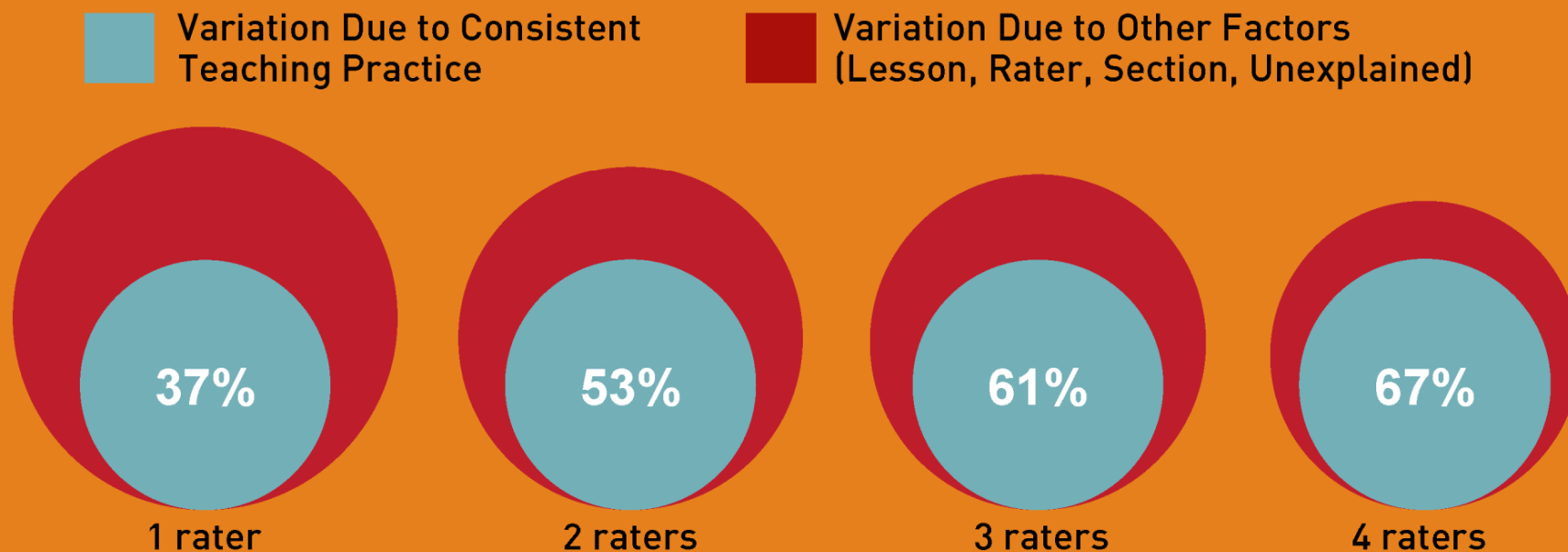
Figure 3. Teachers with Higher Observation Scores Had Students Who Learned More



Source: Measures of Effective Teaching (MET) Project

But High Reliability is Costly

Figure 9. Multiple Observations Led to Higher Reliability



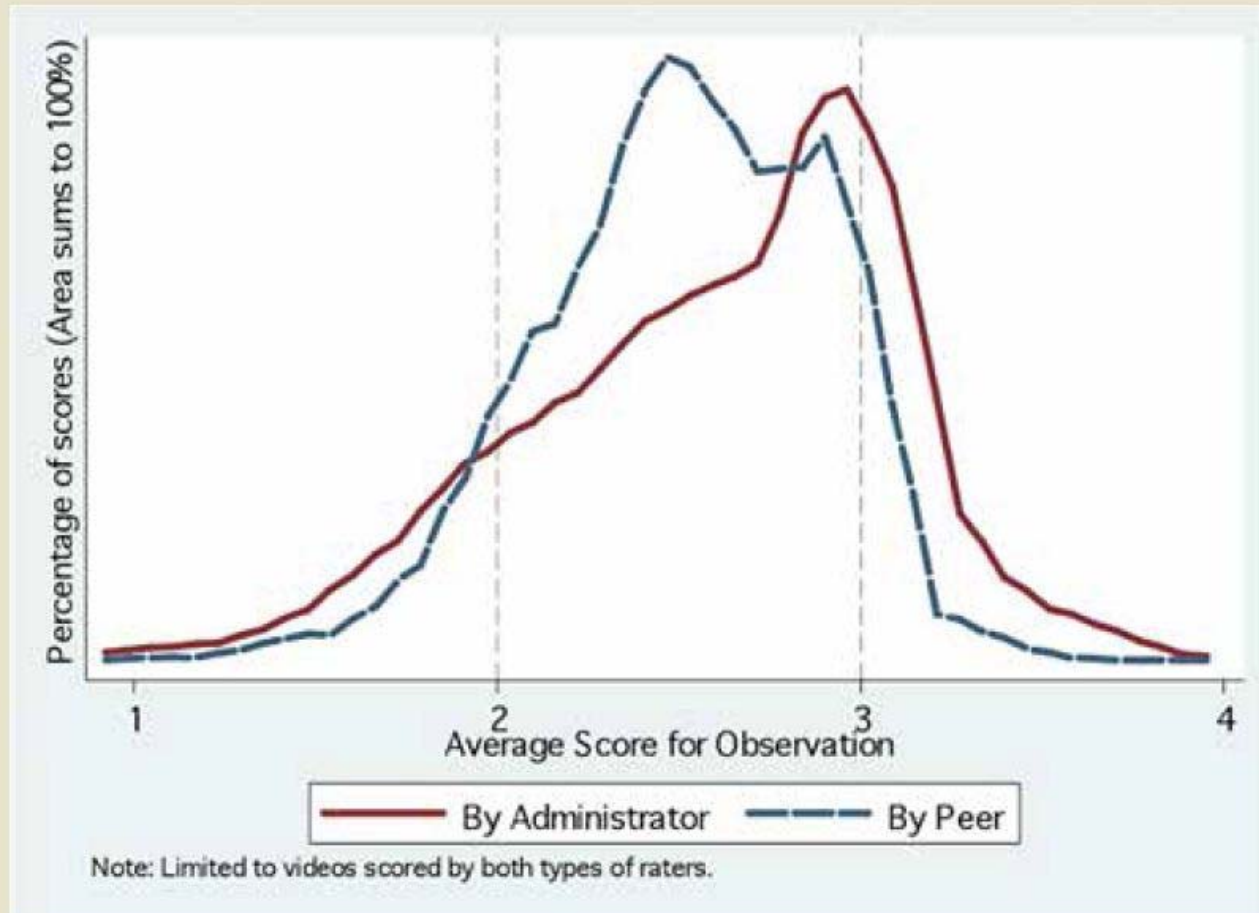
Each rater is observing a different lesson

NOTES: The number in each circle is the percentage of variance in average FFT scores attributable to teacher effects. The area of the inner circle represents the variance in aspects of teachers' practice that is consistent across lessons, while the area of the outer circle adds in variation due to other factors, such as rater disagreement and lesson-to-lesson variance. As the number of observations increases, the variance due to consistent teaching practice remains constant, while the variance due to other factors declines, as it is averaged over more observations.

Administrators vs. Peer Teachers

Figure 4

DISTRIBUTION OF OBSERVATION SCORES BY TYPE OF OBSERVER



Source: Measures of Effective Teaching (MET) Project

Bias in Lesson Selection

Figure 6

COMPARING TEACHER SCORES FOR LESSONS CHOSEN AND NOT CHOSEN



Source: Measures of Effective Teaching (MET) Project

The Role of Classroom Observations

- Measures like classroom observation have useful features that student growth lacks
 - Rely on specific and easily understood rubrics, rather than complicated formulae
 - Provide wide coverage of many grades subjects
 - Can have absolute performance measures, not just relative, given high quality training
 - Can be timely and used for development
 - May capture other dimensions of teaching unrelated to test score growth but valuable
- However, classroom practice is only relevant for performance insofar as it affects students

The Role of Classroom Observations

- Consistent evidence that observation-based evaluations of teaching practice are strongly related to gains in student achievement
 - Validated rubrics (CLASS, FFT, Marzano, etc.)
- But classroom observations have limitations!
 - May narrow teaching practice to fit a specific mold
 - A few lessons \neq a whole year of teaching
 - Correlation in scores across observations is ~ 0.4 , even with highly trained evaluators
 - Bias/preferential treatment could be problematic

Implications for Policy

- Teacher performance is crucial to students' success, academically and more broadly
- Value added is a very useful but incomplete method for evaluating teacher performance
 - Same is true of any other method, like classroom observation, portfolios, surveying students, etc.
- Evaluations should use multiple measures
 - Reliability increases substantially, as errors in measurement offset one another
 - Validity increases by creating a fuller picture of performance in reaching various teaching goals
 - Multiple measures reduce pressure to act in ways that improve evaluation but do not benefit students

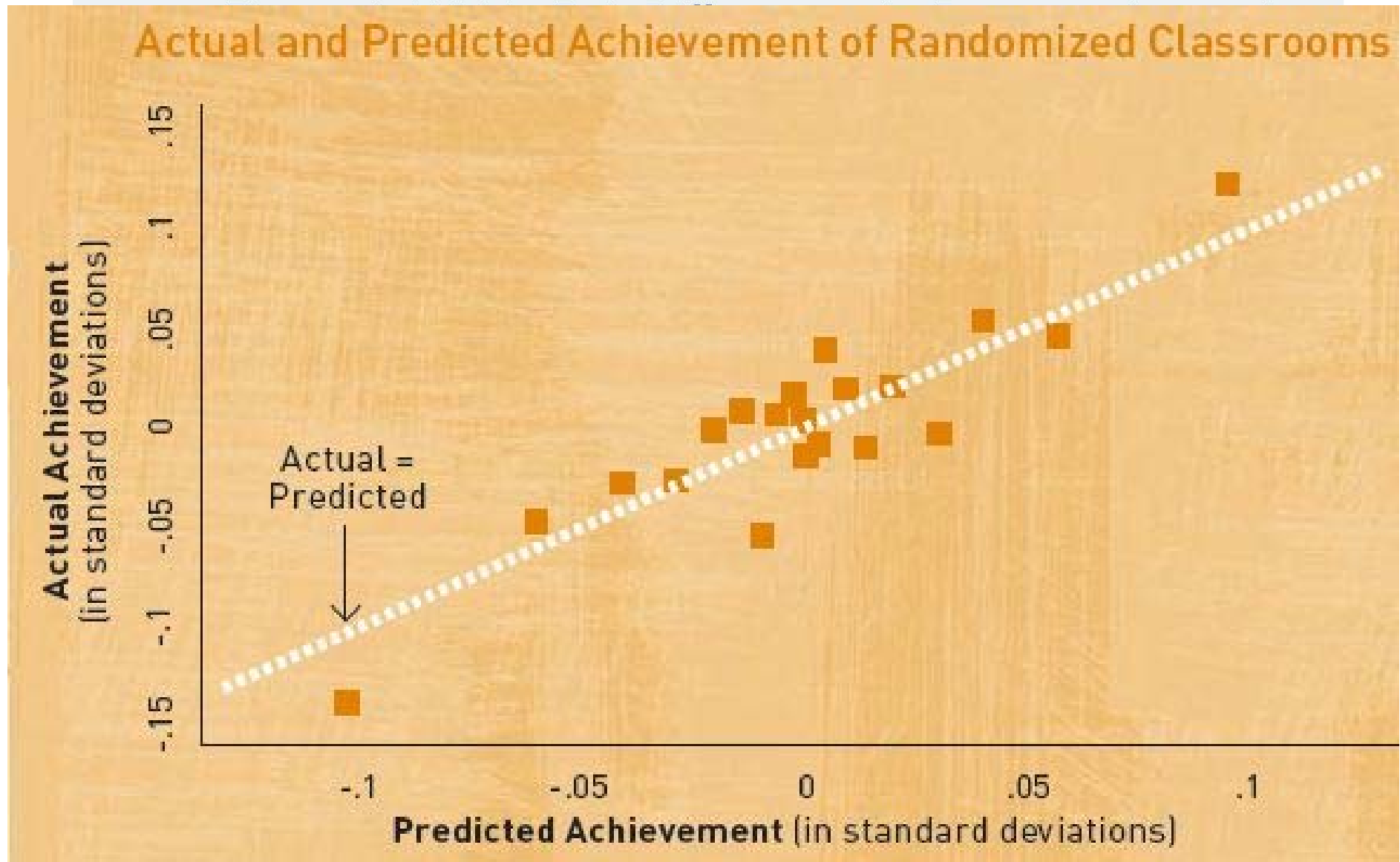
Thank You!

Extra Slides

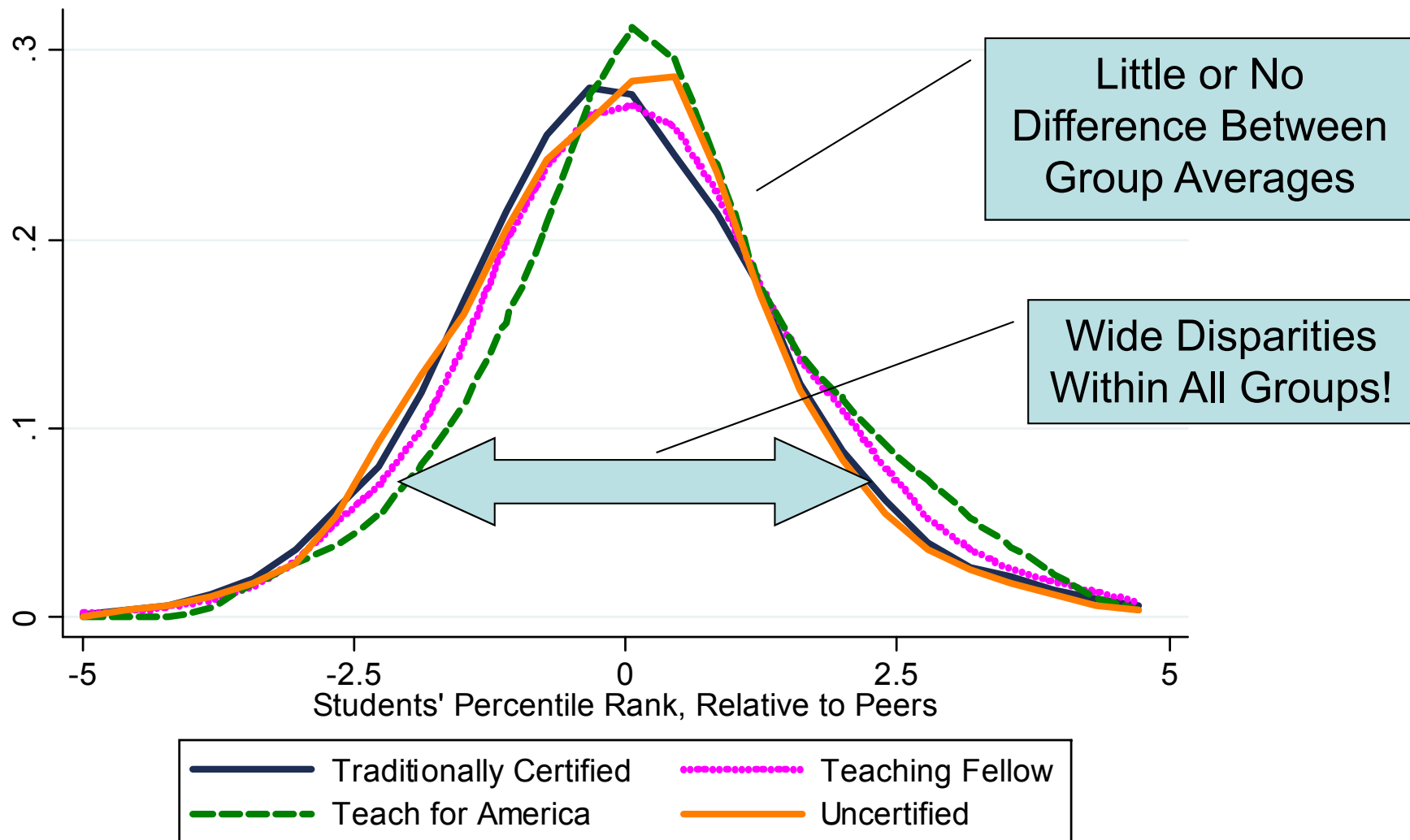
Threats to Validity

- Major concern is systematic student sorting
 - Unfair treatment of teachers that is systematic
 - Example: P's friends get “better” students
- Two random assignment studies find no evidence of bias, but samples are small
- In my recent work, two other pieces of evidence that value-added is not biased
 - Test #1: VA is uncorrelated with parental characteristics omitted from model
 - Test #2: VA predicts grade-level scores when teacher assignments change

Stability in Random Assignment

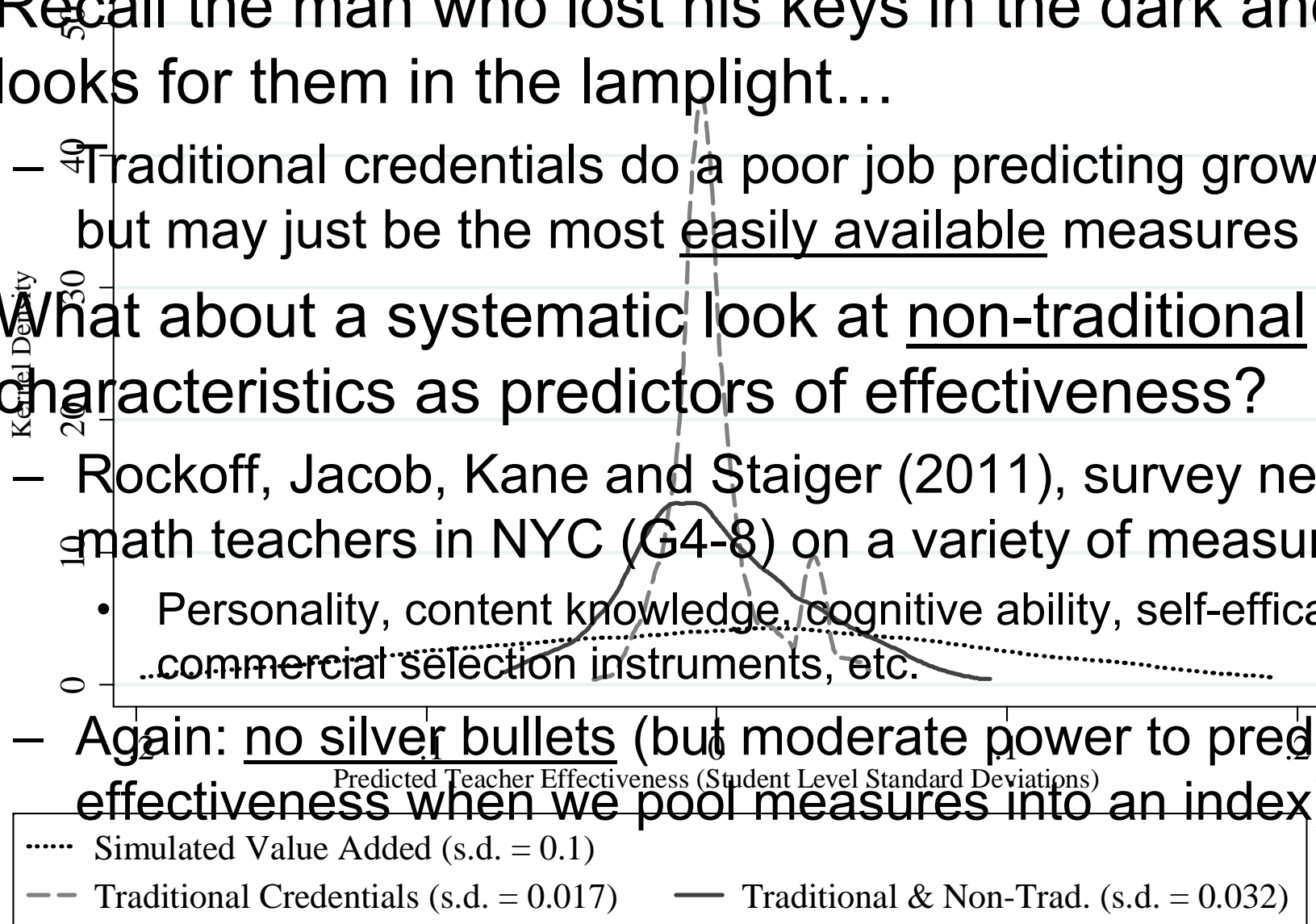


Research on Traditional Credentials

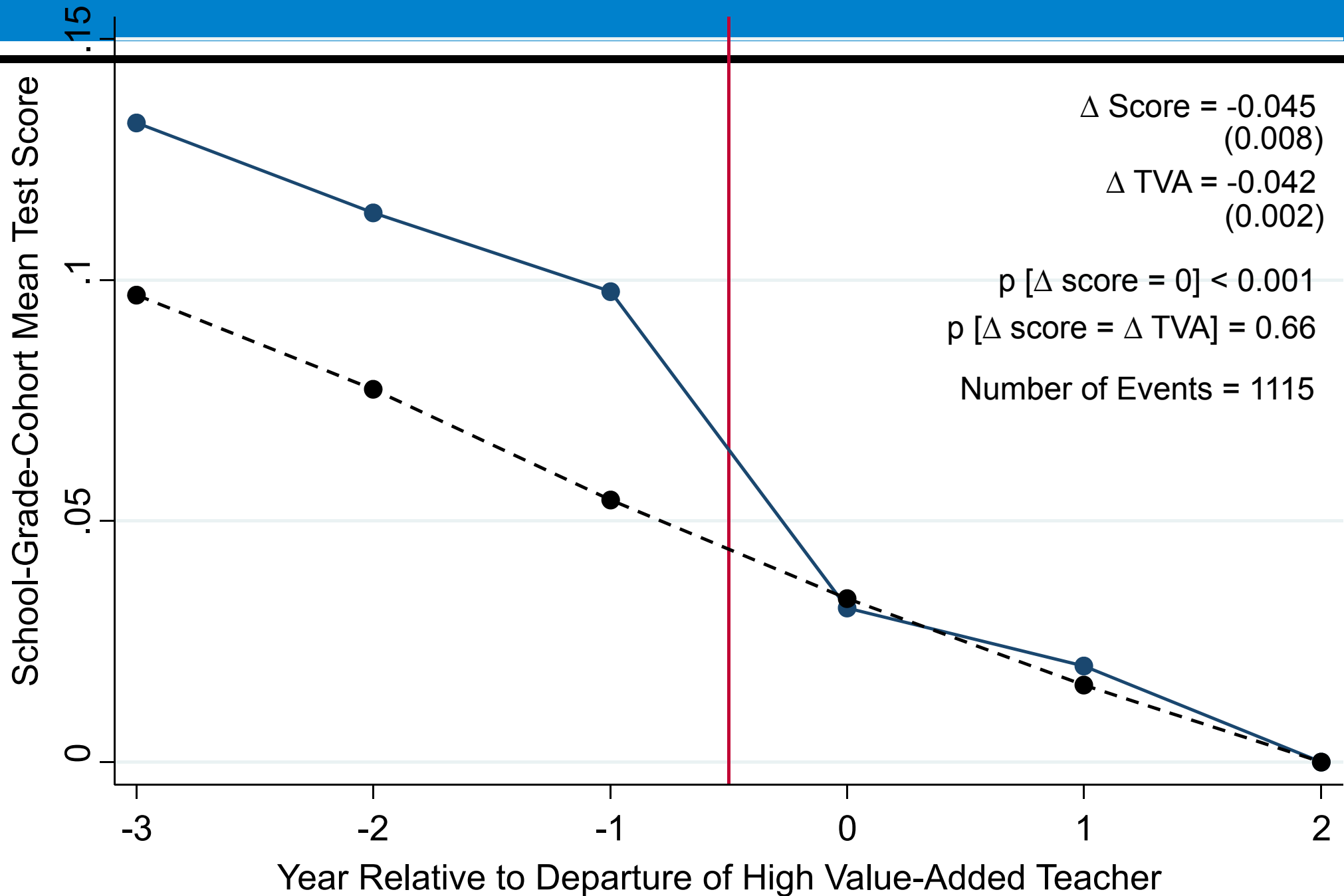


Non-traditional Predictors

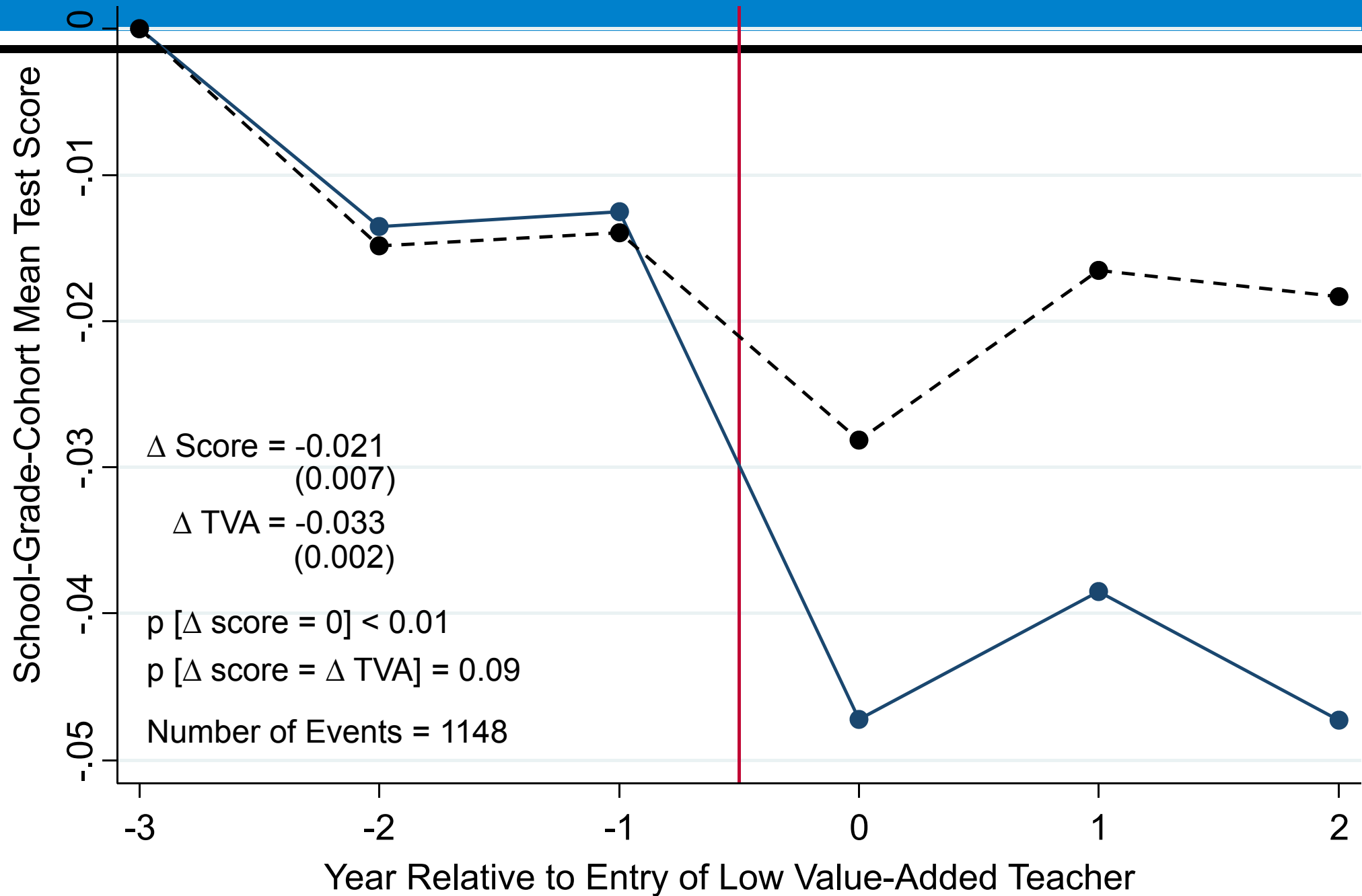
- Recall the man who lost his keys in the dark and looks for them in the lamplight...
 - Traditional credentials do a poor job predicting growth, but may just be the most easily available measures
- What about a systematic look at non-traditional characteristics as predictors of effectiveness?
 - Rockoff, Jacob, Kane and Staiger (2011), survey new math teachers in NYC (G4-8) on a variety of measures
 - Personality, content knowledge, cognitive ability, self-efficacy, ...commercial selection instruments, etc.
 - Again: no silver bullets (but moderate power to predict effectiveness when we pool measures into an index)



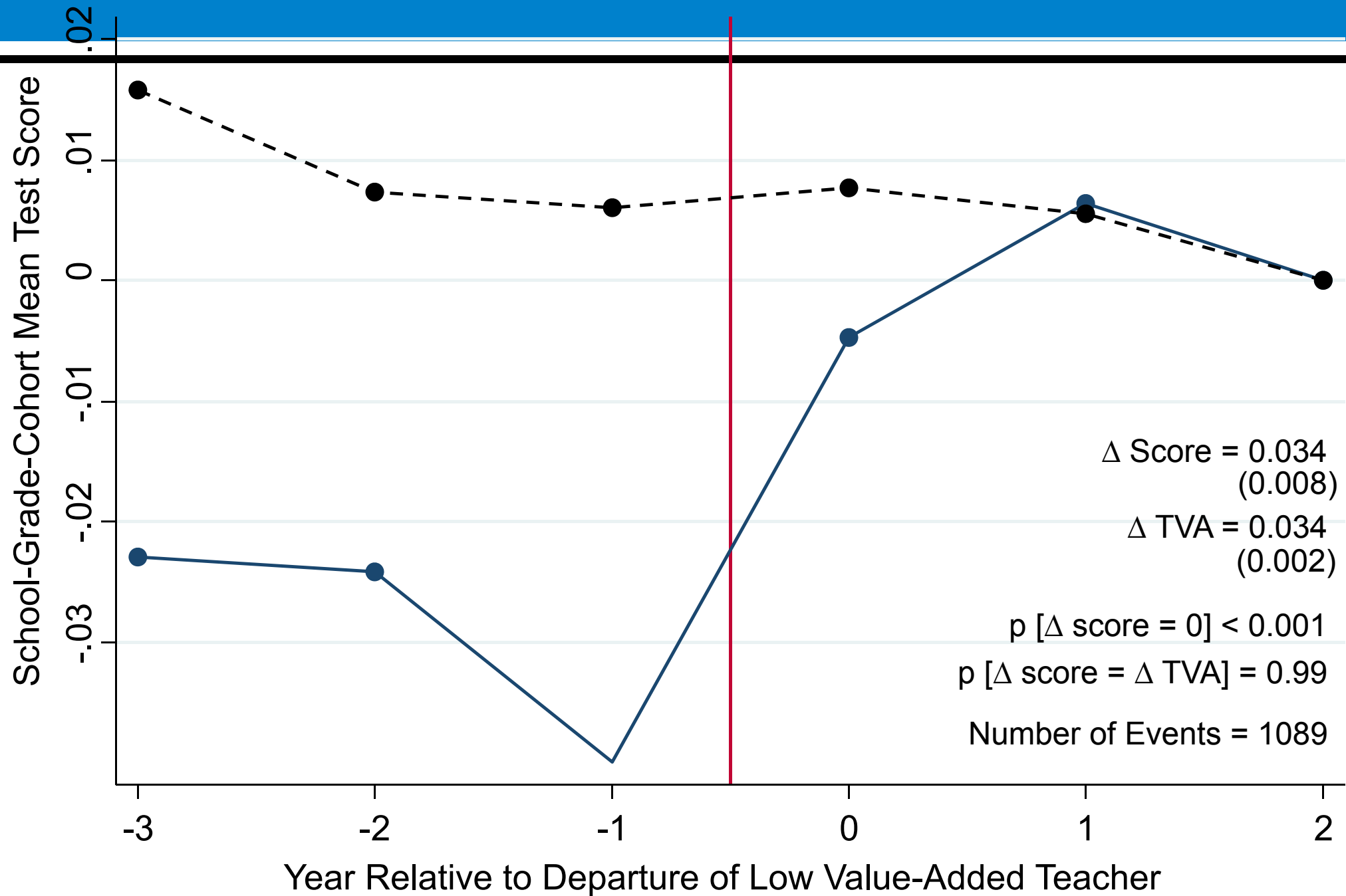
Impact of High Value-Added Teacher Exit on Cohort Test Scores



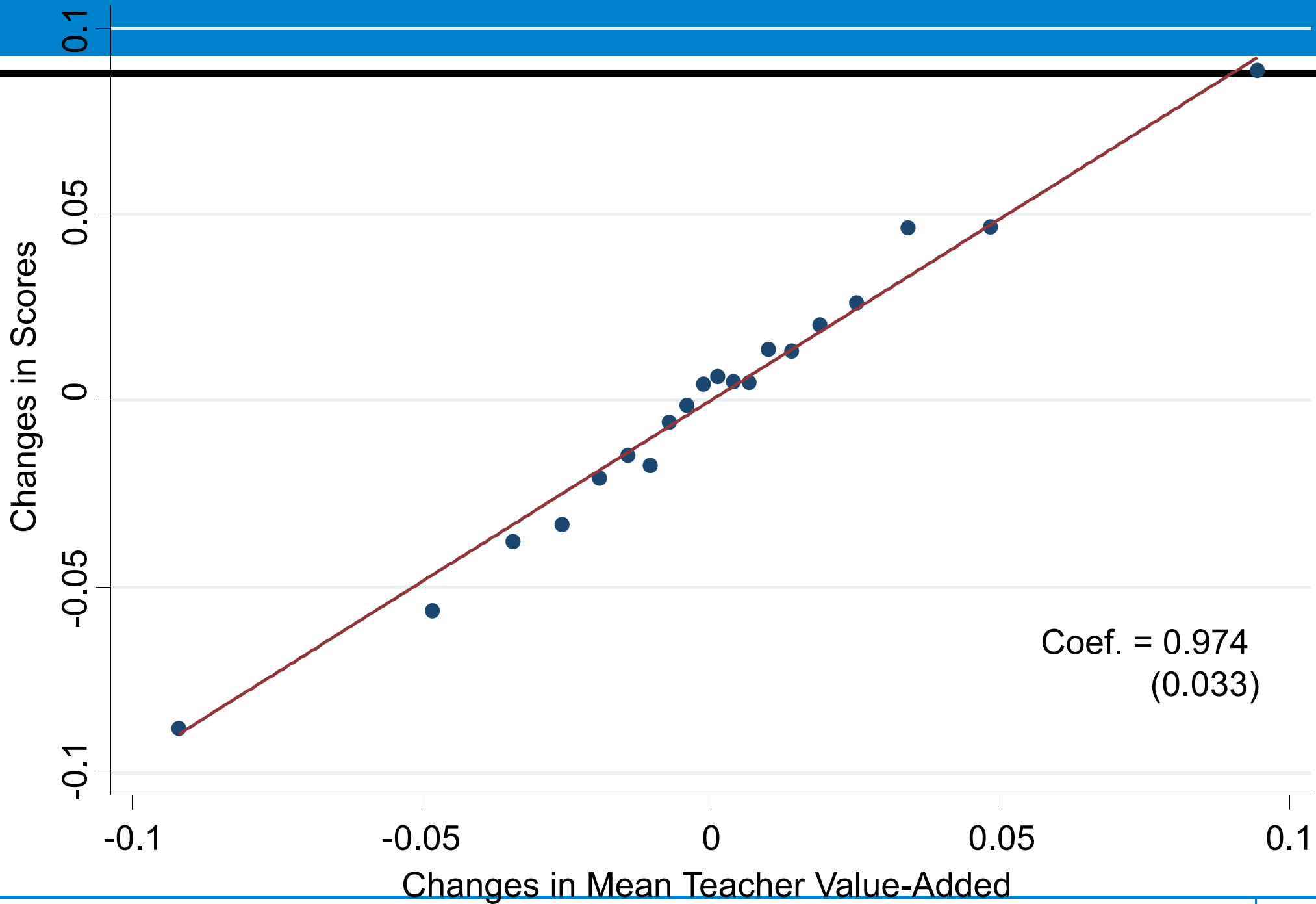
Impact of Low Value-Added Teacher Entry on Cohort Test Scores



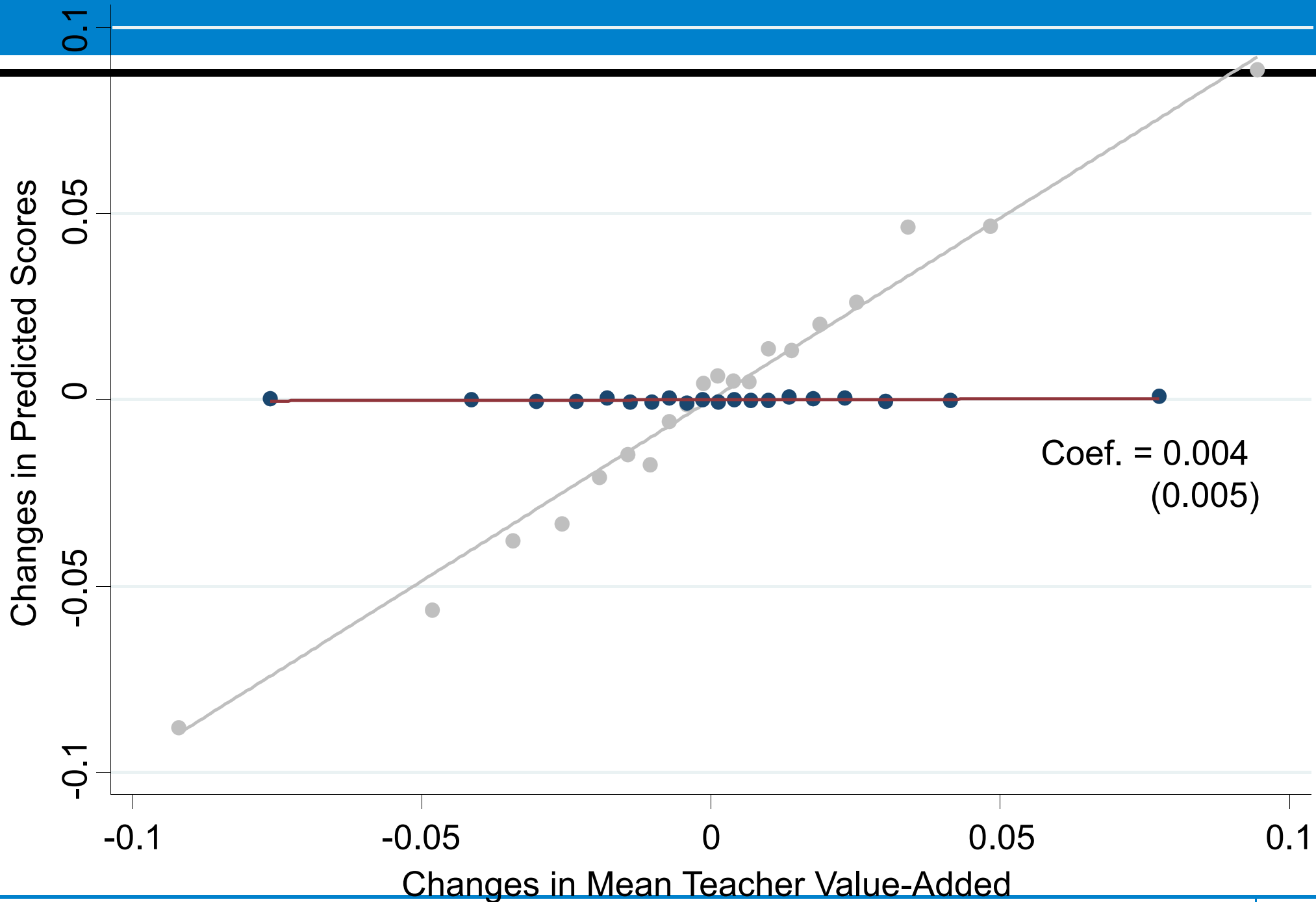
Impact of Low Value-Added Teacher Exit on Cohort Test Scores



Teacher Switchers Design: Changes in Scores vs. Changes in Mean Teacher VA



Changes in Predicted Scores vs. Changes in Mean Teacher VA



Changes in Other-Subject Scores vs. Changes in Mean Teacher VA Middle Schools Only

