NBER WORKING PAPER SERIES

THE LONG-TERM IMPACTS OF TEACHERS: TEACHER VALUE-ADDED AND STUDENT OUTCOMES IN ADULTHOOD

Raj Chetty
John N. Friedman
Jonah E. Rockoff

Working Paper 17699
http://www.nber.org/papers/w17699

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2011

We thank Joseph Altonji, Josh Angrist, David Card, Gary Chamberlain, David Deming, Caroline Hoxby, Guido Imbens, Brian Jacob, Thomas Kane, Lawrence Katz, Adam Looney, Phil Oreopoulos, Jesse Rothstein, Douglas Staiger, Danny Yagan, and seminar participants at the NBER Summer Institute, Stanford, Princeton, Harvard, Univ. of Chicago, Univ. of Pennsylvania, Brookings, Columbia, Univ. of Maryland, Pompeu Fabra, University College London, Univ. of British Columbia, and UC San Diego for helpful discussions and comments. This paper draws upon results from a paper in the IRS Statistics of Income Paper Series entitled "New Evidence on the Long- Term Impacts of Tax Credits on Earnings." Tax microdata were not accessed to write the present paper, as all results using tax data are based on tables contained in the SOI white paper. Peter Ganong, Sarah Griffis, Michal Kolesar, Jessica Laird, and Heather Sarsons provided outstanding research assistance. Financial support from the Lab for Economic Applications and Policy at Harvard and the National Science Foundation is gratefully acknowledged. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research. Publicly available portions of the analysis code are posted at: http://obs.rc.fas.harvard.edu/chetty/va_bias_code.zip

The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood
Raj Chetty, John N. Friedman, and Jonah E. Rockoff
NBER Working Paper No. 17699
December 2011, Revised January 2012
JEL No. I2,J24

ABSTRACT

Are teachers' impacts on students' test scores ("value-added") a good measure of their quality? This question has sparked debate largely because of disagreement about (1) whether value-added (VA) provides unbiased estimates of teachers' impacts on student achievement and (2) whether high-VA teachers improve students' long-term outcomes. We address these two issues by analyzing school district data from grades 3-8 for 2.5 million children linked to tax records on parent characteristics and adult outcomes. We find no evidence of bias in VA estimates using previously unobserved parent characteristics and a quasi-experimental research design based on changes in teaching staff. Students assigned to high-VA teachers are more likely to attend college, attend higher- ranked colleges, earn higher salaries, live in higher SES neighborhoods, and save more for retirement. They are also less likely to have children as teenagers. Teachers have large impacts in all grades from 4 to 8. On average, a one standard deviation improvement in teacher VA in a single grade raises earnings by about 1% at age 28. Replacing a teacher whose VA is in the bottom 5% with an average teacher would increase the present value of students' lifetime income by more than $250,000 for the average class- room in our sample. We conclude that good teachers create substantial economic value and that test score impacts are helpful in identifying such teachers.

Raj Chetty
Department of Economics
Harvard University
1805 Cambridge St.
Cambridge, MA 02138
and NBER
chetty@fas.harvard.edu

Jonah E. Rockoff
Columbia University
Graduate School of Business
3022 Broadway #603
New York, NY 10027-6903
and NBER
jonah.rockoff@columbia.edu

John N. Friedman
Harvard Kennedy School
Taubman 356
79 JFK St.
Cambridge, MA 02138
and NBER
john_friedman@harvard.edu

1 Introduction

Many policy makers advocate increasing the quality of teaching, but there is considerable debate about the best way to measure and improve teacher quality. One prominent method is to evaluate teachers based on their impacts on their students' test scores, commonly termed the "value-added" (VA) approach (Hanushek 1971, Murnane 1975, Rockoff 2004, Rivkin, Hanushek, and Kain 2005, Aaronson, Barrow, and Sander 2007, Kane and Staiger 2008). School districts from Washington D.C. to Los Angeles have begun to publicize VA measures and use them to evaluate teachers. Advocates argue that selecting teachers on the basis of their VA can generate substantial gains in achievement (e.g., Gordon, Kane, and Staiger 2006, Hanushek 2009), while critics contend that VA measures are poor proxies for teacher quality and should play little if any role in evaluating teachers (e.g., Baker et al. 2010, Corcoran 2010).

The debate about teacher VA stems primarily from two unanswered questions.[1] First, do the differences in test-score gains across teachers measured by VA capture causal impacts of teachers or are they driven primarily by student sorting? If students are sorted to teachers in ways that are not accounted for when estimating value-added, VA estimates will incorrectly reward or penalize teachers for the mix of students they get. Researchers have reached conflicting conclusions about the degree of bias in VA (e.g. Kane and Staiger 2008, Rothstein 2010) and there is still disagreement about this important issue. Second, do teachers who raise test scores improve their students' outcomes in adulthood or are they simply better at teaching to the test? Recent work has shown that early childhood education has significant long-term impacts (e.g. Heckman et al. 2010a, 2010b, 2010c, Chetty et al. 2011), but no study has identified the long-term impacts of teacher quality as measured by value-added.

We address these two questions using information from two administrative databases. The first is a dataset on test scores and classroom and teacher assignments in grades 3-8 from a large urban school district in the U.S. These data cover more than 2.5 million students and 18 million tests for math and English (reading) spanning 1989-2009. The second is selected data from United States tax records spanning 1996-2010.[2] These data contain information on student outcomes such as earnings, college attendance, and teenage births as well as parent characteristics such as

[1] There are also other important concerns about VA besides the two we focus on in this paper. For instance, as with other measures of labor productivity, the signal in value-added measures may be degraded by behavioral responses if high-stakes incentives are put in place (Barlevy and Neal 2012).

[2] Tax microdata were not directly used to write the present paper, as all results using tax data are drawn from tables contained in a Statistics of Income paper on the long-term impacts of tax policy (Chetty, Friedman, and Rockoff 2011). We describe the details of how the tax data were analyzed here as a reference.

household income, retirement savings, and mother's age at child's birth. We match nearly 90% of the observations in the school district data to the tax data, allowing us to track a large group of individuals from elementary school to early adulthood.

Our analysis has two parts. In the first part, we develop new tests for bias in VA measures. We estimate teacher value-added using standard Empirical Bayes methods, conditioning on pre-determined variables from the school district data such as lagged test scores (Kane and Staiger 2008, Kane, Rockoff, and Staiger 2008). Our estimates of VA are consistent with prior work: a 1 standard deviation (SD) improvement in teacher VA raises end-of-grade test scores by approximately 0.1 SD on average. To evaluate whether these VA estimates are biased by sorting on observables, we use parent characteristics from the tax data, which are strong predictors of test scores but are omitted from the VA models. We find that these parent characteristics are uncorrelated with teacher value-added *conditional* on the observables used to fit the VA model from the school district data. In addition, lagged test score gains are essentially uncorrelated with current teacher VA conditional on observables. We conclude that sorting on observable dimensions generates little or no bias in standard VA estimates.

To evaluate sorting on unobservables, we develop a quasi-experimental method of testing for bias in VA estimates that exploits changes in teaching assignments at the school-grade level. For example, suppose a high-VA 4th grade teacher moves from school s to another school in 1995. If VA estimates have predictive content, then students entering grade 4 in school s in 1995 should have lower quality teachers on average and their test score gains should be lower on average than the previous cohort. In practice, we find sharp breaks in test score gains around such teacher arrivals and departures at the school-grade-cohort level. Building on this idea, we assess the degree of bias in VA estimates by testing if observed changes in average test scores across cohorts match predictions based on the changes in the mean value-added of the teaching staff.[3] We find that the predicted impacts closely match observed impacts: the point estimate of the bias in forecasted impacts is 2% and statistically insignificant.[4] Although it rests on stronger identifying assumptions than a randomized experiment, our approach of using variation from teacher turnover

[3] This research design is related to recent studies of teacher turnover (e.g., Rivkin, Hanushek, and Kain 2005, Jackson and Bruegmann 2010, Ronfeldt et al. 2011), but is the first direct test of whether the VA of teachers who enter or exit affects mean test scores across cohorts. We discuss how our approach differs from this earlier work in Section 4.4.

[4] This quasi-experimental test relies on the assumption that teacher departures and arrivals are not correlated at a high frequency with student characteristics. We find no evidence of such correlations based on observables such as lagged test scores or scores in other subjects. This is intuitive, as parents are unlikely to immediately switch their children to a different school simply because a single teacher leaves or arrives.

can be implemented in many datasets and yields much more precise estimates of the degree of bias. Our method requires no data other than school district administrative records, and thus provides a simple technique for school districts and education researchers to validate their own value-added models.[5]

As we discuss in greater detail below, our results reconcile the findings of Kane and Staiger (2008) and Rothstein (2010) on bias in VA estimates. Rothstein finds minimal bias in VA estimates due to selection on observables but warns that selection on unobservables could *potentially* be a problem because students are sorted to classrooms based on lagged gains. Like Rothstein, we find minimal selection on observables. We then directly test for selection on unobservables using an approach analogous to Kane and Staiger (2008), but exploiting quasi-experimental variation in lieu of a randomized experiment. Like Kane and Staiger, we find no evidence of selection on unobservables. We therefore conclude that our value-added measures provide unbiased estimates of teachers' causal impacts on test scores despite the grouping of students on lagged gains documented by Rothstein.[6]

In the second part of the paper, we analyze whether high-VA teachers improve their students' outcomes in adulthood. We structure our analysis using a stylized dynamic model of the education production function in which cumulative teacher inputs over all grades affect earnings, as in Todd and Wolpin (2003). We regress outcomes such as earnings for a given set of students on teacher VA estimated using *other* cohorts to account for correlated errors in scores and earnings, as in Jacob, Lefgren, and Sims (2010). The resulting coefficients capture the "reduced form" impact of being assigned a teacher with higher VA in grade g, which includes both the grade g teacher's direct effect and any indirect benefits of being tracked to better teachers or receiving better educational inputs after grade g.

We first pool all grades to estimate the average reduced-form impact of having a better teacher for a single year from grades 4-8. We find that teacher VA has substantial impacts on a broad range of outcomes. A 1 SD improvement in teacher VA in a single grade raises the probability of college attendance at age 20 by 0.5 percentage points, relative to a sample mean of 36%. Improvements in teacher quality also raise the quality of the colleges that students attend, as measured by the average earnings of previous graduates of that college. Changes in the quality of the teaching

[5] STATA code to implement this technique is available at http://obs.rc.fas.harvard.edu/chetty/va_bias_code.zip

[6] Our findings do not contradict Rothstein's results; in fact, we replicate them in our own data. However, while Rothstein concludes that selection on unobservables could potentially generate significant bias, we find that it is actually negligible based on quasi-experimental tests that provide more definitive estimates of the degree of bias.

staff across cohorts generate impacts on college attendance and quality of a similar magnitude, supporting the view that these estimates reflect the causal impact of teachers.

Students who get higher VA teachers have steeper earnings trajectories, with significantly higher earnings growth rates in their 20s. At age 28, the oldest age at which we have a sufficiently large sample size to estimate earnings impacts, a 1 SD increase in teacher quality in a single grade raises annual earnings by about 1% on average. If this impact on earnings remains constant over the lifecycle, students would gain approximately $25,000 on average in cumulative lifetime income from a 1 SD improvement in teacher VA in a single grade; discounting at a 5% rate yields a present value gain of $4,600 at age 12, the mean age at which the interventions we study occur.

We also find that improvements in teacher quality significantly reduce the probability of having a child while being a teenager, increase the quality of the neighborhood in which the student lives (as measured by the percentage of college graduates in that ZIP code) in adulthood, and raise 401(k) retirement savings rates. The impacts on adult outcomes are all highly statistically significant, with the null of no impact rejected with $p < 0.01$.

Under certain strong assumptions about the nature of the tracking process, the net impacts of teacher VA in grade g can be recovered from the reduced-form coefficients by estimating a set of tracking equations that determine how teacher VA in grade g affects VA in subsequent grades. Using this approach, we find that the net impacts of teacher VA are significant and large throughout grades 4-8, showing that improvements in the quality of education can have large returns well beyond early childhood.[7]

The impacts of teacher VA are slightly larger for females than males. A given increase in test scores due to higher teacher quality is worth more in English than math, but the standard deviation of teacher effects is 50% larger in math than English. The impacts of teacher VA are roughly constant in percentage terms by parents' income. Hence, high income households, whose children have higher earnings on average, should be willing to pay larger absolute amounts for higher teacher VA.

The finding that one's teachers in childhood have long-lasting impacts may be surprising given evidence that teachers' impacts on test scores "fade out" very rapidly in subsequent grades (Rothstein 2010, Carrell and West 2010, Jacob, Lefgren, and Sims 2010). We confirm this rapid fade-out in our data, but find that test score impacts stabilize at about 1/3 the original impact after 3

[7] Because we can only analyze the impacts of teacher quality from grades 4-8, we cannot quantify the returns to education at earlier ages. The returns to better education in pre-school or earlier may be much larger than those estimated here (Heckman 2000).

years, showing that some of the achievement gains persist. Despite the fade-out of impacts on scores, the impacts of better teaching on earnings are similar to what one would predict based on the cross-sectional correlation between earnings and contemporaneous test score gains conditional on observables. This pattern of fade-out and re-emergence echoes the findings of recent studies of early childhood interventions (Heckman et al. 2010c, Deming 2009, Chetty et al. 2011).

To illustrate the magnitude of teachers' impacts, we use our estimates to evaluate the gains from selecting teachers based on their estimated VA. We begin by evaluating Hanushek's (2009) proposal to deselect the bottom 5% of teachers based on their value-added. We estimate that replacing a teacher whose true VA is in the bottom 5 percent with an average teacher would increase the present value of students' lifetime income by $267,000 per classroom taught.[8] However, because VA is estimated with noise, the gains from deselecting teachers based on a limited number of classrooms are smaller. We estimate the present value gains from deselecting the bottom 5% of teachers to be approximately $135,000 based on one year of data and $190,000 based on three years of data.

We then evaluate the expected gains from policies that pay bonuses to high-VA teachers in order to increase retention rates. The gains from such policies appear to be only modestly larger than their costs. Although the present value benefit from retaining a teacher whose estimated VA is at the 95th percentile after three years is nearly $200,000 per year, most bonus payments end up going to high-VA teachers who would have stayed even without the additional payment (Clotfelter et al. 2008). Replacing low VA teachers may therefore be a more cost effective strategy to increase teacher quality in the short run than paying bonuses to retain high-VA teachers. In the long run, higher salaries could attract more high VA teachers to the teaching profession, a potentially important benefit that we do not measure here.[9]

It is important to keep two caveats in mind when evaluating the policy implications of our findings. First, teachers were not incentivized based on test scores in the school district and time period we study. The signal content of value-added might be lower when it is used to evaluate teachers because of behavioral responses such as cheating or teaching to the test (Jacob and Levitt 2003, Jacob 2005, Neal and Schanzenbach 2010). Our results quantify the gains from higher VA teachers in an environment without such distortions in teacher behavior.[10] Further work is

[8] This calculation discounts the earnings gains at a rate of 5% to age 12. The total undiscounted earnings gains from this policy are $52,000 per child and more than $1.4 milllion for the average classroom.

[9] Increasing salaries or paying bonuses based on VA could also result in gains to students via changes in teacher effort in the short run. However, a recent experimental study from the U.S. found no significant impacts of this type of incentive program (Springer et al. 2010).

[10] Even in our sample, we find that the top 2% of teachers ranked by VA have patterns of test score gains that are consistent with test manipulation based on the proxy developed by Jacob and Levitt (2003). Correspondingly, these

needed to determine how VA should be used for education policy in a high stakes environment with multitasking and imperfect monitoring (Holmstrom and Milgrom 1991, Barlevy and Neal 2012).

Second, our analysis does not compare value-added with other measures of teacher quality. It is quite plausible that aspects of teacher quality which are not captured by standardized tests have significant long-term impacts. This raises the possibility that other measures of teacher quality (e.g., evaluations based on classroom observation) might be even better predictors of teachers' long-term impacts than value-added scores, though the signal content of these measures in a high stakes environment could also be degraded by behavioral distortions. Further work comparing the long-term impacts of teachers rated on various metrics is needed to determine the optimal method of teacher evaluation. What is clear from this study is that improving teacher quality is likely to yield substantial returns for students; the best way to accomplish that goal is less clear.

The paper is organized as follows. In Section 2, we present a statistical model to formalize the questions we seek to answer and derive estimating equations for our empirical analysis. Section 3 describes the data sources and provides summary statistics as well as cross-sectional correlations between scores and adult outcomes as a benchmark. Section 4 discusses the results of our tests for bias in VA measures. Results on teachers' long-term impacts are given in section 5. Section 6 presents policy calculations and Section 7 concludes.

2 Conceptual Framework

We structure our analysis using a stylized dynamic model of the education production function based on previous work (Todd and Wolpin 2003, Cunha and Heckman 2010, Cunha, Heckman, and Schennach 2010). The purpose of the model is to formalize the identification assumptions underlying our empirical analysis and clarify how the reduced-form parameters we estimate should be interpreted. We therefore focus exclusively on the role of teachers, abstracting from other inputs to the education production function, such as peers or parental investment. Using this model, we (1) define a set of reduced-form treatment effects, (2) present the assumptions under which we can identify these treatment effects, and (3) derive estimating equations for these parameters.

2.1 Structural Model of Student Outcomes

Our model is characterized by three relationships: a specification for test scores, a specification for earnings (or other adult outcomes), and a rule that governs student and teacher assignment to

high VA outlier teachers also have much smaller long-term impacts than one would predict based on their VA.

classrooms. School principals first assign student i in grade g to a classroom $c(i,g)$ based on lagged test scores, prior inputs, and other unobserved determinants of student achievement. Principals then assign a teacher j to each classroom c based on classroom characteristics such as mean lagged scores and class demographics. Let $j(i,g) = j(c(i,g))$ denote student i's teacher in grade g. Let e_j denote teacher j's years of teaching experience.

Student i's test score in grade g, A_{ig}, is a function of current and prior inputs:

$$A_{ig} = \sum_{s=1}^{g} \sigma_{sg} \mu_{j(i,s)} + \lambda_{c(i,g)} + \eta_i + \zeta_{ig} \tag{1}$$

where $\mu_{j(i,g)}$ represents the impact of teacher j on test scores, which we term the teacher's "value-added." We scale teacher quality so that the average teacher has quality $\mu_j = 0$ and the effect of teacher quality in grade g on scores in grade g is $\sigma_{gg} = 1$. For $s < g$, σ_{sg} measures the persistent impact of teacher quality μ in grade s on test scores at the end of grade g. $\lambda_{c(i,g)}$ represents an exogenous transitory classroom-level shock, η_i represents academic ability, and ζ_{ig} represents idiosyncratic noise and other period-specific innovations in individual achievement.

The model for scores in (1) makes two substantive restrictions that are standard in the value-added literature. First, it assumes that teacher quality μ_j is fixed over time, except for the effects of teacher experience, which we model in our empirical specifications. This rules out the possibility that teacher quality fluctuates across years (independent of experience) or that it depends upon the characteristics of the students assigned to the teacher (e.g., high vs low achieving students).[11] Second, our model does not explicitly account for endogenous responses of other inputs such as parental effort in response to changes in teacher quality. We discuss the consequences of these assumptions for our results below.

Earnings Y_i are a function of the inputs over all G grades:

$$Y_i = \sum_{g=1}^{G} \gamma_g \tau^Y_{j(i,g)} + \eta^Y_i \tag{2}$$

where $\tau^Y_{j(i,g)}$ represents teacher j's impact on earnings, γ_g measures the effect of teacher quality in grade g on earnings and η^Y_i reflects individual heterogeneity in earnings ability, which may be correlated with academic ability η_i. This specification assumes that the transitory classroom and individual-level shocks that affect scores have no impact on earnings, a simplification that has no effect on the results below.

[11] One could reinterpret λ in equation 1 as a class-specific component of teacher quality. In that case, the methods we implement below would estimate the component of teacher quality that is constant across years.

2.2 Identifying Teachers' Impacts on Scores

Our first goal is to identify the causal impacts of changing the teacher of class c from teacher j to j' in grade g on test scores and earnings. Define the potential outcome $A_{ig}(j')$ as the test score student i would have in grade g if his teacher were $j(i,g) = j'$. With the normalization $\sigma_{gg} = 1$, the causal effect of replacing teacher j with j' on student i's end-of-year score is simply $A_{ig}(j') - A_{ig}(j) = \mu_{j'} - \mu_j$. In our stylized model, the treatment effect $A_{ig}(j') - A_{ig}(j)$ coincides with the structural impact of teachers on scores. In a more general model with endogenous parent inputs and peer quality, this reduced-form treatment effect combines various structural parameters. For instance, students assigned to a better teacher may get less help on their homework from parents. Though it is not a policy-invariant primitive parameter, the reduced-form parameter μ_j is of direct relevance to certain questions, such as the impacts of retaining teachers on the basis of their VA (Todd and Wolpin 2003).

To estimate μ_j, we begin by estimating the following empirical model for student i's test score in grade g in school year t:

$$
\begin{aligned}
(3) \qquad A_{igt} &= f_{1g}(A_{i,t-1}) + f_2(e_{j(i,g,t)}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g,t)} + \nu_{igt} \\
\text{where } \nu_{igt} &= \mu_{j(i,g,t)} + \theta_{c(i,g,t)} + \varepsilon_{igt}
\end{aligned}
$$

Here $f_{1g}(A_{i,t-1})$ is a control function for individual test scores in year $t-1$, $f_2(e_{j(i,g,t)})$ controls for the impacts of teacher experience, X_{igt} is a vector of student characteristics (such as whether the student is a native English speaker), and $\bar{X}_{c(i,g,t)}$ is a vector of classroom-level characteristics determined before teacher assignment (such as class size or an indicator for being an honors class). We decompose the error term in the empirical model into three components: teacher quality (μ_j), class shocks ($\theta_{c(i,g,t)}$), and idiosyncratic shocks (ε_{igt}). We can distinguish teacher effects μ_j from class shocks $\theta_{c(i,g,t)}$ by observing teachers over many school years.[12] Note that because we control for the effects of teacher experience in (3), μ_j represents the variation in teacher quality that is independent of experience.[13]

The empirical model for test scores in (3) differs from the structural model in (1) because we cannot observe all the terms in (1), such as heterogeneity in individual ability (η_i and ζ_{ig}). Value-

[12] This is the key distinction between our paper and Chetty et al.'s (2011) analysis of the long term impacts of Project STAR using tax data. Chetty et al. observe each teacher in only one classroom and therefore cannot separate teacher and class effects.

[13] To simplify notation, we assume that teachers teach one class per year (as in elementary schools). Because the j and c subscripts become redundant, we drop the c subscript. When teachers are assigned more than one class per year, we treat each class as if it were in a separate year for the purposes of the derivation below.

added models address this problem by controlling for prior-year test scores, which in principle should capture much of the variance in ability because η_i is a component of previous test scores. With these controls, the idiosyncratic error term in the empirical model ε_{igt} reflects unobserved student-level heterogeneity in test scores arising from the components of the structural model in (1) that are orthogonal to lagged scores and other observable characteristics. The class-level error term $\theta_{c(i,g,t)}$ reflects analogous unobserved class-level heterogeneity.

There are various methods one could use to estimate μ_j and the other error components in (3), such as estimating a correlated random effects model, a hierarchical linear model, or implementing an Empirical Bayes procedure. All of these methods rely on the following identification assumption to obtain consistent estimates of μ_j.

Assumption 1 Students are not sorted to teachers on unobservable determinants of test scores:

$$\mathbb{E}\left[\theta_{c(i,g,t)} + \varepsilon_{igt} | j\right] = \mathbb{E}\left[\theta_{c(i,g,t)} + \varepsilon_{igt}\right]$$

Assumption 1 requires that each teacher is no more likely than other teachers to be assigned students who score highly, conditional on the controls in the empirical model (3). If this assumption fails, the estimated teacher effects $\hat{\mu}_j$ will pick up differences in unobserved student characteristics across teachers and not the causal impacts of the teachers themselves. Note that Assumption 1 is not inconsistent with some parents sorting their children to particular teachers. Assumption 1 only requires that the observable characteristics $\left\{A_{i,t-1}, X_{igt}, \bar{X}_{c(i,g,t)}\right\}$ are sufficiently rich so that any remaining unobserved heterogeneity in test scores is balanced across teachers.[14] The first half of our empirical analysis focuses on assessing whether this is the case using two tests that we describe in Section 4.

Empirical Implementation. We estimate μ_j using an Empirical Bayes procedure following Morris (1983) and Kane and Staiger (2008, pp 14-16), which is the most commonly used approach to estimate VA (McCaffrey et al. 2003). We use this approach because of its computational simplicity and because our primary goal is to evaluate the properties of existing VA measures rather than devise new measures. Our procedure for estimating μ_j consists of three steps, which we implement separately for math and English observations:

Step 1: Calculate residual test score gains. We estimate (3) using OLS and compute residuals of student test scores, $\hat{\nu}_{igt}$. We then estimate the variances of the error components σ^2_μ, σ^2_θ, and

[14] For example, suppose motivated parents are able to get their children better teachers. These children would presumably also have had higher test scores in the previous grade. Hence, conditional on prior test scores, the remaining variation in current test scores could be balanced across teachers despite unconditional sorting.

σ^2_ε using equations (2)-(4) in Kane and Staiger (2008). Intuitively, the within-classroom variance identifies σ^2_ε, the within-teacher cross-classroom covariance identifies σ^2_μ, and the remaining variance is due to σ^2_θ.

Step 2: Calculate average teacher effects. Let $\overline{\nu}_{jt}$ denote the mean score residual for the classroom taught by teacher j in year t and n_{jt} the number of students in that class. We estimate each teacher's quality using a precision-weighted average of $\overline{\nu}_{jt}$ across the classes taught by teacher j:

$$\bar{\nu}_j = \sum_t h_{jt}\bar{\nu}_{jt} / \sum_t h_{jt}$$

where $h_{jt} = 1/(\hat{\sigma}^2_\theta + \hat{\sigma}^2_\varepsilon / n_{jt})$ denotes is the inverse of the variance of the estimate of teacher quality obtained from class t.

Step 3: Shrink teacher effect estimates. Finally, we shrink the mean test score impact $\overline{\nu}_j$ toward the sample mean (0) to obtain an estimate of the teacher's quality:

$$\hat{\mu}_j = \overline{\nu}_j \frac{\hat{\sigma}^2_\mu}{\hat{\sigma}^2_\mu + 1/\sum_t h_{jt}} = \overline{\nu}_j \cdot r \tag{4}$$

where $r \equiv \frac{Var(\mu_j)}{Var(\overline{\nu}_j)}$ is commonly termed the "reliability" of the VA estimate.

To understand the purpose of the shrinkage correction, consider an experiment in which we estimate teacher impacts $\overline{\nu}_j$ in year t and then randomly assign students to teachers in year $t+1$. The best (mean-squared error minimizing) linear predictor of student's test scores $A_{ig,t+1}$ based on $\bar{\nu}_j$ is obtained from the OLS regression $A_{ig,t+1} = a + b\bar{\nu}_j$. The coefficients in this regression are $a = 0$ and $b = \frac{cov(A_{ig,t+1}, \bar{\nu}_j)}{var(\overline{\nu}_j)} = \frac{Var(\mu_j)}{Var(\overline{\nu}_j)} = r$, implying that the optimal forecast of teacher j's impact on future scores is $\hat{\mu}_j = \overline{\nu}_j \cdot r$. From a frequentist perspective, the measurement error in $\overline{\nu}_j$ makes it optimal to used a biased but more precise estimate of teacher quality to minimize the mean-squared error of the forecast. From a Bayesian perspective, the posterior mean of the distribution of μ_j with Normal errors is a precision-weighted average of the sample mean ($\overline{\nu}_j$) and the mean of the prior (0), which is $\mathbb{E}\mu_j|\overline{\nu}_j = \overline{\nu}_j \cdot r = \hat{\mu}_j$. Because of these reasons, we follow the literature and use $\hat{\mu}_j$ as our primary measure of teacher quality in our empirical analysis. As a robustness check, we replicate our main results using mean test score residuals ($\bar{\nu}_j$) and show that, as expected, the estimated impacts are attenuated by roughly the mean of the shrinkage factor r.

2.3 Identifying Teachers' Impacts on Earnings

The impact of changing the teacher of class c from j to j' in grade g on mean earnings is:

$$\mu_j^Y - \mu_{j'}^Y = \mathbb{E}Y_i(j(i,g)) - \mathbb{E}Y_i(j'(i,g)) \tag{5}$$

$$= \gamma_g \left(\tau_{j'(i,g)}^Y - \tau_{j(i,g)}^Y \right) + \sum_{s=g+1}^{G} \gamma_s \left(\mathbb{E}\tau_{j(i,s)}^Y|_{j'(i,g)} - \mathbb{E}\tau_{j(i,s)}^Y|_{j(i,g)} \right). \tag{6}$$

Replacing teacher j affects earnings through two channels. The first term in (5) represents the direct impact of the change in teachers on earnings. The second term represents the indirect impact via changes in the expected quality of subsequent teachers to which the student is assigned. For example, a higher achieving student may be tracked into a more advanced sequence of classes taught by higher quality teachers. In a more general model, other determinants of earnings such as parental effort or peer quality might also respond endogenously to the change in teachers.

In principle, one could estimate teacher j's reduced-form causal impact on earnings, μ_j^Y, using an empirical model analogous to the one used above for test scores:

$$\begin{aligned} Y_i &= f_{1g}^Y(A_{i,t-1}) + f_2^Y(e_{j(i,g,t)}) + \phi_1^Y X_{igt} + \phi_2^Y \bar{X}_{c(i,g,t)} + \nu_{igt}^Y \\ \nu_{igt}^Y &= \mu_{j(i,g,t)}^Y + \theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y \end{aligned} \tag{7}$$

Teacher impacts on earnings μ_j^Y can be identified under an assumption about sorting analogous to Assumption 1:

$$\mathbb{E}\left[\theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y \mid j \right] = \mathbb{E}\left[\theta_{c(i,g,t)}^Y + \varepsilon_{igt}^Y \right] \tag{8}$$

This condition, although similar to Assumption 1, is a much stronger requirement in practice. Assumption 1 holds if ε_{igt} is balanced across teachers, which requires that η_i is orthogonal to A_{igt} conditional on lagged test scores and other observables. The condition in (8) holds if ε_{ijt}^Y is balanced across teachers, which requires η_i^Y to be orthogonal to Y_i conditional on lagged test scores and other observables. Because η_i appears directly in $A_{i,t-1}$, it is likely to be absorbed by controlling for lagged scores. In contrast, η_i^Y does *not* appear in lagged scores and hence is unlikely to be absorbed by these controls. If we observed an analog of lagged scores such as lagged expected earnings, we could effectively control for η_i^Y and more plausibly satisfy (8).

As a concrete example, suppose that students have heterogeneous levels of ability, which affects scores and earnings, and family connections, which only affect earnings. Students are sorted to

teachers on the basis of both of these characteristics. While ability is picked up by lagged test scores and thus eliminated from ε_{igt}, family connections are not absorbed by the controls and appear in ε^Y_{igt}. As a result, teachers' impacts on scores can be consistently estimated, but their impacts on earnings cannot because there is systematic variation across teachers in their students' earnings due purely to connections.

In practice, we are unable to account for η^Y_i fully: tests for sorting on pre-determined characteristics analogous to those in Section 4.1 reveal that (8) is violated in our data. Therefore, we cannot identify teachers' total impacts on earnings μ^Y_j despite being able to identify teachers' impacts on test scores. Given this constraint, we pursue a less ambitious objective: estimating the correlation between teachers' impacts on scores and earnings, $cov(\mu_j, \mu^Y_j)$. This yields a lower bound on teacher effects on earnings μ^Y_j, as the standard deviation of μ^Y_j is bounded below by $\beta_g \sigma_\mu$, which measures the portion of $var(\mu^Y_j)$ due to $cov(\mu_j, \mu^Y_j)$.

To see how we can identify $cov(\mu_j, \mu^Y_j)$, consider the following empirical model for earnings as a function of teacher VA for student i in grade g in year t:

$$Y_i = \beta_g \hat{\mu}_{j(i,g)} + f^\mu_{1g}(A_{i,t-1}) + f^\mu_2(e_{j(i,g,t)}) + \phi^\mu_1 X_{igt} + \phi^\mu_2 \bar{X}_{c(i,g,t)} + \varepsilon^\mu_{igt}. \tag{9}$$

The coefficient β_g in this equation represents the mean increase in student earnings from a one unit increase in teacher VA in grade g, as measured using the Empirical Bayes procedure described above. Estimating (9) using OLS yields an unbiased estimate of β_g under the following assumption.

Assumption 2 Teacher value-added is orthogonal to unobserved determinants of earnings:

$$cov\left(\hat{\mu}_j, \varepsilon^\mu_{igt}\right) = 0.$$

Assumption 2 is weaker than (8) because it only requires that there be no correlation between teacher value-added and unobservables.[15] In our example above, it allows students with better family connections η^Y_i to be systematically tracked to certain teachers as long as those teachers do not systematically have higher levels of value-added on test scores, conditional on the controls $\left\{A_{i,t-1}, e_{j(i,g,t)}, X_{igt}, \bar{X}_{c(i,g,t)}\right\}$. While this remains a strong assumption, it may hold in practice because teacher VA was not publicized during the period we study and VA is very difficult to predict based on teacher observables. We evaluate whether conditioning on observables is adequate to

[15] Assumption 2 would be violated if the same observations were used to estimate $\widehat{\mu}_j$ and β because the estimation errors in (3) and (9) are correlated. Students with unobservably high test scores η_i are also likely to have unobservably high earnings η^Y_i. We deal with this technical problem by using a leave-out mean to estimate $\widehat{\mu}_j$ as described in Section 4.

satisfy Assumption 2 using quasi-experimental techniques in Section 5.

The coefficient β_g in (9) represents the reduced-form impact of having a higher VA teacher in grade g and includes the impacts of subsequent endogenous treatments such as better teachers in later grades. While this reduced-form impact is of interest to parents, one may also be interested in identifying the impact of each teacher net of potential tracking to better teachers in later grades. Let $\tilde{\beta}_g$ denote the impact of teacher VA in grade g on earnings holding fixed teacher VA in subsequent grades. One intuitive specification to identify $\tilde{\beta}_g$ is to regress earnings on teacher VA in all grades simultaneously:

$$Y_i = \sum_{g=1}^{G} \tilde{\beta}_g \hat{\mu}_{j(i,g)} + \varepsilon_i^{\mu}. \tag{10}$$

Identifying $\left\{\tilde{\beta}_g\right\}$ in (10) requires the orthogonality condition $Cov\left(\hat{\mu}_{j(i,g)}, \varepsilon_i^{\mu}\right) = 0$. As we discussed above, this assumption does not hold unconditionally because students are assigned to teachers in grade g based on grade $g-1$ test scores $A_{i,g-1}$. Because we must condition on $A_{i,g-1}$ in order to obtain variation in grade g teacher VA $\hat{\mu}_{j(i,g)}$ that is orthogonal to student characteristics, we cannot directly estimate (10), as $A_{i,g-1}$ is endogenous to grade $g-1$ teacher VA $\hat{\mu}_{j(i,g-1)}$.[16] Instead, we develop a simple iterative method of recovering the net impacts $\tilde{\beta}_g$ from our reduced form estimates β_g and estimates of the degree of teacher tracking in Section 6.1.

3 Data

We draw information from two databases: administrative school district records and information on these students and their parents from U.S. tax records. We first describe the two data sources and then the structure of the linked analysis dataset. Finally, we provide descriptive statistics and cross-sectional correlations using the analysis dataset.

3.1 School District Data

We obtain information on students, including enrollment history, test scores, and teacher assignments from the administrative records of a large urban school district. These data span the school years 1988-1989 through 2008-2009 and cover roughly 2.5 million children in grades 3-8. For simplicity, we refer below to school years by the year in which the spring term occurs (e.g., the school

[16] For the same reason, we also cannot estimate the complementarity of teachers across grades. Estimating complementarity requires simultaneous quasi-random assignment of teachers in *both* grades g and $g-1$, but we are only able to isolate quasi-random variation one grade at a time with our research design.

year 1988-89 is 1989).

Test Scores. The data include approximately 18 million test scores. Test scores are available for English language arts and math for students in grades 3-8 in every year from the spring of 1989 to 2009, with the exception of 7th grade English scores in 2002.[17] In the early and mid 1990s, all tests were specific to the district. Starting at the end of the 1990s, the tests in grades 4 and 8 were administered as part of a statewide testing system, and all tests in grades 3-8 became statewide in 2006 as required under the No Child Left Behind law.[18] Because of this variation in testing regimes, we follow prior work on measuring teachers' effects on student achievement, taking the official scale scores from each exam and normalizing the mean to zero and the standard deviation to one by year and grade. The within-grade variation in achievement in the district we examine is comparable to the within-grade variation nationwide, so our results can easily be compared to estimates from other samples.[19]

Demographics. The dataset contains information on ethnicity, gender, age, receipt of special education services, and limited English proficiency for the school years 1989 through 2009. The database used to code special education services and limited English proficiency changed in 1999, creating a break in these series that we account for in our analysis by interacting these two measures with a post-1999 indicator. Information on free and reduced price lunch is available starting in school year 1999.

Teachers. The dataset links students in grades 3-8 to classrooms and teachers from 1991 through 2009.[20] This information is derived from a data management system which was phased in over the early 1990s, so not all schools are included in the first few years of our sample. In addition, data on course teachers for middle and junior high school students—who, unlike students in elementary schools, are assigned different teachers for math and English—are more limited.

[17] We also have data on math and English test scores in grade 2 from 1991-1994 and English test scores in grades 9-10 from 1991-1993, which we use only when estimating teachers' impacts on past and future test scores. Because these observations are a very small fraction of our analysis sample, excluding them has little impact on the placebo tests and fade-out estimates reported in Figure 2.

[18] All tests were administered in late April or May during the early-mid 1990s, and students were typically tested in all grades on the same day throughout the district. Statewide testing dates varied to a greater extent, and were sometimes given earlier in the school year (e.g., February) during the latter years of our data.

[19] The standard deviation of 4th and 8th grade English and math achievement in this district ranges from roughly 95 percent to 105 percent of the national standard deviation on the National Assessment of Educational Progress, based on data from 2003 and 2009, the earliest and most recent years for which NAEP data are available. Mean scores are significantly lower than the national average, as expected given the urban setting of the district.

[20] 5% of students switch classrooms or schools in the middle of a school year. We assign these students to the classrooms in which they took the test to obtain an analysis dataset with one observation per student-year-subject. However, when defining class and school-level means of student characteristics (such as fraction eligible for free lunch), we account for such switching by weighting students by the fraction of the year they spent in that class or school.

Course teacher data are unavailable prior to the school year 1994, then grow in coverage to roughly 60% by school year 1998 and 85% by 2003. Even in the most recent years of the data, roughly 15 percent of the district's students in grades 6 to 8 are not linked to math and English teachers because some middle and junior high schools still do not report course teacher data.

The missing teacher links raise two potential concerns. First, our estimates (especially for grades 6-8) apply to a subset of schools with more complete information reporting systems and thus may not be representative of the district as a whole. Reassuringly, we find that these schools do not differ significantly from the sample as a whole on test scores and other observables. Second, and more importantly, missing data could generate biased estimates. Almost all variation in missing data occurs at the school level because data availability is determined by whether the school utilizes in the district's centralized data management system for tracking course enrollment and teacher assignment. Specifications that exploit purely within-school comparisons are therefore essentially unaffected by missing data and we show that our results are robust to exploiting such variation. Moreover, we obtain similar results for the subset of years when we have complete data coverage in grades 3-5, confirming that missing data does not drive our results.

We obtain information on teacher experience from human resource records. The human resource records track teachers since they started working in the district and hence give us an uncensored measure of within-district experience for the teachers in our sample. However, we lack information on teaching experience outside of the school district.

Sample Restrictions. Starting from the raw dataset, we make a series of sample restrictions that parallel those in prior work to obtain our primary school district sample. First, because our estimates of teacher value-added always condition on prior test scores, we restrict our sample to grades 4-8, where prior test scores are available. Second, we drop the 2% of observations where the student is listed as receiving instruction at home, in a hospital, or in a school serving solely disabled students. We also exclude the 6% of observations in classrooms where more than 25 percent of students are receiving special education services, as these classrooms may be taught by multiple teachers or have other special teaching arrangements. Third, we drop classrooms with less than 10 students or more than 50 students as well as teachers linked with more than 200 students in a single grade, because such students are likely to be mis-linked to classrooms or teachers (0.5% of observations). Finally, when a teacher is linked to students in multiple schools during the same year, which occurs for 0.3% of observations, we use only the links for the school where the teacher is listed as working according to human resources records and set the teacher as missing in the other

schools. After these restrictions, we are left with 15.0 million student-year-subject observations. Of these, 9.1 million records have information on teacher and 7.7 million have information on both teachers and test score gains, which we need to estimate value-added.

3.2 Tax Data

In Chetty, Friedman, and Rockoff (2011), we obtain data on students' adult outcomes and their parents' characteristics from income tax returns. Here, we briefly summarize some key features of the variables used in the analysis below. The year always refers to the tax year (i.e., the calendar year in which the income is earned or the college expense incurred). In most cases, tax returns for tax year t are filed during the calendar year $t+1$. We express all monetary variables in 2010 dollars, adjusting for inflation using the Consumer Price Index.

Earnings. Individual earnings data come from W-2 forms, which are available from 1999-2010. W-2 data are available for *both* tax filers and non-filers, eliminating concerns about missing data. Individuals with no W-2 are coded as having 0 earnings.[21] We cap earnings in each year at $100,000 to reduce the influence of outliers; 1.2% of individuals in the sample report earnings above $100,000 at age 28.

College Attendance. We define college attendance as an indicator for having one or more 1098-T forms filed on one's behalf. Title IV institutions – all colleges and universities as well as vocational schools and other postsecondary institutions – are required to file 1098-T forms that report tuition payments or scholarships received for every student. Because the 1098-T forms are filed directly by colleges, missing data concerns are minimal.[22] Comparisons to other data sources indicate that 1098-T forms accurately capture US college enrollment.[23] We have no information about college completion or degree attainment because the data are based on tuition payments. The 1098-T data are available from 1999-2009.

College Quality. We construct an earnings-based index of college quality as in Chetty et al. (2011). Using the full population of all individuals in the United States aged 20 on 12/31/1999

[21] We obtain similar results using household adjusted gross income reported on individual tax returns. We focus on the W-2 measure because it provides a consistent definition of individual wage earnings for both filers and non-filers. One limitation of the W-2 measure is that it does not include self-employment income.

[22] Colleges are not required to file 1098-T forms for students whose qualified tuition and related expenses are waived or paid entirely with scholarships or grants; however, the forms are generally available even for such cases, perhaps because of automated reporting to the IRS by universities.

[23] See Chetty et al. (2011) for a comparison of total enrollment based on 1098-T forms and statistics from the Current Population Survey. Chetty et al. use this measure to analyze the impacts of Project STAR on college attendance. Dynarski et al. (2011) show that using data on college attendance from the National Clearinghouse yields very similar estimates to Chetty et al.'s findings, providing further confirmation that the 1098-T based college indicator is accurate.

and all 1098-T forms for year 1999, we group individuals by the higher education institution they attended in 1999. We take a 0.25% random sample of those not attending a higher education institution in 1999 and pool them together in a separate "no college" category. For each college or university (including the "no college" group), we then compute average W-2 earnings of the students in 2009 when they are aged 30. Among colleges attended by students in our data, the average value of our earnings index is $42,932 for four-year colleges and $28,093 for two-year colleges.[24] For students who did not attend college, the imputed mean earnings level is $16,361.

Neighborhood Quality. We use data from 1040 forms to identify each household's ZIP code of residence in each year. For non-filers, we use the ZIP code of the address to which the W-2 form was mailed. If an individual did not file and has no W-2 in a given year, we impute current ZIP code as the last observed ZIP code. We construct a measure of a neighborhood's SES using data on the percentage of college graduates in the individual's ZIP code from the 2000 Census.

Retirement Savings. We measure retirement savings using contributions to 401(k) accounts reported on W-2 forms from 1999-2010. We define saving for retirement as an indicator for ever contributing to a 401(k) during this period.

Teenage Birth. We first identify all women who claim a dependent when filing their taxes at any point before the end of the sample in tax year 2010. We observe dates of birth and death for all dependents and tax filers until the end of 2010 as recorded by the Social Security Administration. We use this information to identify women who ever claim a dependent who was born while the mother was a teenager (between the ages of 13 and 19 as of 12/31 the year the child was born). We refer to this outcome as having a "teenage birth," but note that this outcome differs from a direct measure of teenage birth in three ways. First, it does not capture teenage births to individuals who never file a tax return before 2010. Second, the mother must herself claim the child as a dependent at some point during the sample years. If the child is claimed as a dependent by the grandmother for all years of our sample, we would never identify the child. In addition to these two forms of under-counting, we also over-count the number of children because our definition could miscategorize other dependents as biological children. Because most such dependents tend to be elderly parents, the fraction of cases that are incorrectly categorized as teenage births is likely to be small. Even though this variable does not directly measure teenage births, we believe that it is a useful measure of outcomes in adulthood because it correlates with observables as expected (see

[24] For the small fraction of students who attend more than one college in a single year, we define college quality based on the college that received the largest tuition payments on behalf of the student.

Section 5.3). For instance, women who score higher on tests, attend college, or have higher income parents are significantly less likely to have teenage births.

Parent Characteristics. We link students to their parents by finding the earliest 1040 form from 1996-2010 on which the student was claimed as a dependent. We identify parents for 94.7% of students linked with tax records as adults. The remaining students are likely to have parents who did not file tax returns in the early years of the sample when they could have claimed their child as a dependent, making it impossible to link the children to their parents. Note that this definition of parents is based on who claims the child as a dependent, and thus may not reflect the biological parent of the child.

We define parental household income as Adjusted Gross Income (capped at $117,000, the 95th percentile in our sample), averaged over the three years when the child was 19-21 years old.[25] For years in which parents did not file, we impute parental household income from wages and unemployment benefits, each of which are reported on third-party information forms. We define marital status, home ownership, and 401(k) saving as indicators for whether the first primary filer who claims the child ever files a joint tax return, makes a mortgage interest payment (based on data from 1040's for filers and 1099's for non-filers), or makes a 401(k) contribution (based on data from W-2's) during the years when the child is between 19 and 21. We define mother's age at child's birth using data from Social Security Administration records on birth dates for parents and children. For single parents, we define the mother's age at child's birth using the age of the filer who claimed the child, who is typically the mother but is sometimes the father or another relative.[26] When a child cannot be matched to a parent, we define all parental characteristics as zero, and we always include a dummy for missing parents in regressions that include parent characteristics.

3.3 Analysis Dataset

Because most of the adult outcomes we analyze are at age 20 or afterward, we restrict our linked analysis sample to students who would graduate high school in the 2007-08 school year (and thus

[25] To account for changes in marital status, we always follow the primary filer who first claimed the child and define parent characteristics based on the tax returns filed by that parent when the child is between 19 and 21. For instance, if a single mother has a child and gets married when the child was 18, we would define household income as AGI including the mother and her new husband when the child is 19-21. If the child does not turn 21 before 2010, we code the parent characteristics as missing.

[26] We define the mother's age at child's birth as missing for 471 observations in which the implied mother's age at birth based on the claiming parent's date of birth is below 13 or above 65. These are typically cases where the parent does not have an accurate birth date recorded in the SSA file.

turn 20 in 2010) if they progress through school at a normal pace.[27] The school district records were linked to the tax data using an algorithm based on standard identifiers (date of birth, state of birth, gender, and names) described in Appendix A, after which individual identifiers were removed to protect confidentiality. 89.2% of the observations in the school district data were matched to the tax data and match rates do not vary with teacher VA (see Table 2 below).

The linked analysis dataset has one row per student per subject (math or English) per school year, as illustrated in Appendix Table 1. Each observation in the analysis dataset contains the student's test score in the relevant subject test, demographic information, and class and teacher assignment if available. Each row also lists all the students' available adult outcomes (e.g. college attendance and earnings at each age) as well as parent characteristics. We organize the data in this format so that each row contains information on a treatment by a single teacher conditional on pre-determined characteristics, facilitating estimation of equation (3). We account for the fact that each student appears multiple times in the dataset by clustering standard errors as described in section 4.1.

To maximize precision, we estimate teacher value-added using all years for which school district data are available (1991-2009). However, the impacts of teacher VA on test scores and adult outcomes that we report in the main text use only the observations in the linked analysis dataset (i.e., exclude students who would graduate high school after 2008), unless otherwise noted.[28]

3.4 Summary Statistics

The analysis dataset contains 6.0 million student-year-subject observations, of which 4.8 million have information on teachers. Table 1 reports summary statistics for the linked analysis dataset; see Appendix Table 2 for corresponding summary statistics for the full school district data used to estimate teacher value-added. Note that the summary statistics are student-school year-subject means and thus weight students who are in the district for a longer period of time more heavily, as does our empirical analysis. There are 974,686 unique students in our analysis dataset; on average, each student has 6.14 subject-school year observations.

The mean test score in the analysis sample is positive and has a standard deviation below

[27] A few classrooms contain students at different grade levels because of retentions or split-level classroom structures. To avoid dropping a subset of students within a classroom, we include every classroom that has at least one student who would graduate school during or before 2007-08 if he progressed at the normal pace. That is, we include all classrooms in which $\min_i(12 + \text{school year} - \text{grade}_i) \leq 2008$.

[28] Within the analysis data, we use all observations for which the necessary data are available. In particular, when estimating the impacts of VA on scores, we include observations that were not matched to the tax data.

1 because we normalize the test scores in the full population that includes students in special education classrooms and schools (who typically have lower test scores). The mean age at which students are observed is 11.7 years. 76% of students are eligible for free or reduced price lunches. 2.7% of the observations are for students who are repeating the current grade.

The availability of data on adult outcomes naturally varies across cohorts. There are more than 4.6 million observations for which we observe college attendance at age 20. We observe earnings at age 25 for 2.2 million observations and at age 28 for 850,000 observations. Because many of these observations at later ages are for older cohorts of students who were in middle school in the early 1990s, they do not contain information on teachers. As a result, there are only 1.4 million student-subject-school year observations for which we see *both* teacher assignment and earnings at age 25, 376,000 at age 28, and only 63,000 at age 30. The oldest age at which the sample is large enough to obtain reasonably precise estimates of teachers' impacts on earnings turns out to be age 28. Mean earnings at age 28 is $20,327 (in 2010 dollars), which includes zero earnings for 34% of the sample.

For students whom we are able to link to parents, mothers are 28 years old on average when the student was born. The mean parent household income is $35,476, while the median is $27,144 Though our sample includes more low income households than would a nationally representative sample, it still includes a substantial number of higher income households, allowing us to analyze the impacts of teachers across a broad range of the income distribution. The standard deviation of parent income is $31,080, with 10% of parents earning more than $82,630.

As a benchmark for evaluating the magnitude of the causal effects estimated below, Appendix Tables 3-6 report estimates of OLS regressions of the adult outcomes we study on test scores. Both math and English test scores are highly positively correlated with earnings, college attendance, and neighborhood quality and are negatively correlated with teenage births. In the cross-section, a 1 SD increase in test score is associated with a $7,440 (37%) increase in earnings at age 28. Conditional on prior-year test scores and other controls that we use in our analysis below, a 1 SD increase in the current test score is associated with $2,545 (11.6%) increase in earnings on average. We show below that the causal impact of teacher VA on earnings is commensurate to this correlation in magnitude.

4 Does Value-Added Accurately Measure Teacher Quality?

Recent studies by Kane and Staiger (2008) and Rothstein (2010) among others have reached conflicting conclusions about whether VA estimates are biased by student sorting (i.e., whether Assumption 1 in Section 2.2 holds). In this section, we revisit this debate by presenting new tests for bias in VA estimates.

4.1 Empirical Methodology

Throughout our empirical analysis, we regress various outcomes on estimated teacher value-added. In this subsection, we discuss four aspects of our methodology that are relevant for all the regression estimates reported below: (1) leave-out mean estimation of VA, (2) control vectors, (3) standard error calculations, and (4) the treatment of outliers.

First, there is a mechanical correlation between $\hat{\mu}_j$ and student outcomes in a given school year because $\hat{\mu}_j$ is estimated with error and these errors also affect student outcomes.[29] We address this problem by following Jacob, Lefgren and Sims (2010) and use a leave-year-out (jackknife) mean to calculate teacher quality.[30] For example, when predicting teachers' effects on student outcomes in 1995, we estimate $\widehat{\mu}_j^{1995}$ based on all years of the sample *except* 1995. We then regress outcomes for students in 1995 on $\widehat{\mu}_j^{1995}$. More generally, for each observation in year t, we omit score residuals from year t when calculating teacher quality.[31] This procedure is essential to eliminate mechanical biases due to estimation error in $\hat{\mu}_j$ both in our tests for sorting and our estimates of teachers' impacts on adult outcomes.[32]

Second, we use a control vector that parallels existing VA models (e.g., Kane and Staiger 2008)

[29] This problem does not arise when estimating the impacts of treatments such as class size because the treatment is observed; here, the size of the treatment (teacher VA) must itself be estimated, leading to correlated estimation errors.

[30] Because we need at least two classes to define a leave-out mean, our analysis only applies to the population of teachers whom we see teaching two or more classes between 1991 and 2009. Among the classrooms with the requisite controls to estimate value-added (e.g. lagged test scores), we are unable to calculate a leave-out measure of VA for 9% of students because their teachers are observed in the data for only one year. The first-year VA of teachers who leave after one year is 0.01 SD lower than the first-year VA of those who stay for more years. Hence, the mean VA of the subset of teachers in our sample is only 0.001 SD higher than mean VA in the population, suggesting that our estimates are likely to be fairly representative of teacher effects in the full population.

[31] An alternative approach is to split the sample in two, for instance using data after 1995 to estimate teacher VA and data before 1995 to estimate its impacts on outcomes for students who are old enough to be seen in the tax data. We find that such a split-sample approach yields similar but less precise estimates.

[32] Regressing student outcomes on teacher VA without using a leave-out mean effectively introduces the same estimation errors on both the left and right hand side of the regression, yielding biased estimates of teachers' causal impacts. This is the reason that Rothstein (2010) finds that "fifth grade teachers whose students have above average fourth grade gains have systematically lower estimated value-added than teachers whose students underperformed in the prior year."

to estimate student test score residuals using (3):

$$A_{igt} = f_{1g}(A_{i,t-1}) + f_2(e_{j(i,g,t)}) + \phi_1 X_{igt} + \phi_2 \bar{X}_{c(i,g,t)} + \nu_{igt}$$

We parameterize the control function for lagged test scores $f_{1g}(A_{i,t-1})$ using a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English. We interact these cubics with the student's grade level to permit flexibility in the persistence of test scores as students age. We parametrize the control function for teacher experience $f_2(e_{j(i,g,t)})$ using dummies for years of experience from 0 to 5, with the omitted group being teachers with 6 or more years of experience.[33] The student-level control vector X_{igt} consists of the following variables: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, limited English. The class-level control vector $\bar{X}_{c(i,g,t)}$ includes (1) class size and class-type indicators (honors, remedial), (2) cubics in class and school-grade means of prior-year test scores in math and English each interacted with grade, (3) class and school-year means of all the individual covariates X_{igt}, and (4) grade and year dummies. To avoid estimating VA based on very few observations, we follow Kane and Staiger (2008) and exclude classrooms that have fewer than 7 observations with test scores and the full vector of controls X_{igt} (2% of observations). Importantly, the control vectors X_{igt} and $\bar{X}_{c(i,g,t)}$ consist entirely of variables from the school district dataset. We adopt this approach because our goal is to assess properties of value-added estimated without access to information available in tax data, which will not typically be available to school districts.

When estimating the impacts of teacher VA on adult outcomes using (9), we omit the student-level controls X_{igt}. By omitting X_{igt}, we can conduct most of our analysis of long-term impacts using a dataset collapsed to class means, which significantly reduces computational costs. We show in Appendix Table 7d that the inclusion of individual controls has little impact on the coefficients and standard errors of interest for a selected set of specifications.

Third, our outcomes have a correlated error structure because students within a classroom face common class-level shocks and because our analysis dataset contains repeat observations on students in different grades. One natural way to account for these two sources of correlated errors is to cluster standard errors by both student and classroom (Cameron, Gelbach, and Miller 2011). Unfortunately, implementing two-way clustering on a dataset with 6 million observations was infeasible because of computational constraints. We instead cluster standard errors at the

[33] We choose this functional form because prior work (e.g. Rockoff 2004) has shown that the impacts of teacher experience rise sharply and then stabilize after the first three years.

school by cohort level, which adjusts for correlated errors across classrooms and repeat student observations within a school. Clustering at the school-cohort level is convenient because it again allows us to conduct our analysis on a dataset collapsed to class means. We evaluate the robustness of our results to alternative forms of clustering in Appendix Table 7 and show that school-cohort clustering yields more conservative confidence intervals than the more computationally intensive techniques.

Finally, in our baseline specifications, we exclude classrooms taught by teachers whose estimated VA $\hat{\mu}_j^t$ falls in the top two percent for their subject (above 0.21 in math and 0.13 in English) because these teachers' impacts on test scores appear suspiciously consistent with testing irregularities indicative of cheating. Jacob and Levitt (2003) develop a proxy for cheating that measures the extent to which a teacher generates very large test score gains that are followed by very large test score losses for the same students in the subsequent grade. Jacob and Levitt establish that this is a valid proxy by showing that it is highly correlated with unusual answer sequences that directly point to test manipulation. Teachers in the top 2% of our estimated VA distribution are significantly more likely to show suspicious patterns of test scores gains, as defined by Jacob and Levitt's proxy (see Appendix Figure 1).[34] We therefore trim the top 2% of outliers in all the specifications reported in the main text. We investigate how trimming at other cutoffs affects our main results in Appendix Table 8. The qualitative conclusion that teacher VA has long-term impacts is not sensitive to trimming, but including teachers in the top 2% reduces our estimates of teachers' impacts on long-term outcomes by 20-40%. In contrast, excluding the bottom 2% of the VA distribution has little impact on our estimates, consistent with the view that test manipulation to obtain high test score gains is responsible for the results in the upper tail. Directly excluding teachers who have suspect classrooms based on Jacob and Levitt's proxy for cheating yields very similar results to trimming on VA itself.

Because we trim outliers, our baseline estimates should be interpreted as characterizing the relationship between VA and outcomes below the 98th percentiles of VA. This is the relevant range for many questions, such as calculating the gains of switching a child from an average teacher to a teacher 1 SD above the mean. If school districts can identify and eliminate teacher cheating –

[34] Appendix Figure 1 plots the fraction of classrooms that are in the top 5 percent according to Jacob and Levitt's proxy, defined in the notes to the figure, vs. our leave-out-year measure of teacher value-added. On average, classrooms in the top 5 percent according to the Jacob and Levitt measure have test score gains of 0.46 SD in year t followed by mean test score *losses* of 0.43 SD in the subsequent year. Stated differently, teachers' impacts on future test scores fade out much more rapidly in the very upper tail of the VA distribution. Consistent with this pattern, these exceptionally high VA teachers also have very little impact on their students' long-term outcomes.

e.g. by analyzing the persistence of test score gains as suggested by Jacob and Levitt – our estimates would also shed light on the gains from retaining the remaining high-VA teachers. Nevertheless, the fact that high-VA outliers do not have lasting impacts on scores or adult outcomes serves as a warning about the risks of manipulability of VA measures. The signal content of VA measures could be severely reduced if teachers game the system further when VA is actually used to evaluate teachers. This is perhaps the most important caveat to our results and a critical area for further work, as we discuss in the conclusion.

4.2 VA Estimates and Out-of-Sample Forecasts

The first step in our empirical analysis is to estimate leave-year-out teacher effects $\hat{\mu}_j^t$ for each teacher j and year t in our sample. We estimate VA using all years in the school district data for which we have teacher information (1991-2009). The standard deviation of teacher effects is $\sigma_\mu = 0.118$ in math and $\sigma_\mu = 0.081$ in English, very similar to estimates from prior work. Note that these standard deviations measure the dispersion in teacher effects that is orthogonal to teacher experience as well as other controls.[35] Throughout, we scale $\hat{\mu}_j^t$ in units of *student* test scores, i.e., a 1 unit increase in $\hat{\mu}_j^t$ refers to a teacher whose VA is predicted to raise student test scores by 1 SD. Because the standard deviation of teacher effects is approximately 0.1 SD of the student test score distribution (averaging across math and English), a 1 SD increase in teacher VA corresponds to an increase of 0.1 in $\hat{\mu}_j^t$.

We begin our evaluation of the properties of $\widehat{\mu}_j^t$ by verifying that our VA estimates have predictive power for test score gains outside the sample on which they were estimated. Under our assumption in (3) that true teacher effects μ_j are time-invariant, a 1 SD increase in $\widehat{\mu}_j^t$ should be associated with a 1 SD increase in test scores in year t.[36] Figure 1a plots student test scores (combining English and math observations) vs. our leave-year-out estimate of teacher VA in our linked analysis dataset. We condition on the classroom-level controls used when estimating the value-added model in this and all subsequent figures by regressing both the x- and y-axis variables on the vector of controls and then computing residuals. We then bin the student-subject-year residuals into twenty equal-size groups (vingtiles) of $\widehat{\mu}_j^t$ and plot the mean residual score in each

[35] Students assigned to first-year teachers have 0.03 SD lower test score gains, consistent with prior work. Because the impact of experience on scores is small, we have insufficient power to estimate its impacts on adult outcomes; we can rule out neither 0 effects nor effects commensurate to the impacts of VA estimated below. We therefore do not analyze teacher experience further in this paper.

[36] Although the estimation error in value-added leads to attenuation bias, the shrinkage correction we implement in (4) exactly offsets the attenuation bias so that a 1 unit increase in $\widehat{\mu}_j^t$ should raise scores by 1 unit.

bin. Note that these binned scatter plots provide a non-parametric representation of the conditional expectation function but do not show the underlying variance in the individual-level data. The regression coefficient and standard error reported in each figure are estimated on the micro data, with standard errors clustered by school-cohort as described above.

Figure 1a shows that a teacher with $\widehat{\mu}_j^t = 1$ generates a 0.86 SD increase in students' test scores in year t, with a t-statistic over 80 (see also Column 1 of Table 2). This confirms that the VA estimates are highly predictive of student test scores. The coefficient on $\widehat{\mu}_j^t$ is below 1, consistent with the findings of Kane and Staiger (2008), most likely because teacher value-added is not in fact a time-invariant characteristic. For instance, teacher quality may fluctuate when teachers switch schools or grades (Jackson 2010) and may drift over time for other reasons (Goldhaber and Hansen 2010). Such factors reduce the accuracy of forecasts based on data from other years. Because we estimate teacher VA using data from 1991-2009 but only include cohorts who graduate from high school before 2008 in our analysis dataset, the time span between the point at which we estimate VA and analyze test score impacts is especially large in our analysis sample. Replicating Column 1 of Table 2 on the full sample used to estimate teacher VA yields a coefficient on $\widehat{\mu}_j^t$ of 0.96. Because we are forced to use data from more distant years to identify value-added, our estimates of the impacts of teacher quality on adult outcomes may be slightly downward-biased.[37]

The relationship between $\widehat{\mu}_j^t$ and students' test scores in Figure 1a could reflect either the causal impact of teachers on achievement or persistent differences in student characteristics across teachers. For instance, $\widehat{\mu}_j^t$ may forecast students' test score gains in other years simply because some teachers are always assigned students with higher income parents. We now implement two sets of tests for such sorting.

4.3 Test 1: Selection on Observable Characteristics

Value-added estimates consistently measure teacher quality only if they are uncorrelated with unobserved components of student scores. A natural first test of this identifying assumption is to examine the correlation between our estimates of VA and variables omitted from standard VA models.[38] We use two sets of variables to evaluate selection: parent characteristics and prior test

[37] We do not account for variation over time in VA because our primary goal is to assess the properties of teacher VA measures currently being used by school districts. In future work, it would be interesting to develop time-varying measures of VA and evaluate whether they are better predictors of adult outcomes.

[38] Such correlation could arise from either actual selection of students to teachers with higher quality μ_j or sorting across teachers that is unrelated to true quality but generates measurement error in $\hat{\mu}_j^t$ that is correlated with student characteristics. Either of these sources of correlation would violate Assumption 1 and generate biased estimates of VA.

scores.

Parent Characteristics. The parent characteristics from the tax data are ideal to test for selection because they have not been used to fit value-added models in prior work but are strong predictors of student achievement. We collapse the parent characteristics into a single index by regressing test scores on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income.[39] Let A^p_{it} denote the predicted test score for student i in year t in this regression, which we calculate only for students for whom test score data are available. These predicted test scores are an average of the parent characteristics, weighted optimally to reflect their relative importance in predicting test scores. The standard deviation of predicted test scores is 0.26, roughly 30% of the standard deviation of actual test scores in our analysis sample.

Figure 1b plots $\widehat{A}^p_{c,g-1}$ against teacher VA measured using a leave-year-out mean as described above. There is no relationship between predicted scores and teacher VA. At the upper bound of the 95% confidence interval, a 1 standard deviation increase in teacher VA raises predicted scores based on parent characteristics by 0.01 SD (see also Column 2 of Table 2). This compares with an actual score impact of 0.86 SD, showing that very little of the association between teacher VA and actual test scores is driven by sorting on omitted parent characteristics. Note that this result does *not* imply that students from higher vs. lower socioeconomic status families uniformly get teachers of the same quality. Our finding is that controlling for the rich set of observables available in school district databases, such as test scores in the previous grade, is adequate to account for sorting of students to teachers based on parent characteristics. That is, if we take two students who have the same 4th grade test scores, classroom characteristics, ethnicity, suspensions, etc., the student assigned to a teacher with higher estimated VA in grade 5 does not systematically have different parental income or other characteristics.

A second, closely related method of assessing selection on parent characteristics is to control for predicted scores A^p_{it} when estimating the impact of VA on actual scores. Columns 3-4 in Table 2 restrict to the sample in which both score and predicted score are non-missing; the coefficient on $\widehat{\mu}^t_j$ changes only from 0.866 to 0.864 after controlling for predicted scores. Note that parent characteristics have considerable predictive power for test scores even conditional on the controls

[39] We code the parent characteristics as 0 for the 5.4% of matched students for whom we are unable to find a parent, and include an indicator for having no parent matched to the student. We also code mother's age at child's birth as 0 for a small number of observations where we match parents but do not have data on parents' ages, and include an indicator for such cases.

used to estimate the value-added model; the t-statistic on the predicted score A_{it}^p exceeds 60. The fact that parent characteristics are strong predictors of residual test scores yet are uncorrelated with $\widehat{\mu}_j^t$ suggests that the degree of bias in VA estimates is likely to be small (Altonji, Elder, and Taber 2005).

A third approach to evaluating the extent to which the omission of parent characteristics affects VA estimates is to re-estimate $\widehat{\mu}_j^t$, controlling for the parent characteristics to begin with. We repeat the three-step estimation procedure in Section 2.2, controlling for mean parent characteristics by classroom when estimating (3) using the same functional form used above to predict test scores. We then correlate estimates of teacher VA that control for parent characteristics with our original estimates that condition only on school-district observables. The correlation coefficient between the two VA estimates is 0.999, as shown in rows 1 and 2 of Table 3. All three tests show that selection on previously unobserved parent characteristics generates minimal bias in standard VA estimates.

Prior Test Scores. Another set of pre-determined variables that can be used to test for selection are prior test scores (Rothstein 2010). Because value-added models control for $A_{i,t-1}$, one can only evaluate sorting on $A_{i,t-2}$ (or, equivalently, on lagged gains, $A_{i,t-1} - A_{i,t-2}$). The question is whether controlling for additional lags substantially affects VA estimates once one controls for $A_{i,t-1}$. We now present three tests to answer this question that parallel those above for parent characteristics.

We first examine whether twice-lagged test scores are correlated with our baseline estimates of VA. Figure 1c plots twice-lagged scores $A_{i,t-2}$ against teacher VA, following the same methodology used to construct Figure 1a. There is virtually no relationship between VA and twice-lagged score conditional on the controls used to estimate the VA model. As a result, controlling for $A_{i,t-2}$ when estimating the impact of VA on out-of-sample test scores has little effect on the estimated coefficient (columns 6-7 of Table 2). The coefficient on VA is stable despite the fact that $A_{i,t-2}$ has significant predictive power for $A_{i,t}$, even conditional on $A_{i,t-1}$ and $\bar{X}_c$; the t-statistic on $A_{i,t-2}$ exceeds 350. Finally, controlling flexibly for $A_{i,t-2}$ at the individual level (using cubics in math and English scores) when estimating the VA model does not affect estimates significantly. The correlation coefficient between our baseline VA estimates and estimates that control for $A_{i,t-2}$ is 0.975, as shown in row 3 of Table 3. We conclude based on these tests that selection on grade $t-2$ scores generates minimal bias in VA estimates once one conditions on $t-1$ characteristics.

We further develop this test by examining the correlation of our baseline VA measure with

additional leads and lags of test scores. If our VA measures reflect the causal impact of teachers, the correlation between current teacher VA on test scores should jump in the current year. To test this hypothesis, we estimate (9), changing the dependent variable to test scores $A_{i,t+s}$ for $s \in [-4, 4]$, four years before and after the current grade t. Figure 2 plots the coefficients on current teacher VA from each of these regressions.[40] As predicted, teachers' impacts on scores jump at the end of the grade taught by that teacher. A 1 unit increase in teacher VA raises end-of-grade test scores by 0.86 SD, matching the estimate in column 1 of Table 2. In contrast, the same increase in teacher VA in grade g has essentially no impact on test scores prior to grade g. This finding suggests that VA measures capture causal effects of teachers rather than systematic differences across teachers in their students' characteristics, as such characteristics would have to be uncorrelated with past test scores and only affect the current score.

Figure 2 also shows that the impact of current teacher VA fades out in subsequent grades. Prior studies (e.g., Kane and Staiger 2008, Jacob, Sims, and Lefgren 2010, Rothstein 2010) document similar fade-out after one or two years but have not determined whether test score impacts continue to deteriorate after that point. The broader span of our dataset allows us to estimate test score persistence more precisely.[41] In our data, the impact of a 1 SD increase in teacher quality stabilizes at approximately 0.3 SD after 3 years, showing that students assigned to teachers with higher VA achieve long-lasting test score gains.

The last column of Table 2 analyzes the correlation between teacher VA and the probability that a student is matched to the tax data. In this column, we regress an indicator for being matched on teacher VA, using the same specification as in the other columns. There is no significant relationship between VA and match rates, suggesting that our estimates of the impacts of VA on outcomes in adulthood are unlikely to be biased by attrition.

4.4 Test 2: Teacher Switching Quasi-Experiments

The preceding tests show that the bias in VA estimates due to the omission of observables such as parent characteristics and twice-lagged scores is minimal. They do not, however, rule out the

[40] The estimates underlying this figure and their associated standard errors are reported in Appendix Table 9. Naturally, the grades used to estimate each of the points in Figure 2 vary because scores are only available for grades 3-8. We continue to find that VA has an effect on prior test scores that is two orders of magnitude smaller than its impact on current test scores if we restrict to individual grades and use the available leads and lags (e.g. two leads and two lags for grade 6).

[41] For instance, Jacob, Lefgren, and Sims estimate one-year persistence using 32,422 students and two-year persistence using 17,320 students. We estimate one-year persistence using more than 2.8 million student-year-subject observations and four-year persistence using more than 790,000 student-year-subject observations.

possibility that students are sorted to teachers based on unobservable characteristics orthogonal to these variables. The ideal method of testing for selection on unobservables is to evaluate whether VA estimates using observational data accurately predict students' test score gains when students are randomly assigned to teachers. Kane and Staiger (2008) implement such an experiment in Los Angeles involving approximately 3,500 students and 150 teachers. Kane and Staiger's point estimates suggest that there is little bias in VA estimates, but their 95% confidence interval is consistent with bias of up to 50% because of their relatively small sample size (Rothstein 2010). Moreover, Rothstein notes that because certain classes and schools were excluded from the experiment, the external validity of the findings is unclear.

Motivated by these concerns, we develop a quasi-experimental method of estimating the degree of bias due to selection on unobservables. Our approach yields more precise estimates of the degree of bias on a representative sample of a school district's student population.

Research Design. Our research design exploits the fact that adjacent cohorts of students within a school are frequently exposed to teachers with very different levels of VA because of teacher turnover. In our school district dataset, 14.5% of teachers switch to a different grade within the same school the following year, 6.2% of teachers switch to a different school within the same district, and another 6.2% switch out of the district entirely. These changes in the teaching staff from one year to the next generate variation in VA that is "quasi-experimental" in the sense that it is plausibly orthogonal to students' characteristics.

To understand our test, suppose a high-VA teacher moves from 4th grade in school s to another school between 1994 and 1995. Because students entering grade 4 in school s in 1995 have lower VA teachers on average, their mean test scores should be lower than the 1994 cohort if VA estimates capture teachers' causal impacts. Moreover, the size of the change in test scores across these adjacent cohorts should correspond to the change in mean VA. For example, in a school-grade cell with three classrooms, the loss of a math teacher with a VA estimate of 0.3 based on prior data should decrease average math test scores in the entire school-grade cell by 0.1. Importantly, because we analyze the data at the school-grade level, we do *not* exploit information on classroom assignment for this test, eliminating any bias due to non-random assignment of students across classrooms.

Changes in the quality of the teaching staff across school years constitute quasi-experimental variation under the assumption that they are uncorrelated with changes in the quality of students across adjacent cohorts. Let $\Delta\overline{\hat{\mu}}_{sgmt}$ denote the change in mean teacher VA $\hat{\mu}_{sgmt}$ from year

$t-1$ to year t in grade g in subject m (math or reading) in school s, and define mean changes in student unobservables $\Delta\overline{\varepsilon}_{sgmt}$ and $\Delta\overline{\varepsilon}^{\mu}_{sgmt}$ analogously. The identification assumption underlying the quasi-experimental design is

$$Cov\left(\Delta\overline{\hat{\mu}}_{sgmt}, \Delta\overline{\varepsilon}_{sgmt}\right) = 0 \quad \text{and} \quad Cov\left(\Delta\overline{\hat{\mu}}_{sgmt}, \Delta\overline{\varepsilon}^{\mu}_{sgmt}\right) = 0. \tag{11}$$

This assumption requires that the change in mean VA within a school-grade cell is uncorrelated with the change in the average quality of students, as measured by unobserved determinants of scores and earnings. This assumption could potentially be violated by endogenous student or teacher sorting. Student sorting at an annual frequency is minimal because of the costs of changing schools. During the period we study, most students would have to move to a different neighborhood to switch schools, which families would be unlikely to do simply because a single teacher leaves or enters a given grade. While endogenous teacher sorting is plausible over long horizons, the sharp changes we analyze are likely driven by idiosyncratic shocks such as changes in staffing needs, maternity leaves, or the relocation of a spouses. Hence, we believe that (11) is a plausible restriction at high frequencies in our data and we present evidence supporting this assumption below.

Our approach complements recent work analyzing the impacts of teacher turnover on student achievement, but is the first to use turnover to validate VA models directly. Rivkin, Hanushek, and Kain (2005) identify the variance of teacher effects from differences in variances of test score gains across schools with low vs. high teacher turnover. In contrast, we identify the impacts of teachers from first moments – the relationship between changes in mean scores across cohorts and mean teacher quality – rather than second moments. Our approach does not rely on comparisons across schools with different levels of teacher turnover, which may also differ in other unobserved dimensions that could impact earnings directly. For instance, Ronfeldt et al. (2011) show that higher rates of teacher turnover lead to lower student achievement, although they do not assess whether the mean value-added of the teaching staff predicts student achievement across cohorts.[42] Jackson and Bruegmann (2009) document peer effects by analyzing whether the VA of teachers who enter or exit affects the test scores of *other* teachers' students in their school-grade cell, but do not compare changes in mean test scores by cohort to the predictions of VA models.[43]

[42] This is less of a concern in Rivkin, Hanushek, and Kain's analysis of test score impacts because they are able to test whether the variance of test score gains is higher in grades with high turnover, thereby netting out school fixed effects. This is infeasible with outcomes in adulthood, which are observed only after schooling is complete. Rivkin, Hanushek, and Kain are unable to implement the teacher switcher design we develop here because they do not have class assignment data and thus cannot estimate each teacher's individual effect μ_j, which is necessary to construct the school-grade-cohort level mean of teacher quality.

[43] The peer effects documented by Jackson and Bruegmann could in principle affect our validation of VA using

Event Studies. We begin our analysis of teaching staff changes with event studies of scores around the entry and exit of high and low VA teachers (Figure 3). Let year 0 denote the school year that a teacher enters or exits a school-grade-subject cell and define all other school years relative to that year (e.g., if the teacher enters in 1995, year 1992 is -3 and year 1997 is +2). We define an entry event as the arrival of a teacher who did not teach in that school-grade-subject cell for the three preceding years; analogously, we define an exit event as the departure of a teacher who does not return to the same school-grade-subject cell for at least three years. We estimate VA for each teacher using only data *outside* the six-year window used for the event studies to eliminate bias due to correlated estimation errors.[44] We define a teacher as "high VA" if her estimated VA based on years outside the event study window is in the top 5% of the distribution for her subject; a "low VA" teacher has an estimated VA in the bottom 5%.[45] To obtain a balanced sample, we analyze events for which we have data on average test scores at the school-grade-subject level for at least three years before and three years after the event.[46] Because these balanced event studies require data over several years, we use the full school district data spanning 1991-2009 (rather than only the analysis sample linked to the tax data), excluding school-grade-subject cells in which we have no information on teachers.

Figure 3a plots the impact of the entry of a high-VA teacher on mean test scores. The solid series plots school-grade-subject-year means of test scores in the three years before and after a high-VA teacher enters the school-grade-subject cell, with year fixed effects removed to eliminate any secular trends.[47] We do not condition on any other covariates in this figure: each point simply shows average test scores for different cohorts of students within a school-grade-subject cell adjusted for year effects. When a high-VA teacher arrives, end-of-year test scores in the subject and grade

the switcher design. However, peer learning effects are likely to be smaller with teacher exits than entry, provided that knowledge does not deteriorate very rapidly. We find that teacher entry and exit yield broadly similar results, suggesting that spillovers across teachers are not a first-order source of bias for our technique.

[44] More precisely, we calculate VA for each teacher in each year excluding a five year window (two years prior, the current year, and two years post). Coupled with our definitions of entry and exit – which require that the teacher not be present in the school-grade-subject cell for 3 years before or after the event – this ensures that we do not use any data from the relevant cell between event years -3 and +2 to compute teacher VA.

[45] In cases where multiple teachers enter or exit at the same time, we use the teachers' mean VA in decided whether it falls in the top or bottom 5% of the VA distribution. To eliminate potential selection bias, we include high VA outliers in these event studies and our cross-cohort research design more generally; that is, we do not drop the top 2% outliers who may achieve test score gains via manipulation as we do in our baseline analysis that exploits variation across classrooms. Excluding these outliers yields very similar conclusions, as can be seen from Figure 4, which shows that changes in VA predict changes in test scores accurately throughout the value-added distribution.

[46] In school-grade-subject cells with multiple events (e.g. entry of a high VA teacher in both 1995 and 1999), we include all such events by stacking the data and using the three years before and after each event.

[47] We remove year fixed effects in this and all other event study graphs by regressing mean test scores on year dummies, computing residuals, and adding back the mean test score in the estimation sample to facilitate interpretation of the scale.

taught by that teacher rise immediately. The null hypothesis that test scores do not change from year -1 to year 0 is rejected with $p < 0.001$, with standard errors clustered by school-cohort as above. The magnitude of the increase in test scores, which is 0.036 SD from year -1 to year 0, is very similar to what one would forecast based on the change in mean teacher VA. Mean VA rises by 0.044 SD from year -1 to year 0.[48] The estimate in Column 1 of Table 2 based on cross-classroom variation implies that we should expect this increase in teacher VA to increase students' scores by $0.044 \times 0.861 = 0.038$ SD.[49] The hypothesis that the observed change in mean scores of 0.036 equals the predicted change of 0.038 is not rejected ($p = 0.76$).

Figure 3a implies that value-added accurately measures teachers' impacts on students' test scores under the identification assumption in (11). We evaluate this assumption by examining test scores for the same cohort of students in the previous school year. For example, the entry of a high-VA teacher in grade 5 in 1995 should have no impact on the same cohort's 4th grade test scores in 1994. The dashed line in Figure 3a plots test scores in the previous grade for the same cohorts of students. Test scores in the prior grade remain stable across cohorts both before and after the new teacher arrives, supporting our view that school quality and student attributes are not changing sharply around the entry of a high-VA teacher.[50]

The remaining panels of Figure 3 repeat the event study in Panel A for other types of arrivals and departures. Figure 3b examines current and lagged test scores around the departure of a high-VA teacher. There is a smooth negative trend in both current and lagged scores, suggesting that high-VA teachers leave schools that are declining in quality. However, scores in the grade taught by the teacher drop sharply relative to prior scores in the event year, showing that the departure of the high quality teacher lowers the achievement of subsequent cohorts of students. Figures 3c and 3d analyze the arrival and departure of low VA teachers. Test scores in the grade taught by the teacher fall sharply relative to prior-year scores when low VA teachers enter a school-grade cell and rise sharply when low VA teachers leave. In every case, the magnitude of the test score change is significantly different from 0 with $p < 0.001$ but is not significantly different from what one would

[48] When computing this change in mean VA, we weight teachers by the number of students they teach. For teachers who do not have any VA measures from classrooms outside the leave-out window, we impute VA as the mean leave-out VA in the sample. For a small fraction of students for whom we have no teacher information (5% of observations), we also impute teacher VA as the sample mean.

[49] We expect the observed change in scores when a high VA teachers enters to be smaller than the change in mean VA for the same reason that the cross-class coefficient is less than 1 – namely that teacher VA likely changes over time, and we use data from at least three years before or after the event to estimate teacher VA. Hence, the appropriate test for bias is whether the change in test scores matches what one would predict based on the cross-class coefficient of 0.861.

[50] We also find that class size does not change significantly around the entry and exit events we study.

forecast based on the change in mean teacher VA.[51] Together, these event studies provide direct evidence that deselecting low VA teachers and retaining high-VA teachers improves the academic achievement of students.

Teaching Staff Changes. The event studies focus on the tails of the teacher VA distribution and thus exploit only a small fraction of the variation arising from teacher turnover in the data. We now exploit all the variation due to teaching staff changes to obtain a broader estimate of the degree of bias in VA measures. To do so, we first estimate VA for each teacher using data excluding a given pair of adjacent years, $t-1$ and t. We then calculate the change in mean teacher VA for each school-grade-subject-year cell and define $\Delta\overline{\mu}_{sgmt}$ as mean teacher VA in year t minus mean teacher VA in year $t-1$. With this definition, the variation in $\Delta\overline{\mu}_{sgmt}$ is driven purely by changes in the teaching staff and not by changes in the estimated VA for the teachers. This leave-out technique again ensures that changes in mean test scores across cohorts t and $t-1$, which we denote by $\Delta\overline{A}_{sgmt}$, are not spuriously correlated with estimation error in $\Delta\overline{\mu}_{sgmt}$.

Figure 4a plots the changes in mean test scores across cohorts $\Delta\overline{A}_{sgmt}$ against changes in mean teacher value-added $\Delta\overline{\mu}_{sgmt}$. As in the event studies, we remove year fixed effects so that the estimate is identified purely from differential changes in teacher quality across school-grade-subject cells over time. For comparability with the estimates in Table 2, we only use data from the linked analysis sample in this figure. Changes in the quality of the teaching staff strongly predict changes in test scores across consecutive cohorts of students in a school-grade-subject cell. The estimated coefficient on $\Delta\overline{\mu}_{sgmt}$ is 0.843, with a standard error of 0.053 (Table 4, Column 1). This estimate is very similar to the coefficient of 0.861 obtained from the cross-class out-of-sample forecast in Column 1 of Table 2. The point estimate of the degree of bias is 2% and is not statistically distinguishable from 0. At the lower bound of the 95% confidence interval, we reject bias of more than 14%.

Figures 4b through 4d evaluate the identification assumption in (11) underlying our research design using additional placebo tests. Each of these panels replicates Figure 4a with a different dependent variable; the corresponding regression estimates are reported in Columns 2-4 of Table 4. Figure 4b shows that changes in the quality of the teaching staff are unrelated to changes in parent characteristics, as captured by the predicted score measure used in Column 2 of Table 2. In Figures 4c and 4d, we examine the impact of changes in the teaching staff in one subject on mean scores

[51] The event studies in Figure 3 pool variation from teachers switching within schools, across schools, and out of the district. Teacher switches across grades within schools have similar impacts on test scores to teacher switches out of schools.

in the *other* subject. Here, it is important to distinguish between elementary and middle schools. In elementary school, students have one teacher for both math and English. Because elementary school teachers' math and English VA are highly correlated ($r = 0.59$), changes in mean teacher VA across cohorts are highly correlated across the two subjects. But students have different teachers for the two subjects in middle school, and changes in mean VA across cohorts in one subject are thus uncorrelated with changes in mean VA in the other subject. Hence, if (11) holds, we would expect changes in mean teacher VA in English to have much smaller effects on test scores in math (and vice versa) in middle school relative to elementary school. Figures 4c and 4d show that this is indeed the case. In elementary school, changes in mean teacher VA across cohorts strongly predict changes in test scores in the other subject ($t = 11.9$, $p < 0.001$), whereas in middle schools, the coefficient is near zero and statistically insignificant ($t = 0.04$, $p = 0.97$).

Given the results of these placebo tests, any violation of (11) would have to be driven by selection on unobserved determinants of test scores that have no effect on prior test scores and only affect the subject in which teaching staff changes occur. We believe that such selection is implausible given the information available to teachers and students and the constraints they face in sorting across schools at high frequencies.

Finally, we use our quasi-experimental design to evaluate how the choice of controls affects the degree of bias in VA estimates. The results of this analysis are reported in the last column of Table 3. For comparability, we estimate the models on a constant sample of observations for which the covariates required to estimate all the models are available. Row 1 recalculates the degree of bias – defined as the percentage difference between the cross-cohort and cross-class VA coefficients as above – on this sample for the baseline model. Rows 2 and 3 show that the degree of bias is very similar when parental controls and twice-lagged test scores are including in the control vector, consistent with the very high correlations between these VA estimates and the baseline estimates discussed above. In row 4, we include only the controls that are a function of prior-year test scores: cubic polynomials in student, classroom, and school-grade math and English scores interacted with grade level. These VA estimates remain fairly highly correlated with the baseline estimates but have a somewhat larger degree of bias (14%). Finally, row 5 estimates VA without any controls at all, i.e. using raw mean test scores by teacher. These VA estimates are very poorly correlated with the other VA measures and are biased by nearly 90%. We conclude that most of the bias in VA estimates is eliminated by controlling for lagged test scores, and that further controls for demographic variables typically available in school district datasets bring the bias close to zero.

4.5 Relationship to Prior Work

Our results on the validity of VA measures reconcile the conflicting findings of prior work, including Kane and Staiger (2008) and Rothstein (2010). Rothstein reports two important results, both of which we replicate in our data. First, there is significant grouping of students into classrooms based on twice-lagged scores (lagged gains), even conditional on once-lagged scores (Rothstein 2010, Table 4). Second, this grouping on lagged gains generates minimal bias in VA estimates: controlling for twice-lagged scores does not have a significant effect on VA estimates (Rothstein 2010, Table 6; Kane and Staiger 2008, Table 6).[52] The results from our tests in Table 2 and Figure 2 are consistent with Rothstein's conclusions. Therefore, the literature is in agreement that VA measures do not suffer from bias due to selection on observables.

Rothstein quite appropriately emphasizes that his findings raise serious concerns about the *potential* for bias due to selection on unobservable student characteristics.[53] Kane and Staiger's point estimates from a randomized experiment suggest that selection on unobservables is relatively small. Our quasi-experimental tests based on teaching staff changes confirm that the bias due to selection on unobservables turns out to be negligible with greater precision. In future work, it may be useful to explore why the grouping on lagged gains documented by Rothstein is not associated with significant selection on unobservables in practice. However, the findings in this paper and prior work are sufficient to conclude that standard estimates of teacher VA can provide accurate forecasts of teachers' average impacts on students' test scores.

Note that our test, like the experiment implemented by Kane and Staiger, evaluates the accuracy of VA measures on average across teachers. It is conceivable that VA measures are biased against some subgroups of teachers and that this bias is offset by a second source of bias which is negatively correlated with true value-added (Rothstein 2009, page 567). We focus on the accuracy of average forecasts in this paper because our analysis of long-term impacts primarily evaluates the mean impacts of teacher value-added on students. A fruitful direction for future work would be to adapt the methods we propose here to evaluate the accuracy and predictive content of VA measures for

[52] An interesting question is how Rothstein's two findings are consistent with each other. There are two explanations for this pattern. First, the degree of grouping that Rothstein finds on $A_{ig,t-2}$ has small effects on residual test score gains because the correlation between $A_{ig,t-2}$ and A_{igt} conditional on $A_{ig,t-1}$ is relatively small. Second, if the component of $A_{ig,t-2}$ on which there is grouping is not the same as the component that is correlated with $A_{i,t}$, VA estimates may be completely unaffected by grouping on $A_{i,t-2}$. For both reasons, one cannot infer from grouping on $A_{i,t-2}$ that VA estimates are significantly biased by selection on $A_{i,t-2}$. See Goldhaber and Chaplin (2012) for further discussion of these and related issues.

[53] To be clear, this was the original lesson from Rothstein (2010). In personal correspondence, Rothstein notes that his findings are "neither necessary nor sufficient for there to be bias in a VA estimate" and that "if the selection is just on observables, the bias is too small to matter. The worrying scenario is selection on unobservables."

subgroups of the population.

5 Impacts of Value-Added on Outcomes in Adulthood

The results in the previous section show that value-added is a good proxy for a teacher's ability to raise students' test scores. In this section, we analyze whether value-added is also a good proxy for teachers' long run impacts. We do so by regressing outcomes in adulthood Y_i on teacher quality $\hat{\mu}_{j(i,g)}$ and observable characteristics, as in (9). We begin by pooling the data across all grade levels and then present results that estimate grade-specific coefficients on teacher VA. Recall that each student appears in our dataset once for every subject-year with the same level of Y_i but different values of $\mu_{j(i,g)}$. Hence, in this pooled regression, the coefficient estimate β represents the mean impact of having a higher VA teacher for a *single* grade between grades 4-8. We account for the repeated student-level observations by clustering standard errors at the school-cohort level as above.

We first report estimates based on comparisons of students assigned to different teachers, which identifies the causal impact of teachers under Assumption 2. We then evaluate this identification assumption by comparing these estimates to those obtained from the teacher switcher research design, which isolates quasi-experimental variation in teacher VA. We analyze impacts of teacher VA on three sets of outcomes: college attendance, earnings, and other indicators such as teenage birth rates.

5.1 College Attendance

We begin by analyzing the impact of teacher VA on college attendance at age 20, the age at which college attendance rates are maximized in our sample. In all figures and tables in this section, we condition on the standard classroom-level controls as in Figure 1.

Figure 5a plots college attendance rates at age 20 against teacher VA. Being assigned to a higher VA teacher in a single grade raises a student's probability of attending college significantly. The null hypothesis that teacher VA has no effect on college attendance is rejected with a t-statistic above 7 ($p < 0.001$). To interpret the magnitude of the impact, recall that a 1 SD increase in teacher VA raises students' test scores by 0.1 SD on average across math and English. Because we measure teacher quality μ_j in units of student test scores, a 1 unit increase in μ_j corresponds to a 10 SD increase in teacher VA. Hence, dividing the regression coefficients β by 10 yields a rough estimate of the impact of a 1 SD increase in teacher VA on the outcome of interest. In the

case of college attendance, $\beta = 4.92\%$, implying that a 1 SD better teacher in a single grade raises the probability of being in college by 0.49% at age 20, relative to a mean of 37.8%. This impact of a 1.25% increase in college attendance rates for a 1 SD better teacher is roughly similar to the impacts on other outcomes we document below.

To confirm that the relationship in Figure 5a reflects the causal impact of teachers rather than selection bias, we implement tests analogous to those in the previous section in Table 5. As a reference, the first column replicates the OLS regression estimate reported in Figure 5a. In column 2, we replace actual college attendance with predicted attendance based on parent characteristics, constructed in the same way as predicted scores above. The estimates show that one would not have predicted any significant difference in college attendance rates across students with high vs. low VA teachers based on parent characteristics.

To account for potential bias due to unobservables, we exploit quasi-experimental variation from changes in teaching staff as above. Column 3 regresses changes in mean college attendance rates across adjacent cohorts within a school-grade-subject cell on the change in mean teacher VA due to teacher staff changes $\Delta\overline{\mu}_{sgmt}$, defined as in Table 4. As above, we include no controls other than year effects. Students who happen to be in a cohort in their school that is taught by higher VA teachers are significantly more likely to go to college. The estimate of $\beta = 6.1\%$ from this quasi-experimental variation is similar to that obtained from the cross-classroom comparison in column 1, though less precise because it exploits much less variation. The null hypothesis that $\beta = 0$ is rejected with $p < 0.01$, while the hypothesis that β is the same in columns 1 and 3 is not rejected. This finding provides further evidence that teacher VA has a causal impact on college attendance rates and confirms that comparisons across classrooms with high and low VA teachers yield consistent estimates of teachers' impacts.[54]

Next, we analyze whether high-VA teachers also improve the quality of colleges that their students attend. We quantify college quality using the age 30 earnings of students who previously attended the same college, as described in Section 3. Students who do not attend college are assigned the mean earnings of individuals who do not attend college. Figure 5b plots this earnings-based index of college quality (based on the colleges students attend at age 20) vs. teacher VA. Again, there is a highly significant relationship between the quality of colleges students attend and

[54] This result rules out bias due to omitted variables that affect long-term outcomes but not test scores. For instance, one may be concerned that students who are assigned to better teachers in one subject are also assigned to better teachers in other subjects or better extracurricular activities, which would inflate estimates of long-term impacts. The cross-cohort research design rules out such biases because fluctuations in teaching staff are highly subject-specific and are uncorrelated with other determinants of student outcomes, as shown in Figure 4d.

the quality of the teachers they had in grades 4-8 ($t = 9.5$, $p < 0.001$). A 1 SD improvement in teacher VA (i.e., an increase of 0.1 in μ_j) raises college quality by \$164 (0.66%) on average (Column 4 of Table 5). Column 5 shows that exploiting the cross-cohort teacher switcher variation again yields similar estimates of the impact of teacher VA on college quality.

The \$164 estimate combines intensive and extensive margin responses because it includes the effect of increased college attendance rates on projected earnings. Isolating intensive margin responses is more complicated because of selection bias: students who are induced to go to college by a high-VA teacher will tend to attend lower-quality colleges, pulling down mean earnings conditional on attendance. We take two approaches to overcome this selection problem and identify intensive-margin effects. First, we define an indicator for "high quality" colleges as those with average earnings above the median among colleges that students attend in our sample, which is \$39,972. We regress this indicator on teacher VA in the full sample, including students who do not attend college. Column 6 of Table 5 shows that high-VA teachers increase the probability that students attend high quality colleges. A 1 SD increase in teacher VA raises the probability of attending a high quality college by 0.36%, relative to a mean of 17%. This increase is most consistent with an intensive margin effect, as students would be unlikely to jump from not going to college at all to attending a high quality college. Second, we derive a lower bound on the intensive margin effect by assuming that those who are induced to attend college attend a college of average quality. The mean college quality conditional on attending college is \$38,623, while the quality for all those who do not attend college is \$16,361. Hence, at most $(38,623 - 16,361) \times 0.49\% = \$109$ of the \$164 impact is due to the extensive margin response, confirming that teachers improve the quality of colleges that students attend.

Figure 5c shows the impact of teachers on college attendance at other ages. Teacher VA has a significant impact on the college attendance rate through age 25, partly reflecting attendance of graduate or professional schools. The impacts on college attendance at age 25 are smaller in magnitude (0.28% per 1 SD of teacher VA) than at age 20 because the mean college attendance rate at age 25 is 18.1% in this sample (Column 7 of Table 5). These continued impacts on college attendance in the mid 20's affect our analysis of earnings impacts, to which we now turn.

5.2 Earnings

The correlation between annual earnings and lifetime income rises rapidly as individuals enter the labor market and begins to stabilize only in the late twenties. We therefore begin by analyzing the

impacts of teacher VA on earnings at age 28, the oldest age at which we have a sufficiently large sample of students to obtain precise estimates.[55] Figure 6 plots earnings at age 28 against teacher VA, conditioning on the same set of classroom-level controls as above. Being assigned to a higher value-added teacher has a clear, statistically significant impact on earnings, with the null hypothesis of $\beta = 0$ rejected with $p < 0.01$. A 1 SD increase in teacher VA in a single grade increases earnings at age 28 by \$182, 0.9% of mean earnings in the regression sample. This regression estimate is also reported in Column 1 of Table 6. Column 2 shows the effect on wages at age 30. The point estimate is slightly larger than that at age 28, but because the sample is only one-sixth the size, the 95% confidence interval for the estimate is very wide. We therefore focus on earnings impacts up to age 28 for the remainder of our analysis.

To interpret the magnitude of the effect of teacher VA on earnings at age 28, we calculate the lifetime earnings impact of having a 1 SD higher VA teacher in a single grade. We assume that the percentage gain in earnings remains constant at 0.9% over the life-cycle and that earnings are discounted at a 3% real rate (i.e., a 5% discount rate with 2% wage growth) back to age 12, the mean age in our sample. Under these assumptions, the mean present value of lifetime earnings at age 12 in the U.S. population is approximately \$522,000.[56] Hence, the financial value of having a 1 SD higher VA teacher (i.e., a teacher at the 84th percentile instead of the median) is $0.9\% \times \$522,000 \simeq \$4,600$ per grade.[57] Another useful benchmark is the increase in earnings from an additional year of schooling, which is around 6% per year (see e.g., Oreopoulos 2006). Having a teacher in the first percentile of the value-added distribution (2.33 SD below the mean) for one year thus has an earnings impact equivalent to attending school for about 60% of the school year. This magnitude is plausible, insofar as attending school even with very low quality teaching is likely to have some returns due to benefits from peer interaction and other factors.

Next, we analyze how teacher value-added affects the trajectory of earnings by examining earnings impacts at each age from 20 to 28. We run separate regressions of earnings at each age on teacher VA and the standard vector of classroom controls. Figure 7a plots the coefficients from these regressions (which are reported in Appendix Table 10), divided by average earnings at each

[55] Although individuals' earnings trajectories remain quite steep at age 28, earnings levels at age 28 are highly correlated with earnings at later ages (Haider and Solon 2006), a finding we confirm in the tax data (Chetty et al. 2011, Appendix Table I).

[56] We calculate this number using the mean wage earnings of a random sample of the U.S. population in 2007 to obtain an earnings profile over the lifecycle, and then inflate these values to 2010 dollars (see Chetty et al. 2011 for details).

[57] The undiscounted earnings gains (assuming a 2% growth rate but 0% discount rate) are approximately \$25,000 per student.

age to obtain percentage impacts. As above, we multiply the estimates by 0.1 to interpret the effects as the impact of a 1 SD increase in teacher VA. The impact of teacher quality on earnings rises almost monotonically with age. At early ages, the impact of higher VA is *negative* and significant, which is consistent with our finding that higher VA teachers induce their students to go to college. As these students enter the labor force, they have steeper earnings trajectories and eventually earn significantly more than students who had lower VA teachers in grades 4-8. The earnings impacts become positive and statistically significant starting at age 26. By age 28, the earnings impact is nearly 1% of earnings, as in Figure 7. Stated differently, higher teacher VA increases the growth rate of earnings when students are in their 20s. In column 3 of Table 6, we verify this result by regressing the change in earnings from age 22 to age 28 on teacher VA. As expected, a 1 SD increase in teacher VA increases earnings growth by $180 (1.3%) over this period.

We obtain further insight into the role of college in mediating these changes in earnings trajectories by comparing the impacts of teacher VA on students who attend grade schools with low vs. high college attendance rates. We divide the sample into two groups: students who attend schools with an age 20 college attendance rate above vs. below 35%, the sample mean. In schools with low college attendance rates at age 20, few students are in college at age 25. As a result, teacher VA does not have a significant impact on college attendance rates at age 25 for students in these schools, as shown in Column 4 of Table 6. In contrast, in schools with high college attendance rates, a 1 SD increase in teacher VA raises college attendance rates by 0.47 percentage points even at age 25. If college attendance masks earnings impacts, we should expect the effects of teacher VA on wage growth to be higher in these high college attendance schools.

Figure 7b tests this hypothesis by plotting the effect of value-added on earnings by age for students who attended schools with above- and below- average college attendance rates. As expected, the impacts of teacher VA on earnings rise much more sharply with age for students who attended grade schools with high college attendance rates. Teacher VA has a negative impact on earnings in the early 20's for students who attended such schools, whereas its impacts are always positive for students who attended schools with low college attendance rates. The positive impacts of teacher VA on earnings even in subgroups that are unlikely to attend college indicates that better teaching has direct returns in the labor market independent of its effects on college attendance. Columns 6 and 7 of Table 6 confirm that the effect of teacher VA on wage growth from age 22 to 28 is much larger for students who attended schools with high college attendance rates.

The results in Figure 7 suggest that the 0.9% mean earnings impact per SD of teacher VA

at age 28 may understate the impact on lifetime earnings, particularly for high SES groups. To gauge how much further the earnings impacts might rise over time, we use the cross-sectional correlation between test scores and earnings, which we can estimate with greater precision up to age 30. Appendix Table 4 lists coefficients from OLS regressions of earnings at each age on test scores. These regressions pool all grades, control for the same variables used to estimate the baseline value-added model, and use a constant sample of students for whom we observe earnings from 20-30 to eliminate cohort effects. The correlation between test scores and earnings is roughly 20% higher at age 30 than at age 28. If the causal impacts of teacher VA match these cross-sectional patterns by age, the lifetime earnings impact of a 1 SD improvement in teacher VA in a single grade would likely exceed 1.1%.

The cross-sectional relationship between test scores and earnings reported in Appendix Table 4 implies that a 0.1 SD increase in test scores is associated with a 1.1% increase in earnings at age 28. Hence, the impact of teacher VA is similar to the impact one would have predicted based on the impact of VA on end-of-grade test scores and the cross-sectional relationship between test scores and earnings. This result aligns with previous evidence that improvements in education raise contemporaneous scores, then fade out in later scores (as shown in Figure 2), only to reemerge in adulthood (Deming 2009, Heckman et al. 2010c, Chetty et al. 2011).

5.3 Other Outcomes

We now analyze the impacts of teacher VA on other outcomes, starting with our "teenage birth" measure, which is an indicator for filing a tax return and claiming a dependent who was born while the mother was a teenager (see Section 3). We first evaluate the cross-sectional correlations between this proxy for teenage birth and test scores as a benchmark. Students with a 1 SD higher test score are 3.8 percentage points less likely to have a teenage birth relative to a mean of 8% (Appendix Table 3). Conditional on lagged test scores and other controls, a 1 SD increase in test score is associated with a 1 percentage point reduction in teenage birth rates. These correlations are significantly larger for populations that have a higher risk of teenage birth, such as minorities and low-income students (Appendix Table 5). These cross-sectional patterns support the use of this measure as a proxy for teenage births even though we can only identify children who are claimed as dependents in the tax data.

Column 1 of Table 7 analyzes the impact of teacher VA on the fraction of female students who have a teenage birth. Having a 1 SD higher VA teacher in a single year from grades 4 to 8 reduces

the probability of a teen birth by 0.099 percentage points, a reduction of roughly 1.25%, as shown in Figure 8a. This impact is very similar to the cross-sectional correlation between scores and teenage births, echoing our results on earnings and college attendance.

Column 2 of Table 7 analyzes the impact of teacher VA on the socio-economic status of the neighborhood in which students live at age 25, measured by the percent of college graduates living in that neighborhood. A 1 SD increase in teacher VA raises neighborhood SES by 0.063 percentage points (0.5% of the mean) by this metric, as shown in Figure 8b. Column 3 shows that this impact on neighborhood quality more than doubles at age 28, consistent with the growing earnings impacts documented above.

Finally, we analyze impacts on retirement savings. Teacher VA does not have a significant impact on 401(k) savings at age 25 in the pooled sample (not reported). However, Column 4 shows that for students who attended schools with low college attendance rates (defined as in Column 4 of Table 6), a 1 SD increase in teacher VA raises the probability of having a 401(k) at age 25 by 0.19 percentage points (1.6% of the mean). In contrast, Column 5 shows that for students in high college-attendance schools, the point estimate of the impact is negative. These results are consistent with the impacts on earnings trajectories documented above. In schools with low college attendance rates, students who get high-VA teachers find better jobs by age 25 and are more likely to start saving in 401(k)'s. In schools with high college attendance rates, students who get high-VA teachers are more likely to be in college at age 25 and thus may not obtain a job in which they begin saving for retirement until they are older.

5.4 Heterogeneity Analysis

In Table 8, we analyze whether teacher value-added has heterogeneous effects across demographic groups and subjects. We study impacts on college quality at age 20 rather than earnings because the heterogeneity analysis requires large samples and because the college quality measure provides a quantitative metric based on projected earnings gains.

Panel A studies impact heterogeneity across population subgroups. Each number in the first row of the table is a coefficient estimate from a separate regression of college quality on teacher VA, with the same classroom-level controls as in the previous sections. Columns 1 and 2 consider heterogeneity by gender. Columns 3 and 4 consider heterogeneity by parental income, dividing students into groups above and below the median level of parent income in the sample. Columns 5 and 6 split the sample into minority and non-minority students.

Two lessons emerge from Panel A of Table 8. First, the point estimates of the impacts of teacher VA are larger for girls than boys, although one can reject equality of the impacts only at a 10% significance level. Second, the impacts are larger for higher-income and non-minority households in absolute terms. For instance, a 1 SD increase in VA raises college quality by $123 for children whose parents have below-median income, compared with $209 for those whose parents have above-median income. However, the impacts are much more similar as a percentage of mean college quality: 0.56% for low-income students vs. 0.75% for high-income students.

The larger dollar impact for high socioeconomic students could be driven by two channels: a given increase in teacher VA could have larger impacts on the test scores of high SES students or a given increase in scores could have larger long-term impacts. The second row of coefficient estimates of Table 8 shows that the impacts of teacher VA on scores are virtually identical across all the subgroups in the data. In contrast, the correlation between scores and college quality is significantly larger for higher SES students (Appendix Table 5). Although not conclusive, these findings suggest that the heterogeneity in teachers' long term impacts is driven by the second mechanism, namely that high SES students benefit more from test score gains. Overall, the heterogeneity in treatment effects indicates that teacher quality is complementary to family inputs and resources, i.e. the marginal value of better teaching is *larger* for students from high SES families. An interesting implication of this result is that higher income families should be willing to pay more for teacher quality.

Panel B of Table 8 analyzes differences in teachers' impacts across subjects. For these regressions, we split the sample into elementary (Columns 1-3) and middle (Columns 4-6) schools. We first analyze the effects of teacher VA in each subject separately on a constant sample with a fixed set of controls and then include both math and English teacher VA in the same regression. In all the specifications, the coefficients on VA are larger in English than math. An English teacher who raises her students' test scores by 1 SD has a larger long-term impact than a math teacher who generates a commensurate test score gain. However, it is important to recall that the variance of teacher effects is larger in math than English: a 1 SD improvement in teacher VA raises math test scores by approximately 0.118 SD, compared with 0.081 SD in English. Hence, a 1 SD increase in the quality of a math teacher actually has a relatively similar impact to a 1 SD increase in the quality of an English teacher.

Including both English and math VA in the same regression has very different effects in elementary vs. middle school. As discussed above, students have one teacher for both subjects in

elementary school but not middle school. Because a given teacher's math and English VA are highly correlated ($r = 0.59$), the magnitude of the two subject-specific coefficients drops by nearly 40% when included together in a single regression for elementary school (Column 3). Intuitively, when math VA is included by itself in elementary school, it partly picks up the effect of having better teaching in English as well. In contrast, including both math and English teacher VA in middle school has a much smaller effect on the estimates, as shown in Column 6.

5.5 Robustness Checks

We conclude our empirical analysis by assessing the robustness of our results to alternative empirical specifications, focusing on the simplifications we made for computational tractability.

First, we assess the robustness of our statistical inferences to alternative forms of clustering standard errors. Appendix Table 7 reports alternative standard error calculations for three of our main specifications: the impact of teacher VA on scores, college attendance at age 20, and earnings at age 28. We estimate each of these models using the baseline control vector used in Table 2. Panels A of Appendix Table 7 shows that a block bootstrap at the student level, which accounts for repeated student observations, yields narrower confidence intervals than school-cohort clustering. Panel B shows that in smaller subsamples of our data, two-way clustering by class and student yields slightly smaller standard errors than school-cohort clustering. Panel C shows that school-cohort clustering is also conservative relative to clustering by classroom in a sample that includes only the first observation for each student.

Second, we assess the robustness of our estimates to alternative control vectors (Panel D of Appendix Table 7). Including the student-level controls used when estimating the VA model in addition to the baseline classroom-level control vector used to estimate the regressions in Tables 2, 5, and 6 has virtually no impact on the coefficients or standard errors. The last row of the table evaluates the impacts of including school by year fixed effects. In this row, we include school by year effects both when estimating VA and in the second-stage regressions of VA on adult outcomes. The inclusion of school by year fixed effects does not affect our qualitative conclusion that teacher VA has substantial impacts on adult outcomes, but the estimated impact on college attendance at age 20 falls, while the impact on earnings at age 28 rises.[58]

Finally, we replicate the baseline results using raw estimates of teacher quality without the

[58] We did not include school-year fixed effects in our baseline specifications because school districts typically seek to rank teachers within their districts rather than within schools. Moreover, our tests in Section 4 suggest that such fixed effects are not necessary to obtain unbiased estimates of the impacts of teacher VA.

Empirical Bayes shrinkage correction, denoted by $\overline{\nu}_j$ in Section 2. We again exclude the current year when estimating $\overline{\nu}_j$ to account for correlated estimation error as above. In columns 1-4 of Appendix Table 11, we estimate specifications analogous to (9) using OLS, with a leave-year-out measure $\overline{\nu}_j^t$ on the right hand side instead of $\widehat{\mu}_j^t$. The estimated coefficients are roughly half of those reported above, reflecting the substantial attenuation from measurement error in teacher quality. The shrinkage correction implemented in our baseline measure of teacher VA is one approach to correct for this measurement error. As an alternative approach, we regress each outcome on test scores, instrumenting for scores using the raw teacher effects $\overline{\nu}_j$. The resulting two-stage least squares coefficients are reported in Columns 5-7 of Appendix Table 11. These 2SLS estimates are very similar to our baseline results, confirming that our findings are not sensitive to the way in which correct for measurement error in teacher quality.

6 Policy Calculations

In this section, we use our estimates to answer two policy questions. First, do teachers matter more in some grades relative to others? Second, what are the expected earnings gains from retaining or deselecting teachers based on their estimated VA?

6.1 Impacts of Teachers by Grade

The reduced-form estimates in the previous section identify the impacts of replacing a single teacher j with another teacher j' in one classroom. While this question is of interest to parents, policymakers are typically interested in the impacts of reforms that improve teacher quality more broadly. As shown in (5), the reduced-form impact of changing the teacher of a single classroom includes the impacts of being tracked to a better teacher in subsequent grades. While a parent may be interested in the reduced-form impact of teacher VA in grade g (β_g), a policy reform that raises teacher quality in grade g will not allow every child to get a better teacher in grade $g+1$. In this section, we estimate teachers' net impacts in each grade, holding fixed future teacher VA ($\tilde{\beta}_g$), to shed light on this policy question.

Because we have no data after grade 8, we can only estimate teachers' net effects holding fixed teacher quality up to grade 8.[59] We therefore set $\tilde{\beta}_8 = \beta_8$. We recover $\tilde{\beta}_g$ from estimates of β_g by subtracting out the impacts of future teachers on earnings iteratively. Consider the effect of

[59] If tracking to high school teachers is constant across all grades in elementary school, our approach accurately recovers the relative impacts of teachers in grades 4-8.

teacher quality in 7th grade. Our reduced-form estimate of β_7, obtained by estimating (9) using only grade 7, can be decomposed into two terms:

$$\beta_7 = \tilde{\beta}_7 + \rho_{78}\tilde{\beta}_8$$

where ρ_{78} is the extent to which teacher VA in grade 7 increases teacher VA in grade 8. We can estimate $\hat{\rho}_{78}$ using an OLS regression that parallels (9) with future teacher VA as the dependent variable:

$$\hat{\mu}_{j(i,8)} = \alpha + \hat{\rho}_{78}\hat{\mu}_{j(i,7)} + f_1(A_{i,t-1}) + f_2(e_{j(i,7,t)}) + \phi_1 X_{i7t} + \phi_2 \bar{X}_{c(i,7,t)} + \eta^{\mu}_{it78}.$$

Combining these two equations shows that the net impact of the grade 7 teacher is simply her reduced-form impact minus her indirect impact via tracking to a better 8th grade teacher:

$$\tilde{\beta}_7 = \beta_7 - \hat{\rho}_{78}\beta_8.$$

Iterating backwards, we can calculate $\tilde{\beta}_6$ by estimating $\hat{\rho}_{68}$ and $\hat{\rho}_{67}$ and so on until we obtain the full set of net impacts. We show formally that this procedure recovers net impacts $\tilde{\beta}_g$ in Appendix B.

This approach to calculating teachers' net impacts has three important limitations. First, it assumes that all tracking to future teachers occurs via teacher VA on test scores. We allow students who have high-VA teachers in grade g to be tracked to higher VA ($\mu_{j(i,g+1)}$) teachers in grade $g+1$, but *not* to teachers with higher unobserved earnings impacts μ^Y. We are forced to make this strong assumption because we have no way to estimate teacher impacts on earnings that are orthogonal to VA, as discussed in Section 2. Second, $\tilde{\beta}_g$ does not net out potential changes in other factors besides teachers, such as peer quality or parental inputs. Hence, $\tilde{\beta}_g$ cannot be interpreted as the "structural" impact of teacher quality holding fixed all other inputs in a general model of the education production function (e.g., Todd and Wolpin 2003). Finally, our approach assumes that teacher effects are additive across grades. We cannot identify complementarities in teacher VA across grades because our identification strategy forces us to condition on lagged test scores, which are endogenous to the prior teacher's quality. It would be valuable to relax these assumptions in future work to obtain a better understanding of how the sequence of teachers one has in school affects outcomes in adulthood.

Figure 9 displays our estimates of β_g and $\tilde{\beta}_g$, which are also reported in Appendix Table 12. We use college quality (projected earnings at age 30 based on college enrollment at age 20) as

the outcome to have sufficient precision to identify grade-specific effects. We estimate β_g using specifications analogous to Column 4 of Table 5 for each grade separately. Because the school district data system did not cover many middle schools in the early and mid 1990s, we cannot analyze the impacts of teachers in grades 6-8 for more than half the students who are in 4th grade before 1994. To obtain a more balanced sample for comparisons across grades, we restrict attention to cohorts who would have been in 4th grade during or after 1994 for this analysis.

Figure 9 has two lessons. First, the net impacts $\tilde{\beta}_g$ are close to the reduced-form impacts. This is because the tracking coefficients $\rho_{g,g'}$ are generally quite small, as shown in Appendix Table 13. Tracking is slightly larger in middle school, as one would expect, but still has a relatively small impact on $\tilde{\beta}_g$. Second, teachers' long-term impacts are large and significant in all grades. Although the estimates in each grade have relatively wide confidence intervals, there is no systematic trend in the impacts. This pattern is consistent with the cross-sectional correlations between test scores and adult outcomes, which are also relatively stable across grades (Appendix Table 6).

One issue that complicates cross-grade comparisons is that teachers spend almost the entire school day with their students in elementary school (grades 4-5 as well as 6 in some schools), but only their subject period (Math or English) in middle school (grades 7-8). If teachers' skills are correlated across subjects – as is the case with math and English value-added, which have a correlation of 0.59 for elementary school teachers – then a high-VA teacher should have a greater impact on earnings in elementary school than middle school because they spend more time with the student. The fact that high-VA math and English teachers continue to have substantial impacts even in middle school underscores our conclusion that higher quality education has substantial returns well beyond early childhood.

6.2 Impacts of Selecting Teachers on VA

In this section, we use our estimates to predict the potential earnings gains from selecting and retaining teachers on the basis of their VA. The primary objective of these calculations is to illustrate the magnitudes of teachers' impacts rather than evaluate selection as a policy to improve teacher quality.

We make three assumptions in our calculations. First, we assume that the percentage impact of a 1 unit improvement in teacher VA on earnings observed at age 28, which we denote by b, remains constant over the life-cycle. Second, we do not account for general equilibrium effects that may reduce wages if all children are better educated or for non-monetary returns to education such

as reductions in teenage birth rates (Oreopoulos and Salvanes 2010). Third, we follow Krueger (1999) and discount earnings gains at a 3% real annual rate (consistent with a 5% discount rate and 2% wage growth) back to age 12, the average age in our sample. Under this assumption, the present value of earnings at age 12 for the average individual in the U.S. population is $\$522,000$, as noted above.

We first evaluate Hanushek's (2009, 2011) proposal to replace the 5 percent of teachers with the lowest value-added with teachers of average quality. To calculate the impacts of such a policy, note that a teacher in the bottom 5% of the true VA distribution is on average 2.04 standard deviations below the mean teacher quality. Therefore, replacing a teacher in the bottom 5% with an average teacher generates a gain per student of

$$\$522,000 \times 2.04 \times b\sigma_\mu$$

where σ_μ denotes the standard deviation of teacher effects. We set $b = \$1,815/20,362 = 8.9\%$ based on the estimate in Column 1 of Table 6 and $\sigma_\mu = (0.081 + 0.118)/2$, the average of the SD of teacher effects across math and English. With these values, replacing a teacher in the bottom 5% with an average teacher generates earnings gains of $\$9,422$ per student in present value at age 12, or $\$267,000$ for a class of average size (28.3 students). The undiscounted cumulative earnings gains from deselection are 5.5 times larger than these present value gains ($52,000 per student and $1.48 million per classroom), as shown in Appendix Table 14.[60] These calculations show that improving teacher VA – whether by selection, better training, or other methods – is likely to have substantial returns for students.

The $\$267,000$ present value gain is based on selecting teachers based on their *true* VA μ_j. In practice, we only observe a noisy estimate of μ_j based on a small number of classrooms. To calculate the gains from deselecting the bottom 5% of teachers based on their *estimated* VA, note that (4) implies that $\sigma_{\hat{\mu}} = \sigma_\mu \sqrt{r(n_c)}$ where $r(n_c)$ is the reliability of VA estimates based on n_c classrooms of data. Hence, with n_c years of data, the bottom 5 percent of teachers ranked on $\hat{\mu}_j$ have a mean forecasted quality of $2.04\sigma_\mu\sqrt{r(n_c)}$. The gain from deselecting the lowest 5% of teachers based on n_c classrooms of data is thus $G(n_c) = \$267,000 \cdot \sqrt{r(n_c)}$.[61]

[60] These calculations assume that deselected teachers are replaced by teachers with the same amount of experience rather than rookies. Rookie teachers' test score impacts are 0.03 SD below those of experienced teachers, on average. However, given that the median teacher remains in our data for 6 years, the expected benefits of deselection would be reduced by less than 3% ($\frac{0.03/6}{2.04\sigma_\mu}$) from hiring inexperienced teachers to replace those deselected.

[61] This calculation accounts for estimation error due to noise but ignores drift in VA over time (except for drift due to teacher experience, which we control for in our analysis). Drift affects the calculation in two ways. First, our

Figure 10 plots $G(n_c)$ assuming a constant class size of 28.3 students; see Appendix Table 14 for the values underlying this figure. It yields three lessons. First, the gains from deselecting low quality teachers on the basis of very few years of data are much smaller than the maximum attainable gain of $267,000 because of the noise in VA estimates. With one year of data, the gains are about half as large ($135,000). This is because reliability with one class of students is approximately $r(1) = \frac{1}{4}$ in our data, consistent with prior work on teacher effects (Staiger and Rockoff 2010, McCaffrey et al. 2009). That is, one-quarter of the variance in the mean test score residual for a single classroom is driven by teacher quality, with the remaining variance due to classroom and student level noise. Second, the gains grow fairly rapidly with more data in the first 3 years but the marginal gains from additional information are small. With three years of data, one can achieve more than 70% of the maximum impact ($190,000). Waiting for three more years would increase the gain by $30,000 but has an expected cost of $3 \times \$190,000 = \$570,000$. The marginal gains from obtaining one more year of data are outweighed by the expected cost of having a low VA teacher on the staff even after the first year (Staiger and Rockoff 2010). Third, because VA estimates are noisy, there could be substantial gains from using other signals of quality to complement VA estimates, such as principal evaluations or other subjective measures based on classroom observation.

An alternative approach to improving teacher quality is to increase the retention of high-VA teachers. Retaining a teacher at the 95th percentile of the estimated VA distribution (using 3 classrooms of data) for an extra year would yield present value earnings gains of $\$522,000 \times 1.96 \times b\sigma_\mu\sqrt{r(3)} = \$182,000$. In our data, roughly 9% of teachers in their third year do not return to the school district for a fourth year.[62] Clotfelter et al. (2008) estimate that a $1,800 bonus payment in North Carolina reduces attrition rates by 17%. Based on this estimate, a one time bonus payment of $1,800 to high-VA teachers who return for a fourth year would increase retention rates in the next year by 1.5 percentage points and generate an average benefit of $2,730. The expected benefit of offering a bonus to even an excellent (95th percentile) teacher is only modestly larger than the cost because for every extra teacher retained, one must pay bonuses to 60 (91/1.5) additional teachers.

estimate of b uses estimated VA from other years and thereby understates the impact of a 1 unit increase in true VA on earnings. This leads us to understate the $267,000 gain. Second, if true VA is mean reverting, deselecting teachers based on their current VA will yield smaller gains in subsequent years, because some of the low VA teachers improve over time. An interesting direction for future research is to estimate the process that VA follows and then identify the expected gains from selecting teachers based on their true VA over various horizons.

[62] The rate of attrition bears little or no relation to VA, consistent with the findings of Boyd et al. (2009).

One important caveat to these calculations is that they assume that teacher effectiveness μ_j does not vary with classroom characteristics. Our estimates of VA only identify the component of teacher quality that is orthogonal to lagged test scores and the other characteristics that we control for to account for sorting. That is, teachers are evaluated relative to the average quality of teachers with similar students, not relative to the population. Thus, while we can predict the effects of selecting teachers among those assigned to a sub-population of similar students, we cannot predict the impacts of policies that reassign teachers to randomly selected classrooms from the population (Rubin, Stuart, and Zanutto 2004). This is a limitation in all existing value-added measures of teacher quality and could have significant implications for their use if teaching quality interacts heavily with student attributes. Lockwood and McCaffrey (2009) argue that such interactions are small relative to the overall variation in teacher VA. In addition, our estimates based on teaching staff changes suggest that VA is relatively stable as teachers switch to different grades or schools. Nevertheless, further work is needed on this issue if a policymaker is considering reassigning teachers across classrooms and seeks a global ranking of their relative quality.

7 Conclusion

This paper has presented evidence that existing value-added measures are informative about teachers' long-term impacts. However, two important issues must be resolved before one can determine whether VA should be used to evaluate teachers. First, using VA measures in high-stakes evaluations could induce responses such as teaching to the test or cheating, eroding the signal in VA measures. This question can be addressed by testing whether VA measures from a high stakes testing environment provide as good of a proxy for long-term impacts as they do in our data.[63] If not, one may need to develop metrics that are more robust to such responses, as in Barlevy and Neal (2012). Districts may also be able to use data on the persistence of test score gains to identify test manipulation, as in Jacob and Levitt (2003), and thereby develop a more robust estimate of VA. Second, one must weigh the cost of errors in personnel decisions against the mean benefits from improving teacher value-added. We quantified mean earnings gains from selecting teachers on VA but did not quantify the costs imposed on teachers or schools from the turnover generated by such policies.

[63] As we noted above, even in the low-stakes regime we study, some teachers in the upper tail of the VA distribution have test score impacts consistent with test manipulation. If such behavior becomes more prevalent when VA is actually used to evaluate teachers, the predictive content of VA as a measure of true teacher quality could be compromised.

Whether or not VA should be used as a policy tool, our results suggest that parents would place great value on having their child in the classroom of a high value-added teacher. Consider a teacher whose true VA is 1 SD above the median who is contemplating leaving a school. Each child would gain approximately $25,000 in total (undiscounted) lifetime earnings from having this teacher instead of the median teacher. With an annual discount rate of 5%, the parents of a classroom of average size should be willing to pool resources and pay this teacher approximately $130,000 ($4,600 per parent) to stay and teach their children during the next school year. Our analysis of teacher entry and exit directly confirms that retaining such a high-VA teacher would improve students' outcomes.

While these calculations show that good teachers have great value, they do not by themselves have implications for optimal teacher salaries or merit pay policies. The most important lesson of this study is that finding policies to raise the quality of teaching – whether via the use of value-added measures, changes in salary structure, or teacher training – is likely to have substantial economic and social benefits in the long run.

References

1. Aaronson, Daniel, Lisa Barrow, and William Sander. 2007. "Teachers and Student Achievement in Chicago Public High Schools." *Journal of Labor Economics* 24(1): 95-135.

2. Altonji, Joseph, Todd Elder, and Christopher Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1): 151-184.

3. Baker, Eva L., Paul E. Barton, Linda Darling-Hammond, Edward Haertel, Helen F. Ladd, Robert L. Linn, Diane Ravitch, Richard Rothstein, Richard J. Shavelson, and Lorrie A. Shepard. 2010. "Problems with the Use of Student Test Scores to Evaluate Teachers." Economic Policy Institute Briefing Paper #278.

4. Barlevy, Gadi and Derek Neal. 2012. "Pay for Percentile." *American Economic Review* (forthcoming).

5. Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff "Who Leaves? Teacher Attrition and Student Achievement." *Economics of Education Review* (forthcoming).

6. Cameron, Colin A., Jonah B. Gelbach, and Douglas Miller. 2011. "Robust Inference with Multi-way Clustering," *Journal of Business and Economic Statistics* 29 (2): 238-249.

7. Carrell, Scott E. and James E. West. 2010. "Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors," *Journal of Political Economy* 118(3): 409-432.

8. Chetty, Raj, John N. Friedman, Nathaniel Hilger, Emmanuel Saez, Diane Schanzenbach, and Danny Yagan. 2011. "How Does Your Kindergarten Classroom Affect Your Earnings? Evidence from Project STAR" *Quarterly Journal of Economics* 126(4): 1593-1660, 2011.

9. Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. 2011. "New Evidence on the Long-Term Impacts of Tax Credits." IRS Statistics of Income White Paper.

10. Clotfelter, Charles, Elizabeth Glennie, Helen Ladd, and Jacob Vigdor. 2008. "Would higher salaries keep teachers in high-poverty schools? Evidence from a policy intervention in North Carolina." *Journal of Public Economics* 92: 1352-70.

11. Corcoran, Sean P. 2010. "Can Teachers be Evaluated by Their Students' Test Scores? Should they Be? The Use of Value-Added Measures of Teacher Effectiveness in Policy and Practice." Report for the Annenberg Institute for School Reform, Education Policy for Action Series.

12. Cunha, Flavio and James J. Heckman. 2010. "Investing in our Young People." NBER Working Paper 16201.

13. Cunha, Flavio, James J. Heckman, and Susanne M. Schennach. 2010. "Estimating the Technology of Cognitive and Noncognitive Skill Formation." *Econometrica* 78(3): 883–931.

14. Deming, David. 2009. "Early Childhood Intervention and Life-Cycle Development: Evidence from Head Start." *American Economic Journal: Applied Economics* 1(3): 111-134.

15. Dynarski, Susan, Joshua M. Hyman, and Diane Whitmore Schanzenbach. 2011. "Experimental Evidence on the Effect of Childhood Investments on Postsecondary Attainment and Degree Completion." NBER Working Paper 17533.

16. Goldhaber, Dan and Duncan Chaplin, 2012. "Assessing the 'Rothstein Test': Does It Really Show Teacher Value-Added Models Are Biased?" University of Washington Working Paper.

17. Goldhaber, Dan and Michael Hansen. 2010. "Using Performance on the Job to Inform Teacher Tenure Decisions." *American Economic Review* 100(2): 250-255.

18. Gordon, Robert, Thomas J. Kane, and Douglas O. Staiger. 2006. "Identifying Effective Teachers Using Performance on the Job," The Hamilton Project White Paper 2006-01.

19. Haider, Steven, and Gary Solon. 2006. "Life-cycle variation in the Association Between Current and Lifetime Earnings." *American Economic Review* 96: 1308-1320.

20. Hanushek, Eric A. 1971. "Teacher Characteristics and Gains in Student Achievement: Estimation Using Micro Data." *American Economic Review Papers and Proceedings* 61(2): 280-88.

21. Hanushek, Eric A. 2009. "Teacher Deselection." in Creating a New Teaching Profession, ed. Dan Goldhaber and Jane Hannaway, 165–80. Washington, DC: Urban Institute Press.

22. Hanushek, Eric A. 2011. "The Economic Value of Higher Teacher Quality." *Economics of Education Review* 30: 466–479.

23. Hanushek Eric A., John F. Kain and Steven G. Rivkin. 2004. "Why Public Schools Lose Teachers," *Journal of Human Resources* 39(2): 326-354

24. Heckman, James J. 2002. "Policies to Foster Human Capital." *Research in Economics* 54(1): 3-56.

25. Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam. Yavitz. 2010a. "Analyzing Social Experiments as Implemented: A Reexamination of the Evidence from the HighScope Perry Preschool Program." *Quantitative Economics* 1(1): 1-46.

26. Heckman, James J., Seong H. Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz. 2010b. "The Rate of the Return to the High Scope Perry Preschool Program." *Journal of Public Economics* 94: 114-128.

27. Heckman, James J., Lena Malofeeva, Rodrigo Pinto, and Peter A. Savelyev. 2010c. "Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes," University of Chicago, unpublished.

28. Holmstrom, Bengt and Paul Milgrom. 1991. "Multitask Principal-Agent Analyses: Incentive Contracts, Asset Ownership, and Job Design," *Journal of Law, Economics, and Organizations* 7: 24–52.

29. Internal Revenue Service. 2010. *Document 6961: Calendar Year Projections of Information and Withholding Documents for the United States and IRS Campuses 2010-2018,* IRS Office of Research, Analysis, and Statistics, Washington, D.C.

30. Jackson, C. Kirabo. 2010. "Match Quality, Worker Productivity, and Worker Mobility: Direct Evidence From Teachers." NBER Working Paper No. 15990.

31. Jackson, C. Kirabo, and Elias Bruegmann. 2009. "Teaching Students and Teaching Each Other: The Importance of Peer Learning for Teachers," *American Economic Journal: Applied Economics* 1(4): 85–108.

32. Jacob, Brian A. 2005. "Accountability, Incentives and Behavior: The Impact of High-Stakes Testing in the Chicago Public Schools." *Journal of Public Economics* 89(6): 761–796.

33. Jacob, Brian A. and Steven D. Levitt. 2003. "Rotten Apples: An Investigation Of The Prevalence And Predictors Of Teacher Cheating." *The Quarterly Journal of Economics* 118(3): 843-877.

34. Jacob, Brian A., Lars Lefgren, and David P. Sims. 2010. "The Persistence of Teacher-Induced Learning Gains," Journal of Human Resources, 45(4): 915-943.

35. Jacob, Brian A. and Jonah E. Rockoff. 2011. "Organizing Schools to Improve Student Achievement: Start Times, Grade Configurations, And Teaching Assignments" Hamilton Project Discussion Paper 2011-08.

36. Kane, Thomas J., and Douglas O. Staiger. 2008. "Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation," NBER Working Paper No. 14607.

37. Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger. 2008. "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City," *Economics of Education Review* 27: 615–631

38. Kane, Thomas J., Eric S. Taylor, John H. Tyler, and Amy L. Wooten. 2011. "Identifying Effective Classroom Practices Using Student Achievement Data." *Journal of Human Resources* 46(3): 587-613.

39. Krueger, Alan B. 1999. "Experimental Estimates of Education Production Functions." *Quarterly Journal of Economics* 114(2): 497-532.

40. Lockwood, J.R. and Daniel F. McCaffrey. 2009. "Exploring Student-Teacher Interactions in Longitudinal Achievement Data," *Education Finance and Policy* 4(4): 439-467.

41. McCaffrey, Daniel F., Tim R. Sass, J.R. Lockwood, and Kata Mihaly. 2009. "The Intertemporal Variability of Teacher Effect Estimates," *Education Finance and Policy* 4(4): 572-606.

42. Morris, Carl. 1983. "Parametric Empirical Bayes Inference: Theory and Applications" *Journal of the American Statistical Association* 78: 47-55.

43. Murnane, Richard J. 1975. *The Impact of School Resources on the Learning of Inner City Children.* Cambridge, MA: Ballinger.

44. Neal, Derek A. and Diane Whitmore Schanzenbach. 2010. "Left Behind by Design: Proficiency Counts and Test-Based Accountability," *Review of Economics and Statistics* 92(2): 263-283.

45. Oreopoulos, Philip. 2006. "Estimating Average and Local Average Treatment Effects of Education when Compulsory School Laws Really Matter." American Economic Review 96(1): 152-175.

46. Oreopoulos, Philip, and Kjell G. Salvanes. 2010. "Priceless: The Nonpecuniary Benefits of Schooling." *Journal of Economic Perspectives* 25(1): 159–84.

47. Rivkin, Steven. G., Eric. A. Hanushek, and John F. Kain. 2005. "Teachers, Schools and Academic Achievement." *Econometrica* 73: 417–458.

48. Rockoff, Jonah E. 2004. "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data," *American Economic Review* 94: 247-252.

49. Rockoff, Jonah E., Douglas O. Staiger, Thomas J. Kane, and Eric S. Taylor. 2011. "Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools." *American Economic Review*, forthcoming.

50. Rockoff, Jonah E. and Cecilia Speroni. 2011. "Subjective and Objective Evaluations of Teacher Effectiveness: Evidence from New York City," *Labour Economics* 18: 687–696

51. Ronfeldt, Matthew, Hamilton Lankford, Susanna Loeb, James Wyckoff. 2011. "How Teacher Turnover Harms Student Achievement," NBER Working Paper No. 17176.

52. Rothstein, Jesse. 2009. "Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables," *Education Finance and Policy* 4(4), 537-571.

53. Rothstein, Jesse. 2010. "Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement," *Quarterly Journal of Economics* 125(1): 175-214.

54. Rubin, Donald B., Elizabeth A. Stuart and Elaine L. Zanutto. 2004. "A Potential Outcomes View of Value-Added Assessment in Education." *Journal of Educational and Behavioral Statistics*, 29(1): 103-116.

55. Springer, Matthew G., Ballou, Dale, Hamilton, Laura, Le, Vi-Nhuan, Lockwood, J.R., McCaffrey, Daniel F., Pepper, Matthew, and Brian M. Stecher. 2010. "Teacher Pay for Performance: Experimental Evidence from the Project on Incentives in Teaching." Nashville, TN: National Center on Performance Incentives at Vanderbilt University.

56. Staiger, Douglas O., and Jonah E. Rockoff. 2010. "Searching for Effective Teachers with Imperfect Information." *Journal of Economic Perspectives* 24: 97-117.

57. Todd, Petra E. and Kenneth I. Wolpin. 2003. "On the Specification and Estimation of the Production Function for Cognitive Achievement" *The Economic Journal* 113(485): F3-F33.

58. U.S. Census Bureau. 2006-2008. "American Community Survey."
ACS 3-year data. http://www.census.gov.

59. U.S. Census Bureau. 2010. "School Enrollment–Social and Economic Characteristics of Students: October 2008, Detailed." Washington, D.C.
http://www.census.gov/population/www/socdemo/school.html.

Appendix A: Matching Algorithm

We follow the matching algorithm developed in Chetty et al. (2011) to link the school district data to tax records. The algorithm was designed to match as many records as possible using variables that are not contingent on ex post outcomes. Date of birth, gender, and last name in the tax data are populated by the Social Security Administration using information that is not contingent on ex post outcomes. First name and ZIP code in tax data are contingent on observing some ex post outcome. First name data derive from information returns, which are typically generated after an adult outcome like employment (W-2 forms), college attendance (1098-T forms), and mortgage interest payment (1098 forms). The ZIP code on the claiming parent's 1040 return is typically from 1996 and is thus contingent on the ex post outcome of the student not having moved far from her elementary school for most students in our analysis sample.

Chetty et al. (2011) show that the match algorithm outlined below yields accurate matches for approximately 99% of cases in a school district sample that can be matched on social security number. Note that identifiers were used solely for the matching procedure. After the match was completed, the data were de-identified (i.e., individual identifiers such as names were stripped) and the statistical analysis was conducted using the de-identified dataset.

Step 1 [Date of Birth, Gender, Last Name]: We begin by matching each individual from the school-district data to Social Security Administration (SSA) records. We match individuals based on exact date of birth, gender, and the first four characters of last name. We only attempt to match individuals for which the school records include a valid date of birth, gender, and at least one valid last name. SSA records all last names ever associated in their records with a given individual; in addition, there are as many as three last names for each individual from the school files. We keep a potential match if any of these three last names match any of the last names present in the SSA file.

Step 2 [Rule Out on First Name]: We next check the first name (or names) of individuals from the school records against information from W2 and other information forms present in the tax records. Since these files reflect economic activity usually after the completion of school, we use this information in Step 2 only to "rule out" possible matches in order to minimize selection bias. In particular, we disqualify potential matches if none of the first names on the information returns match any of the first names in the school data. As before, we use only the first four characters of a first name. For many potential matches, we find no first name information in the tax information records; at this step we retain these potential matches. After removing potential matches that are mismatched on first name, we isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality (MQ) of matches identified at this stage as $MQ = 1$.

Step 3 [Dependent ZIP code]: For each potential match that remains, we find the household that claimed the individual as a dependent (if the individual was claimed at all) in each year. We then match the location of the claiming household, identified by the 5-digit ZIP code, to the home address ZIP code recorded in the school files. We classify potential matches based on the best ZIP code match across all years using the following tiers: exact match, match within 10 (e.g., 02139 and 02146 would qualify as a match), match within 100, and non-match. We retain potential matches only in the highest available tier of ZIP code match quality. For example, suppose there are 5 potential matches for a given individual, and that there are no exact matches on ZIP code, two matches within 10, two matches within 100, and one non-match. We would retain only the two that matched within 10. After this procedure, we isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match

pool. We classify the match quality of matches identified at this stage as $MQ = 2$.

Step 4 [Place of Birth]: For each potential match that remains, we match the state of birth from the school records with the state of birth as identified in SSA records. We classify potential matches into three groups: state of birth matches, state of birth does not match but the SSA state is the state where the school district is, and mismatches. Note that we include the second category primarily to account for the immigrants in the school data for whom the recorded place of birth is outside the country. For such children, the SSA state-of-birth corresponds to the state in which they received the social security number, which is often the first state in which they lived after coming to the country. We retain potential matches only in the best available tier of place-of-birth match quality. We then isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 3$.

Step 5 [Rule In on First Name]: After exhausting other available information, we return to the first name. To recall, in step 2 we retained potential matches that either matched on first name or for which there was no first name available. In this step, we retain only potential matches that match on first name, if such a potential match exists for a given student. We also use information on first name present on 1040 forms filed by potential matches as adults to identify matches at this stage. We then isolate students for whom only one potential match remains in the tax records. We declare such cases a match and remove them from the match pool. We classify the match quality of matches identified at this stage as $MQ = 4$.

Step 6 [Fuzzy Date-of Birth]: In previous work (Chetty et al. 2011), we found that 2-3% of individuals had a reported date of birth that was incorrect. In some cases the date was incorrect only by a few days; in others the month or year was off by one, or the transcriber transposed the month and day. To account for this possibility, we take all individuals for whom no eligible matches remained after step 2. Note that if any potential matches remained after step 2, then we would either settle on a unique best match in the steps that follow or find multiple potential matches even after step 5. We then repeat step 1, matching on gender, first four letters of last name, and fuzzy date-of-birth. We define a fuzzy DOB match as one where the absolute value of the difference between the DOB reported in the SSA and school data was in the set $\{1, 2, 3, 4, 59, 10, 18, 27\}$ in days, the set $\{1, 2\}$ in months, or the set $\{1\}$ in years. We then repeat steps 2 through 5 exactly as above to find additional matches. We classify matches found using this fuzzy-DOB algorithm as $MQ = 5.X$, where X is the corresponding MQ from the non-fuzzy DOB algorithm. For instance, if we find a unique fuzzy-DOB match in step 3 using dependent ZIP codes, then $MQ = 5.2$.

The following table shows the distribution of match qualities for all student-test-score observations. In all, we match 89.2% of student-subject observations in the analysis sample. We match 90.0% of observations in classes for which we are able to estimate VA for the teacher. Unmatched students are split roughly evenly among those for whom we found multiple matches and those for whom we found no match.

Match Quality (MQ)	Frequency	Percent	Cumulative Match Rate
1	3327727	55.63%	55.63%
2	1706138	28.52%	84.15%
3	146256	2.44%	86.59%
4	64615	1.08%	87.67%
5.1	84086	1.41%	89.08%
5.2	6450	0.11%	89.19%
5.3	747	0.01%	89.20%
5.4	248	0.00%	89.20%
Multiple Matches	304436	5.09%	
No Matches	341433	5.71%	

Appendix B: Identifying Teachers' Net Impacts

This appendix shows that the iterative method described in Section 6.1 recovers the net impacts of teacher VA, $\tilde{\beta}_g$, defined as the impact of raising teacher VA in grade g on earnings, holding fixed VA in subsequent grades.

We begin by estimating the following equations using OLS for $g \in [4, 8]$:

$$(12) \quad Y_i = \beta_g \hat{\mu}_{j(i,g)} + f^{\mu}_{1g}(A_{i,t-1}) + f^{\mu}_{2g}(e_{j(i,g,t)}) + \phi^{\mu}_{1g} X_{igt} + \phi^{\mu}_{2g} \bar{X}_{c(i,g,t)} + \varepsilon^{\mu}_{igt}$$

$$(13) \; \hat{\mu}_{j(i,g')} = \rho_{gg'} \hat{\mu}_{j(i,g)} + f^{g'}_{1g}(A_{i,t-1}) + f^{g'}_{2}(e_{j(i,g,t)}) + \phi^{g'}_{1g} X_{igt} + \phi^{g'}_{2g} \bar{X}_{c(i,g,t)} + \eta_{itgg'} \quad \forall g' > g$$

The first set of equations estimates the reduced form impact of teacher VA in grade g on earnings. The second set of equations estimates the impact of teacher VA in grade g on teacher VA in future grade g'. Denote by $\mathbb{X}$ the vector of controls in equations (12) and (13). Note that identification of the tracking coefficients $\rho_{gg'}$ using (6.1) requires the following variant of Assumption 2:

Assumption 2A Teacher value-added in grade g is orthogonal to unobserved determinants of future teacher value-added:

$$Cov\left(\hat{\mu}_{j(i,g)}, \eta_{itgg'} \mid \mathbb{X}\right) = 0.$$

After estimating $\{\beta_g\}$ and $\{\rho_{gg'}\}$, we recover the net impacts $\tilde{\beta}_g$ as follows. Under our definition of $\tilde{\beta}_g$, earnings can be written as $\sum_{g=1}^{G} \tilde{\beta}_g \hat{\mu}_{j(i,g)} + \varepsilon^{\mu}_i$. Substituting this definition of Y_i into (12) and noting that $\rho_{gg'} = Cov\left(\hat{\mu}_{j(i,g')}, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right) / Var\left(\hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)$ yields

$$\beta_g = \frac{Cov\left(\sum_{g'=1}^{G} \tilde{\beta}_{g'} \hat{\mu}_{j(i,g')} + \varepsilon^{Y}_i, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)}{Var\left(\hat{\mu}_{j(i,g)} \mid \mathbb{X}\right)} = \sum_{g'=1}^{G} \rho_{gg'} \tilde{\beta}_{g'}.$$

One implication of Assumption 2, the orthogonality condition needed to identify earnings impacts, is that

$$Cov\left(\hat{\mu}_{j(i,g')}, \hat{\mu}_{j(i,g)} \mid \mathbb{X}\right) = 0 \quad \text{for } g' < g$$

since past teacher quality $\hat{\mu}_{j(i,g')}$ is one component of the error term ε^{μ}_{igt} in (12). Combined with

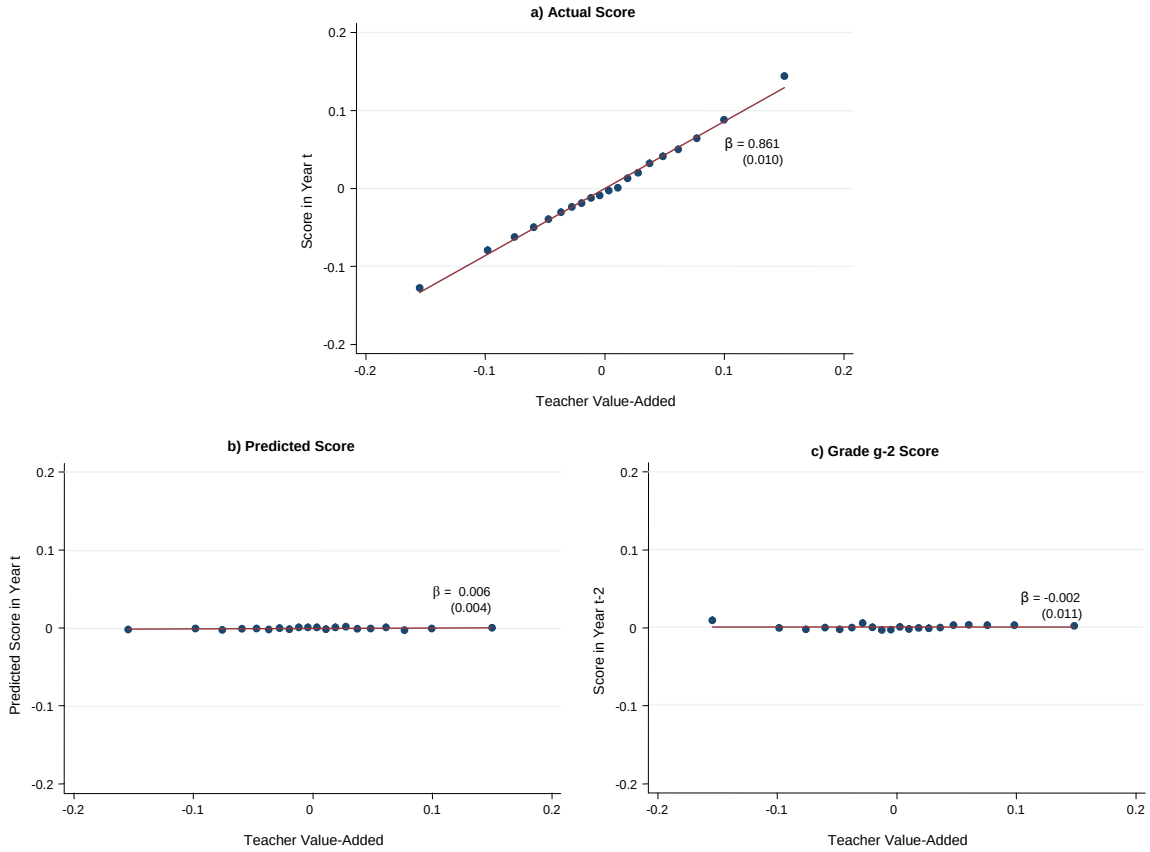
the fact that $\rho_{gg} = 1$ by definition, these equations imply that

$$\begin{aligned} \beta_g &= \tilde{\beta}_g + \sum_{g'=g+1}^{G} \rho_{gg'} \tilde{\beta}_{g'} \quad \forall g < G \\ \beta_G &= \tilde{\beta}_G. \end{aligned}$$

Rearranging this triangular set of equations yields the following system of equations, which can be solved by iterating backwards as in Section 6.1:

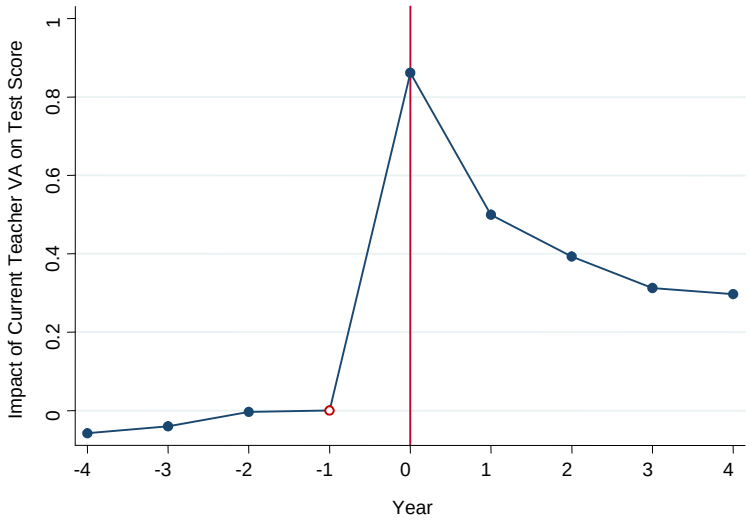
(14)
$$\begin{aligned} \tilde{\beta}_G &= \beta_G \\ \tilde{\beta}_g &= \beta_g - \sum_{g'=g+1}^{G} \rho_{gg'} \tilde{\beta}_{g'} \quad \forall g < G. \end{aligned}$$

FIGURE 1
Effects of Teacher Value-Added on Actual, Predicted, and Lagged Scores



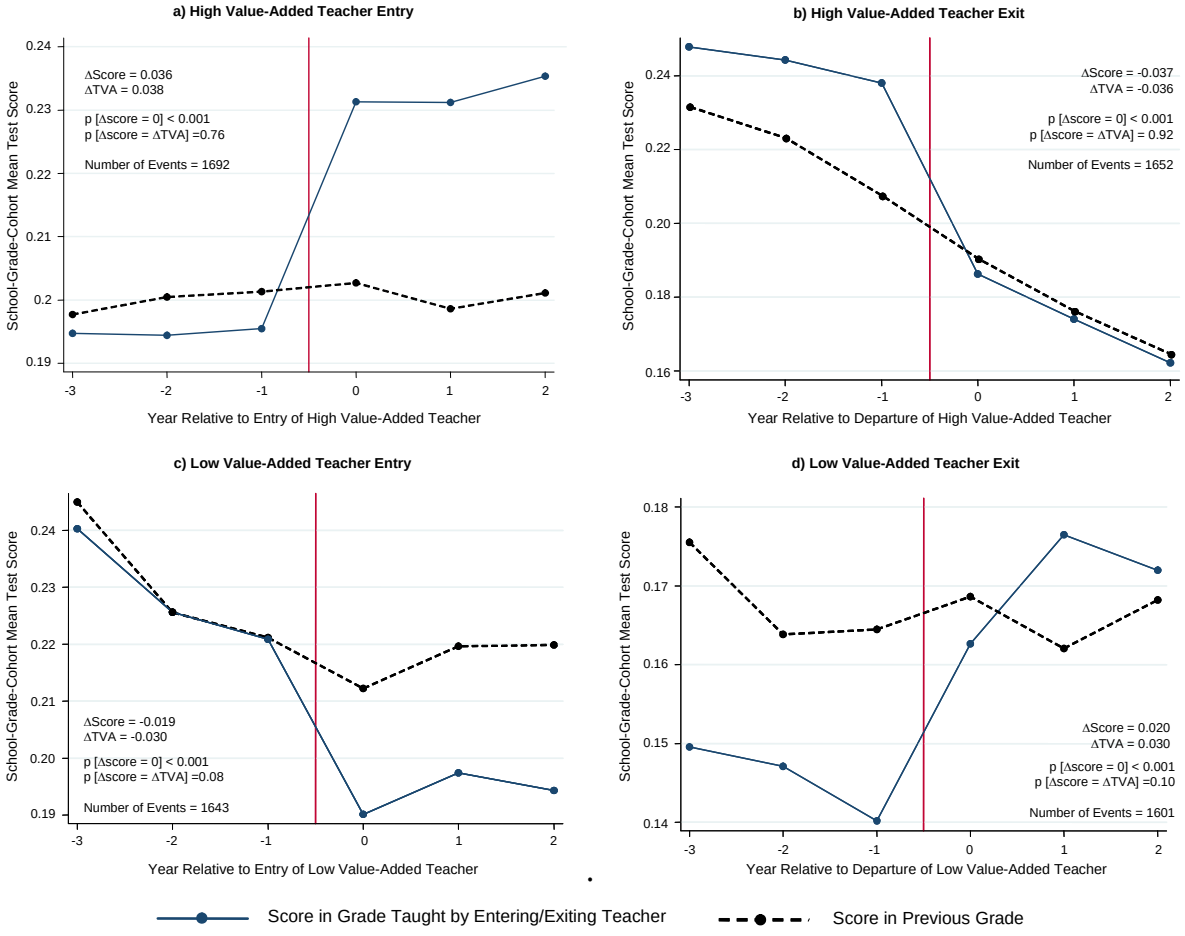
Notes: These figures plot student scores, scaled in standard deviation units, vs. our leave-year-out measure of teacher value-added, which is also scaled in units of student test score standard deviations. The figures are drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. In Panel A, the y variable is actual end-of-grade student scores; in Panel B, it is the predicted score based on parent characteristics; and in Panel C, it is the score two years before in the same subject. Predicted score is based on the fitted values from a regression of test score on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income (see Section 4.3 for details). All three figures control for the following classroom-level variables: school year and grade dummies, class-type indicators (honors, remedial), class size, and cubics in class and school-grade means of lagged test scores in math and English each interacted with grade. They also control for class and school-year means of the following student characteristics: ethnicity, gender, age, lagged suspensions, lagged absences, and indicators for grade repetition, special education, limited English. We use this baseline control vector in all subsequent figures unless otherwise noted. To construct each binned scatter plot, we first regress both the y- and x-axis variable on the control vector and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 2
Impacts of Teacher Value-Added on Lagged, Current, and Future Test Scores



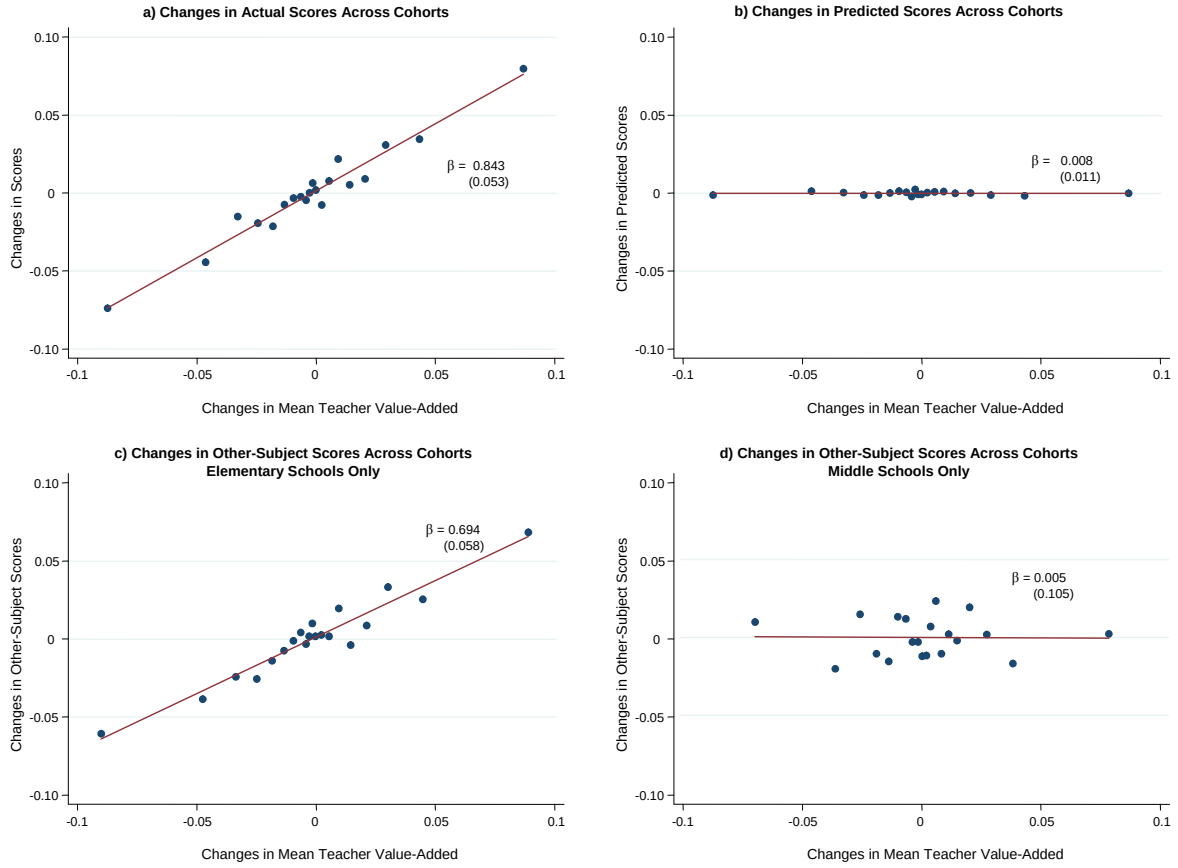
Notes: This figure shows the effect of teacher value-added in year $t = 0$ on student scores from four years prior to assignment to the teacher of interest to four years after. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Each point shows the coefficient estimate from a separate OLS regression of test scores (including all available grades and subjects) on teacher value-added and the baseline control vector used in Figure 1. The points for $t < -1$ represent placebo tests for selection on observables, while points for $t > 0$ show the persistence of teachers' impacts on test scores. The point at $t = 0$ corresponds to the regression coefficient in Panel A of Figure 1. The point at $t = -1$ is equal to zero by construction, because we control for lagged test scores. Teacher value-added is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The coefficients from the regressions along with their associated standard errors are reported in Appendix Table 9.

FIGURE 3
Impacts of Teacher Entry and Exit on Average Test Scores by Cohort



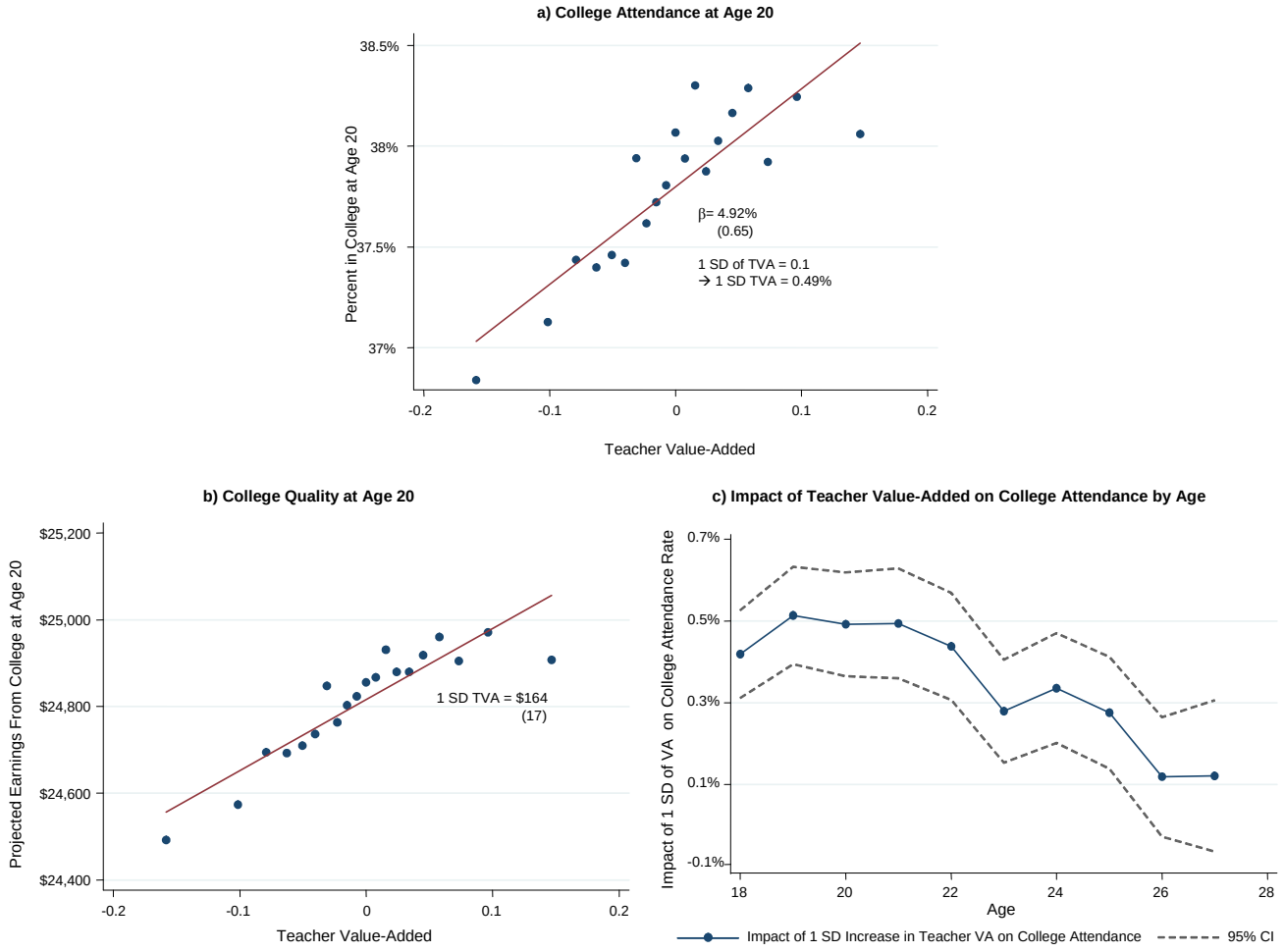
Notes: These figures plot event studies of current scores (solid line) and prior-year scores (dashed line) by cohort as teachers enter or leave a school-grade-subject cell in year $t = 0$. Panels A and B analyze the entry and exit of a high-VA teacher (teachers with VA in the top 5% of the distribution); Panels C and D analyze the entry and exit of a low-VA (bottom 5%) teacher. All panels are plotted using a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. To construct each panel, we first estimate each teacher's VA using data from classes taught outside the years $t \in [-3, 2]$. We then plot mean scores in the subject taught by the teacher for students in the entire school-grade-subject cell in the years before and after the arrival or departure of the teacher. We remove year fixed effects by regressing the y variable on year indicators and plotting the mean of the residuals, adding back the sample mean of each variable to facilitate interpretation of the scale. Each point therefore shows the mean score of a different cohort of students within a single school-grade-subject cell, removing secular time trends. Each panel reports the change in mean score gains (mean scores minus mean lag scores) from $t = -1$ to $t = 0$. We also report the change in mean teacher VA multiplied by 0.861, the cross-class coefficient of score on VA (Column 1 of Table 2). We multiply the change in mean VA by this factor to forecast the change in test scores implied by the change in mean VA. We report p values from F tests of the hypotheses that the change in score gains from $t = -1$ to $t = 0$ equals 0 and equals the change in mean VA times 0.861. Mean teacher VA is calculated using a student-weighted average, imputing the sample mean for teachers who do not have data outside the $t \in [-3, 2]$ window.

FIGURE 4
Effect of Changes in Teaching Staff on Scores Across Cohorts



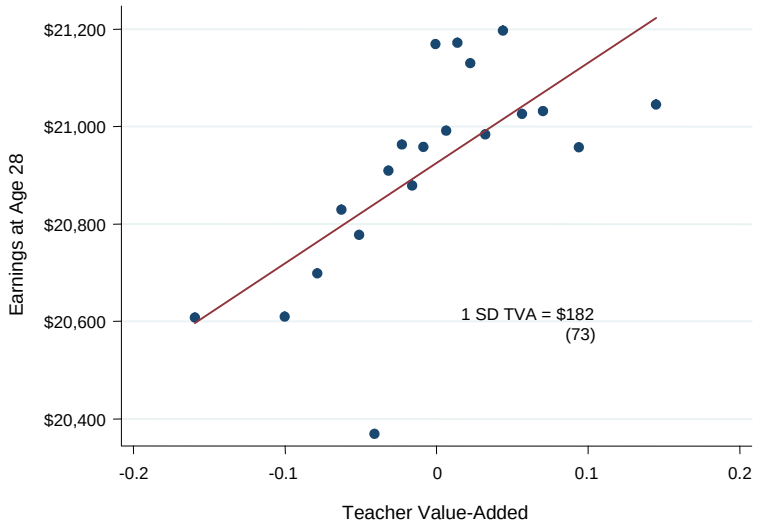
Notes: This figure plots changes in average test scores across cohorts versus changes in average teacher VA across cohorts, generalizing the event study in Figure 3 to include all changes in teaching staff. All panels are plotted using a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. We calculate changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows. First, we calculate teacher value-added for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean value-added across all teachers, weighting by the number of students they teach and imputing the sample mean VA for those for teachers for whom we have no estimate of VA. Finally, we compute the difference in mean teacher VA (year *t* minus year $t-1$) to obtain the x axis variable. The y axis variables are defined by calculating the change in the mean of the dependent variable (year t minus year t-1) within a school-grade-subject cell. In Panel A, the y-axis variable is the change in end-of-grade scores across cohorts in the relevant subject. In Panel B, the y-axis variable is the change in predicted test scores based on parent characteristics, defined as Figure 1b. In Panels C and D, the y-axis variable is the change in test scores in the other subject (e.g. math scores when analyzing English teachers' VA) for observations in elementary and middle school, respectively. To construct each binned scatter plot, we first regress both the y- and x-axis variable on year dummies and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin. The solid line shows the best linear fit estimated on the underlying school-grade-subject-year data estimated using an unweighted OLS regression. The coefficients show the estimated slope of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 5
Effects of Teacher Value-Added on College Attendance



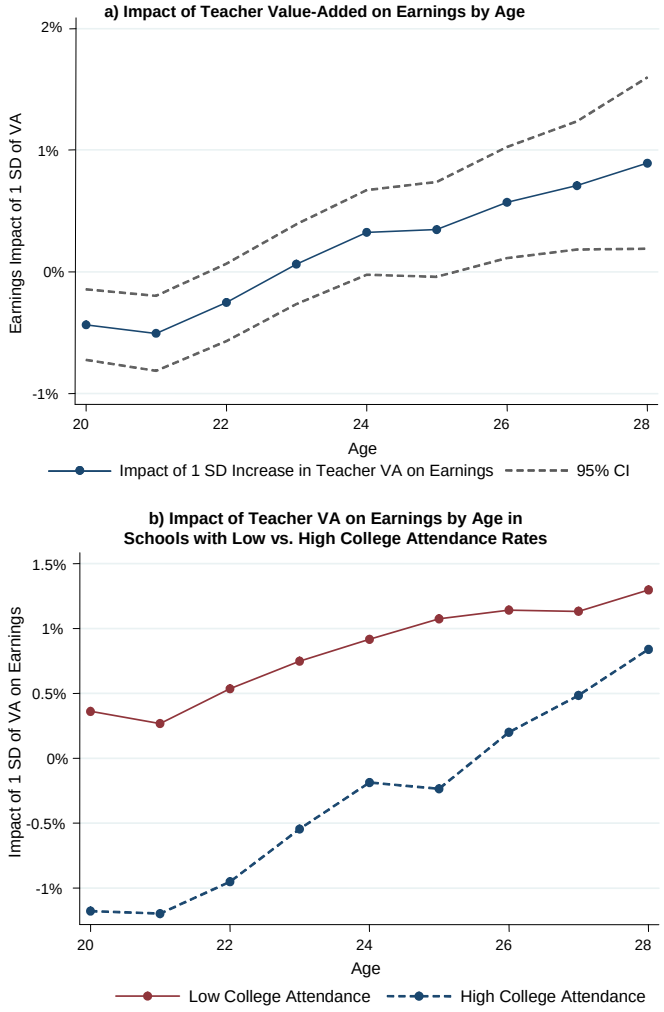
Notes: Panel A plots the relationship between teacher VA and college attendance rates at age 20. College attendance is measured by receipt of a 1098-T form in the year during which a student turned 20. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct the binned scatter plot, we first regress both the x- and y-variables on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the residual of the x variable and scatter the means of the y- and x-variable residuals within each bin, adding back the sample means of both variables to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficient shows the estimated slope of the best-fit line, with the standard error clustered at the school-cohort level reported in parentheses. Panel B replicates Panel A, changing the y variable to our earnings-based index of college quality at age 20. College quality is constructed using the average wage earnings at age 30 in 2009 for all students attending a given college at age 20 in 1999. For individuals who did not attend college, we calculate mean wage earnings at age 30 in 2009 for all individuals in the U.S. aged 20 in 1999 who did not attend any college. Panel C replicates the regression specification in Panel A and plots the resulting coefficients on college attendance from ages from 18 to 27. Each point represents the coefficient estimate on teacher value-added from a separate regression. The dashed lines show the boundaries of the 95% confidence intervals for the effect of value-added on college attendance at each age.

FIGURE 6
Effect of Teacher Value-Added on Earnings at Age 28



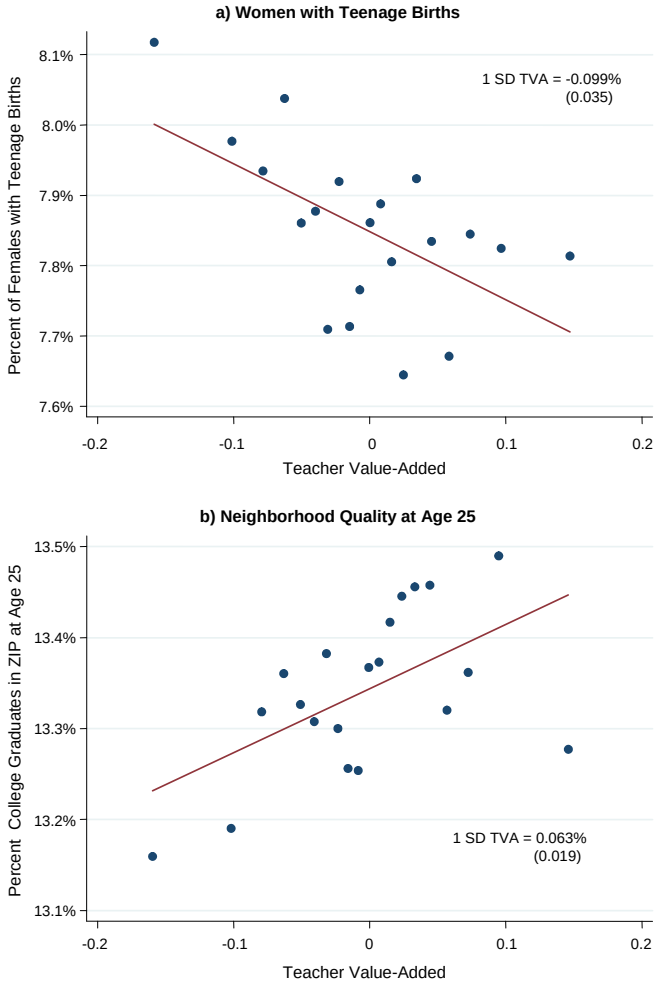
Notes: This figure plots the effect of teacher value-added on wage earnings at age 28, computed using data from W-2 forms issued by employers. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct the binned scatter plot, we first regress both earnings and value-added on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the value-added residual and scatter the means of the earnings and value-added residuals within each bin, adding back the sample means of earnings and value-added to facilitate interpretation of the scale. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficient shows the estimated slope of the best-fit line, with the standard error clustered at the school-cohort level reported in parentheses.

FIGURE 7
Effect of Teacher Value-Added on Earnings by Age



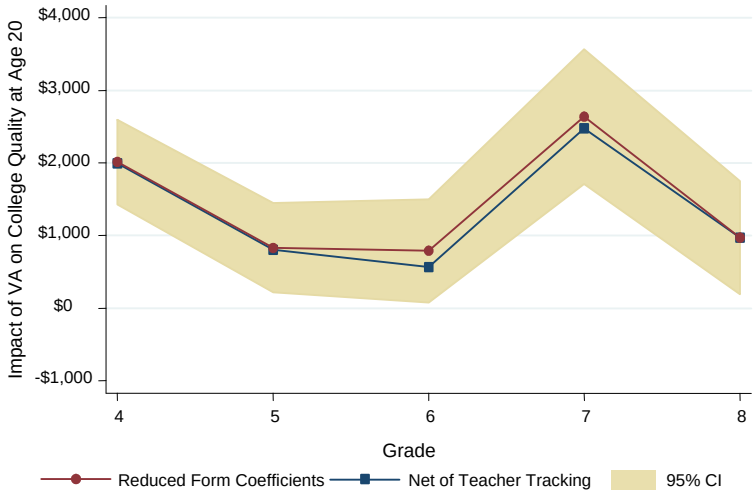
Notes: This figure shows the effect of a 1 SD increase in teacher value-added on earnings at each age, expressed as a percentage of mean earnings at that age. The figure is drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. To construct the figure, we first run a separate OLS regression of earnings at each age (using all observations for which the necessary data are available) on teacher value-added, following exactly the specification used in Figure 7 . We then divide this regression coefficient by 10 to obtain an estimate of the impact of a 1 SD increase in teacher VA on earnings. Finally, we divide the rescaled coefficient by the mean earnings level in the estimation sample at each age to obtain the percentage impact of a 1 SD increase in VA on earnings by age. Panel A shows the results for the full sample. The dashed lines represent the 95% confidence interval, computed using standard errors clustered at the school-cohort level. Panel B replicates Panel A, splitting the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. The solid series includes schools with attendance rates below 35% while the dashed series includes schools with attendance rates above 35%. The coefficients and standard errors underlying these figures are reported in Appendix Table 10.

FIGURE 8
Effects of Teacher Value-Added on Other Outcomes in Adulthood



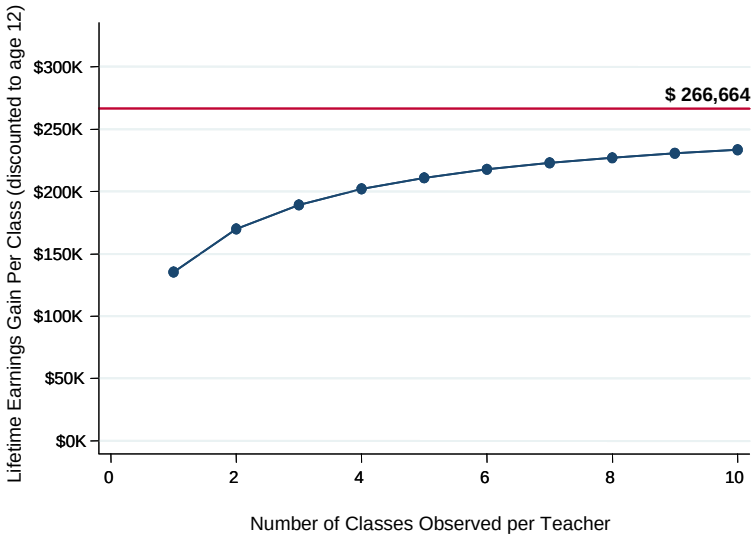
Notes: These figures plot the effect of teacher value-added on teenage births (for females only) and neighborhood quality. We define a teenage birth as an individual claiming a dependent who was born when she was between the ages of 13 and 19 on the 1040 tax form in any year in our sample (see Section 3.2 for details). We define neighborhood quality as the fraction of residents with a college degree in the ZIP code where the individual resides. The figures are drawn using the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. To construct each binned scatter plot, we first regress both the y- and x-axis variables on the baseline control vector used in Figure 1 and calculate residuals. We then group the observations into twenty equal-sized (5 percentile-point) bins based on the x-axis residual and scatter the means of the y- and x-axis residuals within each bin, adding back the sample means of x- and y- axis variables to facilitate interpretation of the scales. The solid line shows the best linear fit estimated on the underlying micro data estimated using OLS. The coefficients show the estimated slopes of the best-fit line, with standard errors clustered at the school-cohort level reported in parentheses.

FIGURE 9
Impacts of Teacher Value-Added on College Quality by Grade



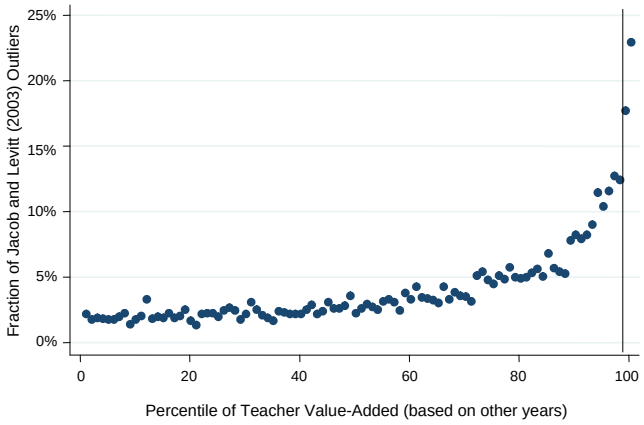
Notes: This figure plots the impact of a 1 SD increase in teacher VA in each grade from 4-8 on our earnings-based index of college quality (projected earnings at age 30 based on the college in which the student is enrolled at age 20). The figure is drawn using the linked analysis sample described in section 3.3. The upper (circle) series shows the reduced-form effect of improved teacher quality in each grade, including both the direct impact of the teacher on earnings and the indirect effect through improved teacher quality in future years. Each point in this series represents the coefficient on teacher value-added from a separate regression of college quality at age 20 on teacher VA for a single grade. We use the same specification as in Figure 5c but limit the sample to cohorts who would have been in 4th grade during or after 1994 to obtained a balanced sample across grades. The shaded area represents a 95% confidence interval, calculated based on standard errors clustered by school-cohort. The lower (square) series plots the impact of teachers in each grade on college quality netting out the impacts of increased future teacher quality. We net out the effects of future teachers using the tracking coefficients reported in Appendix Table 13 and solving the system of equations in Section 6.1.

FIGURE 10
Earnings Impacts of Deselecting Low Value-Added Teachers



Notes: This figure displays the present value of lifetime earnings gains for a single classroom of students from deselecting teachers whose estimated value-added is in the bottom 5% of the distribution. The horizontal line shows the gain that could be achieved by deselecting the bottom 5% of teachers based on their true VA μ_j, measured noiselessly using an infinite number of classes per teacher. The increasing series plots the feasible gains from deselection of the bottom 5% of teachers when their VA is estimated based on the number of classes shown on the *x* axis, accounting for finite-sample error in VA estimates. Appendix Table 14 lists the values that are plotted as well as undiscounted cumulative earnings gains, which are approximately 5.5 times larger in magnitude. To obtain the values in the figure, we first calculate the present value of average lifetime earnings per student using the cross-sectional life-cycle earnings profile for the U.S. population in 2007, discounting earnings back to age 12 using a 3% net discount rate (equivalent to a 5% discount rate with 2% wage growth). Column 1 of Table 6 implies that a 1 SD increase in VA raises earnings by 0.9% at age 28. We assume that this 0.9% earnings gain remains constant over the life cycle and calculate the impacts of a 1 SD improvement in teacher quality on mean lifetime earnings, averaging across English and math teachers. Finally, we multiply the mean lifetime earnings impact by 28.3, the mean class size in our analysis sample.

APPENDIX FIGURE 1
Jacob and Levitt (2003) Proxy for Test Manipulation vs. Value-Added Estimates



Notes: This figure plots the relationship between our leave-out-year measure of teacher value added and Jacob and Levitt's proxy for cheating. The figure is drawn using the linked analysis sample described in section 3.3. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The y- axis variable is constructed as follows. Let $\Delta\bar{A}_{c,t} = \bar{A}_{c,t} - \bar{A}_{c,t-1}$ denote the change in mean test scores from year $t-1$ to year t for students in classroom c. Let $R_{c,t}$ denote the ordinal rank of classroom c in $\Delta\bar{A}_{c,t}$ among classrooms in its grade, subject, and school year and $r_{c,t}$ the ordinal rank as a fraction of the total number of classrooms in that grade, subject, and school year. Jacob and Levitt's (2003) measure for cheating in each classroom is $JL_c = (r_{c,t})^2 + (1 - r_{c,t+1})^2$. Higher values of this proxy indicate very large test score gains followed by very large test score losses, which Jacob and Levitt show is correlated with a higher chance of having suspicious patterns of answers indicative of cheating. Following Jacob and Levitt, we define a classroom as an outlier if its value of JL_c falls within the top 5% of classrooms in the data. To construct the binned scatter plot, we group classrooms into percentiles based on their teacher's estimated value-added, ranking math and English classrooms separately. We then compute the fraction of Jacob-Levitt outliers within each percentile bin and scatter these fractions vs. the percentiles of teacher VA. Each point thus represents the fraction of Jacob-Levitt outliers at each subject-specific percentile of teacher VA, where VA for each teacher is estimated using data from other years. The dashed vertical line depicts the (subject-specific) 98th percentile of the value-added distribution. We exclude classrooms with estimated VA above this threshold in our baseline specifications because they have much higher frequencies of Jacob-Levitt outliers. See Appendix Table 8 for results with trimming at other cutoffs.

TABLE 1
Summary Statistics for Linked Analysis Dataset

Variable	Mean	S.D.	Observations
	(1)	(2)	(3)
Student Data:			
Class size (not student-weighted)	28.3	5.8	211,371
Number of subject-school years per student	6.14	3.16	974,686
Teacher experience (years)	8.08	7.72	4,795,857
Test score (SD)	0.12	0.91	5,312,179
Female	50.3%	50.0%	5,336,267
Age (years)	11.7	1.6	5,976,747
Free lunch eligible (1999-2009)	76.0%	42.7%	2,660,384
Minority (Black or Hispanic)	71.8%	45.0%	5,970,909
English language learner	10.3%	30.4%	5,813,404
Special education	3.4%	18.1%	5,813,404
Repeating grade	2.7%	16.1%	5,680,954
Student match rate to adult outcomes	89.2%	31.0%	5,982,136
Student match rate to parent chars.	94.6%	22.5%	5,329,715
Adult Outcomes:			
Annual wage earnings at age 20	4,796	6,544	5,255,599
Annual wage earnings at age 25	15,797	18,478	2,282,219
Annual wage earnings at age 28	20,327	23,782	851,451
In college at age 20	36.2%	48.1%	4,605,492
In college at age 25	17.3%	37.8%	1,764,179
College Quality at age 20	24,424	12,834	4,605,492
Contribute to a 401(k) at age 25	14.8%	35.5%	2,282,219
ZIP code % college graduates at age 25	13.2%	7.1%	1,919,115
Had a child while a teenager (for women)	8.4%	27.8%	2,682,644
Parent Characteristics:			
Household income (child age 19-21)	35,476	31,080	4,396,239
Ever owned a house (child age 19-21)	32.5%	46.8%	4,396,239
Contributed to a 401k (child age 19-21)	25.1%	43.3%	4,396,239
Ever married (child age 19-21)	42.1%	49.4%	4,396,239
Age at child birth	27.6	7.4	4,917,740
Predicted Score	0.16	0.26	4,669,069

Notes: All statistics reported are for the linked analysis dataset described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. The sample has one observation per student-subject-school year. Student data are from the administrative records of a large urban school district in the U.S. Adult outcomes and parent characteristics are from 1996-2010 federal income tax data. All monetary values are expressed in real 2010 dollars. All ages refer to the age of an individual as of December 31 within a given year. Teacher experience is the number of years of experience teaching in the school district. Test score is based on standardized scale scores, as described in Section 3.1. Free lunch is an indicator for receiving free of reduced-price lunches. Earnings are total wage earnings reported on W-2 forms, available from 1999-2010; those who are matched to tax data but have no W-2 are coded as having zero earnings. College attendance is measured by the receipt of a 1098-T form, available from 1999-2009. For a given college, "college quality" is defined as the average wage earnings at age 30 in 2009 for the subset of the U.S. population enrolled in that college at age 20 in 1999. For individuals who do not attend college, college quality is defined as the mean earnings at age 30 in 2009 of all individuals in the U.S. population not in college at age 20 in 1999. 401(k) contributions are reported on W-2 forms. ZIP code of residence is taken from either the address reported on 1040 or W-2 forms; for individuals without either in a given year, we impute location forward from the most recent non-missing observation. Percent college graduates in the ZIP code is based on data from the 2000 Census. Teenage births are measured only for females, by the claiming of a dependent, at any time in our sample, who was born when the claiming parent was between 13 and 19 years old. We link students to their parents by finding the earliest 1040 form from 1998-2010 on which the student is claimed as a dependent. We are unable to link 5.4% of matched students to their parents; the summary statistics for parents exclude these observations. Parent income is average adjusted gross income during the three tax-years when a student is aged 19-21. For parents who do not file, household income is defined as zero. Home ownership is measured by reporting mortgage interest payments on a 1040 or 1099 form. Marital status is measured by whether the claiming parent files a joint return while the child is between 19 and 21. Predicted score is predicted from a regression of scores on parent characteristics using the estimating equation in Section 4.3.

TABLE 2
Tests for Balance Using Parent Characteristics and Lagged Scores

Dep. Var.:	Score in year t	Predicted Score	Score in year t	Score in year t	Score in year t-2	Score in year t	Score in year t	Percent Matched
	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(SD)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Teacher VA	0.861	0.006	0.866	0.864	-0.002	0.803	0.804	0.160
	(0.010)	(0.004)	(0.011)	(0.011)	(0.011)	(0.011)	(0.011)	(0.280)
	[82.68]	[1.49]	[75.62]	[75.85]	[-0.21]	[72.74]	[70.63]	[0.562]
Pred. score using par. chars.				0.175				
				(0.012)				
				[62.70]				
Year t-2 Score							0.521	
							(0.001)	
							[363.3]	
Observations	3,721,120	2,877,502	2,877,502	2,877,502	2,771,865	2,771,865	2,771,865	4,018,504

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses and t-statistics in square brackets. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher VA is scaled in units of student test score standard deviations. VA is estimated using data from classes taught by the same teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. In this and all subsequent tables, we exclude outlier observations with teacher VA in the top 2% of the VA distribution unless otherwise noted. In columns 1, 3, 4, 6, and 7, the dependent variable is the student's test score in a given year and subject. In column 2, the dependent variable is the predicted value generated from a regression of test score on mother's age at child's birth, indicators for parent's 401(k) contributions and home ownership, and an indicator for the parent's marital status interacted with a quartic in parent's household income. See Section 4.3 for details of the estimating equation for predicted scores. In column 5, the dependent variable is the score two years earlier in the same subject. The dependent variable in column 8 is an indicator for being matched to the tax data. The second independent variable in each of columns 4 and 7 is the same as the dependent variables in columns 2 and 5, respectively. All specifications control for the following classroom-level variables: school year and grade dummies, class-type indicators (honors, remedial), class size, and cubics in class and school-grade means of lagged test scores in math and English each interacted with grade. They also control for class and school-year means of the following student characteristics: ethnicity, gender, age, lagged suspensions, lagged absences, and indicators for grade repetition, special education, limited English. We use this baseline control vector in all subsequent tables unless otherwise noted.

TABLE 3
Sensitivity of Teacher Value-Added Measures to Controls

	(1) baseline	(2) add parent chars.	(3) add t-2 scores	(4) t-1 scores only	(5) no controls	(6) Quasi-Experimental Estimate of Bias
(1) baseline	1.000					3.1% (7.6)
(2) add parent chars.	0.999	1.000				2.6% (7.6)
(3) add t-2 scores	0.975	0.974	1.000			1.4% (7.4)
(4) t-1 scores only	0.945	0.943	0.921	1.000		14.3% (6.9)
(5) no controls	0.296	0.292	0.279	0.323	1.000	87.8% (1.4)

Notes: Columns 1-5 of this table report correlations between teacher value-added estimates from five models, each using a different control vector. The correlations are weighted by the number of years taught by each teacher. The models are estimated on a constant subsample of 89,673 classrooms from the linked analysis dataset for which the variables needed to estimate all five models are available. For each model, we estimate student test score residuals using equation (3) using the relevant control vector and then implement the remaining steps of the Empirical Bayes procedure in Section 2.2 identically. Model 1 uses the student- and class-level control vector used to estimate value-added in our baseline specifications. This control vector includes a cubic polynomial in prior-year scores in math and a cubic in prior-year scores in English interacted with the student's grade level, dummies for teacher experience, as well as the following student-level controls: ethnicity, gender, age, lagged suspensions and absences, and indicators for grade repetition, special education, limited English. The control vector also includes the following classroom-level controls: class-type indicators (honors, remedial), class size, cubics in class and school-grade means of lagged test scores in math and English each interacted with grade, class and school-year means of all the student-level controls, and school year and grade dummies. Model 2 adds classroom-level means of the following parental characteristics to model 1: parent's age at child's birth, mean parent household income, and indicators for whether the parent owned a house, invested in a 401k, or was married while child was 19-21, and an indicator for whether no parent was found for the child in the tax data. Model 3 adds a cubic in twice-lagged test scores in the same subject to model 1. Model 4 controls for only lagged scores, using cubics in student's prior-year math and English scores interacted with grade level and cubics in the mean prior-year math and English scores for the classroom and school-grade cell also interacted with grade level. Model 5 includes no controls. In column 6, we estimate the degree of bias in the VA estimates produced by each model using quasi-experimental changes in teaching staff as described in Section 4.4. To calculate the degree of bias, we first estimate the effect of changes in mean VA on changes in test scores across cohorts using the specification in Column 1 of Table 4. We then estimate the effect of differences in teacher VA across classrooms on test scores, using the specification in Column 1 of Table 2 but with the control vector corresponding exactly to that used to estimate the value-added model. Finally, we define the degree of bias as the percentage difference between the cross-cohort and cross-class coefficients. Standard errors for the bias calculation are calculated as the standard error of the coefficient in the cross-cohort regression divided by the cross-class estimate; this calculation ignores the error in the cross-class estimate, which is negligible, as shown in Column 1 of Table 2.

TABLE 4
Impacts of Quasi-Experimental Changes in Teaching Staff on Test Scores

Dependent Variable:	Δ Score	Δ Predicted Score	Δ Other Subj. Score	Δ Other Subj. Score
	(SD)	(SD)	(SD)	(SD)
	(1)	(2)	(3)	(4)
Changes in mean teacher VA across cohorts	0.843	0.008	0.694	0.005
	(0.053)	(0.011)	(0.058)	(0.105)
	[15.95]	[0.74]	[11.90]	[0.04]
Grades	4 to 8	4 to 8	Elem. Sch.	Middle Sch.
Number of school x grade x subject x year cells	24,887	25,073	20,052	4,651

Notes: Each column reports coefficients from an unweighted OLS regression, with standard errors clustered by school-cohort in parentheses and t-statistics in square brackets. The regressions are estimated on a dataset containing school x grade x subject x year means from the linked analysis sample described in section 3.3. We calculate changes in mean teacher VA across consecutive cohorts within a school-grade-subject cell as follows. First, we calculate teacher value-added for each teacher in a school-grade-subject cell in each adjacent pair of school years using information excluding those two years. We then calculate mean value-added across all teachers, weighting by the number of students they teach and imputing the sample mean VA for those for teachers for whom we have no estimate of VA. Finally, we compute the difference in mean teacher VA (year t minus year t-1) to obtain the independent variable. We do not exclude teachers whose estimated VA is in the top 2% of the distribution when computing mean VA. The dependent variables are defined by calculating the change in the mean of the dependent variable (year t minus year t-1) within a school-grade-subject cell. In column 1, the dependent variable is the change in mean scores in the corresponding subject. In column 2, it is the change in the predicted score, constructed as described in the notes to Table 2. In Columns 3 and 4, the dependent variable is the change in the score in the other subject (e.g. math scores for English teachers). Column 3 restricts the sample to elementary schools, where math and English are taught by the same teacher; column 4 restricts the sample to middle schools, where different teachers teach the two subjects. All specifications include no controls except year fixed effects.

TABLE 5
Impacts of Teacher Value-Added on College Attendance

Dep. Var.:	College at Age 20	Pred. College at Age 20	Changes in Age 20 Coll. Attendance	College Quality at Age 20	Changes in Age 20 Coll. Quality	High Quality Coll. at Age 20	College at Age 25
	(%)	(%)	(%)	($)	($)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher VA	4.917	0.463		1,644		3.588	2.752
	(0.646)	(0.261)		(173)		(0.612)	(0.697)
Changes in mean VA across cohorts			6.101		1,319		
			(2.094)		(539)		
Controls	x	x		x		x	x
Source of Variation	X-Class	X-Class	X-Cohort	X-Class	X-Cohort	X-Class	X-Class
Observations	3,095,822	3,097,322	25,073	3,095,822	24,296	3,095,822	985,500
Mean of Dep. Var.	37.8	37.8	35.9	24,815	24,293	19.8	18.1

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. Columns 1, 2, 4, 6, and 7 use cross-class variation, while columns 3 and 5 use cross-cohort variation. For specifications that use cross-class variation, teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the baseline control vector in model 1 of Table 3. The dependent variable in column 1 is an indicator for college attendance at age 20. The dependent variable in column 2 is the predicted value generated from a regression of college attendance at age 20 on parent characteristics, using the same specification as for predicted score described in the notes to Table 2. The dependent variable in column 4 is the earnings-based index of college quality, defined in the notes to Table 1. The dependent variable in column 6 is an indicator for attending a college whose quality is greater than the median college quality among those attending college, which is $39,972. The dependent variable in column 7 is an indicator for college attendance at age 25. All cross-class regressions include the baseline class-level control vector used in Table 2. For the specifications that exploit cross-cohort variation in columns 3 and 5, we use changes in mean teacher value-added as the main independent variable, defined exactly as in Table 4. The dependent variables in Columns 3 and 5 are changes in mean college attendance and quality across consecutive cohorts within a school-grade-subject cell. Columns 3 and 5 include no controls except year fixed effects.

TABLE 6
Impacts of Teacher Value-Added on Earnings

Dep. Var.:	Earnings at Age 28	Earnings at Age 30	Wage Growth Ages 22-28	College at Age 25	College at Age 25	Wage Growth Ages 22-28	Wage Growth Ages 22-28
	($)	($)	($)	(%)	(%)	($)	($)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Teacher VA	1,815	2,058	1,802	0.526	4.728	1,403	2,838
	(729)	(1953)	(636)	(0.789)	(1.152)	(661)	(1,118)
Observations	368,427	61,639	368,405	528,065	457,435	201,933	166,472
Schools	All	All	All	Low Coll.	High Coll.	Low Coll.	High Coll.
Mean of Dep. Var.	20,912	22,347	14,039	14.30	22.43	10,159	18,744

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The dependent variable in columns 1 and 2 are the individual's wage earnings reported on W-2 forms at ages 28 and 30, respectively. The dependent variable in columns 3, 6, and 7 is the change in wage earnings between ages 22 and 28. The dependent variable in columns 4 and 5 is an indicator for attending college at age 25. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. Columns 1-3 use the entire analysis sample. In columns 4-7, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Columns 4 and 6 use schools with attendance rates below 35% while columns 5 and 7 use schools with attendance rates above 35%.

TABLE 7
Impacts of Teacher Value-Added on Other Outcomes

Dep. Var.:	Teenage Birth	Percent College Grads in ZIP at Age 25	Percent College Grads in ZIP at Age 28	401(k) at Age 25	401(k) at Age 25
	(%)	(%)	(%)	(%)	(%)
	(1)	(2)	(3)	(4)	(5)
Value-Added	-0.991	0.628	1.439	1.885	-1.780
	(0.353)	(0.194)	(0.310)	(0.680)	(0.987)
Observations	1,826,742	1,168,965	310,638	725,140	646,955
Schools	All	All	All	Low Coll.	High Coll.
Mean of Dep. Var.	7.9	13.3	13.6	12.1	19.2

Notes: Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. The dependent variable in column 1 is an indicator for having a teenage birth, defined as in Table 1. The dependent variable in columns 2 and 3 is the fraction of residents in an individual's zip code of residence at ages 25 and 28 with a college degree or higher, based on data from the 2000 Census. ZIP code is obtained from either 1040 or W-2 forms filed in the current year or imputed from past years for non-filers. The dependent variable in columns 4 and 5 is an indicator for whether an individual made a contribution to a 401(k) plan at age 25. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. Columns 1-3 use the entire analysis sample. In columns 4 and 5, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Columns 4 and 6 use schools with attendance rates below 35% while columns 5 and 7 use schools with attendance rates above 35%.

TABLE 8
Heterogeneity in Impacts of Teacher Value-Added

Panel A: Impacts by Demographic Group

	Girls	Boys	Low Income	High Income	Minority	Non-Minority
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: College Quality at Age 20 ($)						
Teacher VA	1,903	1,386	1,227	2,087	1,302	2,421
	(211)	(203)	(174)	(245)	(154)	(375)
Mean of Dep. Var.	25,509	24,106	21,950	27,926	21,925	31,628
Dependent Variable: Test Score (SD)						
Teacher VA	0.856	0.863	0.843	0.865	0.846	0.889
	(0.012)	(0.013)	(0.014)	(0.013)	(0.012)	(0.018)
Mean of Dep. Var.	0.191	0.161	-0.010	0.324	-0.037	0.663

Panel B: Impacts by Subject

	Elementary School			Middle School		
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent Variable: College Quality at Age 20 ($)						
Math Teacher VA	1095		638	1,648		1,374
	(176)		(219)	(357)		(347)
English Teacher VA		1,901	1,281		2,896	2,543
		(303)	(376)		(586)	(574)

Notes: Each cell reports a coefficient from a separate OLS regression of an outcome on teacher value-added, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. All regressions exploit variation across classrooms and include the baseline class-level control vector used in Table 2. In Panel A, there is one observation for each student-subject-school year; in Panel B, the data are reshaped so that both subjects (math and English) are in the same row, with one observation for each student-school year. The dependent variable in the top half of Panel A and in Panel B is the earnings-based index of college quality (see Table 1 for details). The dependent variable in the second half of Panel A is the student's test score. In Panel A, we split the sample in columns 1 and 2 between boys and girls. We split the sample in columns 3 and 4 based on whether a student's parental income is higher or lower than median in sample, which is $26,961. We split the sample in columns 5 and 6 based on whether a student belongs to an ethnic minority (Black or Hispanic). In Panel B, we split the sample into elementary schools (schools where the student is taught by the same teacher for both math and English) and middle schools (which have different teachers for each subject). All specifications in Panel B control for the baseline class-level variables described in Table 2 in both the student's math and English classrooms.

APPENDIX TABLE 1
Structure of Linked Analysis Dataset

Student	Subject	Year	Grade	Class	Teacher	Test Score	Matched to Tax Data?	In college at Age 20?	Earnings at Age 28	Parent Income
			…							
Bob	Math	1992	4	1	Jones	0.5	1	1	$27K	$95K
Bob	English	1992	4	1	Jones	-0.3	1	1	$27K	$95K
Bob	Math	1993	5	2	Smith	0.9	1	1	$27K	$95K
Bob	English	1993	5	2	Smith	0.1	1	1	$27K	$95K
Bob	Math	1994	6	3	Harris	1.5	1	1	$27K	$95K
Bob	English	1994	6	4	Adams	0.5	1	1	$27K	$95K
Nancy	Math	2002	3	5	Daniels	0.4	0	.	.	
Nancy	English	2002	3	5	Daniels	0.2	0	.	.	
Nancy	Math	2003	4	6	Jones	-0.1	0	.	.	
Nancy	English	2003	4	6	Jones	0.1	0	.	.	
			…							

Notes: This table illustrates the structure of the analysis dataset, which combines information from the school district database and the tax data. The linked analysis data includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one row for each student-subject-school year, with 5,928,136 rows in total. Individuals who were not linked to the tax data have missing data on adult outcomes and parent characteristics. The values in this table are not real data and for illustrative purposes only.

APPENDIX TABLE 2
Summary Statistics for School District Data Used to Estimate Value-Added

Variable	Mean (1)	S.D. (2)	Observations (3)
Class size (not student-weighted)	27.5	5.1	318,812
Number of subject-school years per student	5.58	2.98	1,375,552
Teacher Experience (years)	7.3	7.3	7,675,495
Test score (SD)	0.17	0.88	7,675,495
Female	50.8%	50.0	7,675,288
Age (years)	11.4	1.5	7,675,282
Free lunch eligible (1999-2009)	79.6%	40.3%	5,046,441
Minority (Black or Hispanic)	71.6%	45.1%	7,672,677
English language learner	5.1%	22%	7,675,495
Special education	1.9%	13.7%	7,675,495
Repeating grade	1.7%	13.0%	7,675,495

Notes: Statistics reported are for the set of observations used to estimate teacher value-added. These are observations from the full school district dataset spanning 1991-2009 described in section 3.1 that have information on test scores, teachers, and all the control variables (such as lagged test scores) needed to estimate the baseline value-added model in Table 3. We exclude observations from classrooms that have fewer than 7 students with the necessary information to estimate value-added. The sample has one observation per student-subject-school year. See notes to Table 1 for definitions of variable and additional details.

APPENDIX TABLE 3
Cross-Sectional Correlations Between Outcomes in Adulthood and Test Scores

Dep. Var.:	Earnings at Age 28	College at Age 20	College Quality at Age 20	Teenage Birth	Percent College Grads in ZIP at Age 25
	($)	(%)	($)	(%)	(%)
	(1)	(3)	(2)	(4)	(5)
No Controls	7,601	18.33	6,030	-3.84	1.85
	(28)	(0.02)	(6)	(0.02)	(0.01)
With Controls	2,539	5.66	2,009	-1.03	0.37
	(76)	(0.05)	(13)	(0.04)	(0.01)
Math	2,813	5.97	2,131	-0.88	0.34
Full Controls	(104)	(0.07)	(18)	(0.06)	(0.02)
English	2,194	5.27	1,843	-1.21	0.38
Full Controls	(112)	(0.07)	(18)	(0.06)	(0.02)
Mean of Dep. Var.	20,867	37.2	24,678	8.25	13.2

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on test scores measured in standard deviation units, with standard errors reported in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The dependent variable is wage earnings at age 28 in column 1, an indicator for attending college at age 20 in column 2, our earnings-based index of college quality in column 3, an indicator for having a teenage birth (defined for females only) in column 4, and the fraction of residents in an individual's zip code of residence with a college degree or higher at age 25 in column 5. See notes to Table 1 for definitions of these variables. The regressions in the first row include no controls. The regressions in the second row include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The regressions in the third and fourth row both include the full vector of controls and split the sample into math and English test score observations.

APPENDIX TABLE 4
Cross-Sectional Correlations Between Test Scores and Earnings by Age

	Dependent Variable: Earnings ($)										
	20	21	22	23	24	25	26	27	28	29	30
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
No Controls	435	548	1,147	2,588	3,516	4,448	5,507	6,547	7,440	8,220	8,658
	(15)	(19)	(24)	(30)	(36)	(42)	(49)	(56)	(63)	(68)	(72)
With Controls	178	168	354	942	1,282	1,499	1,753	2,151	2,545	2,901	3,092
	(46)	(60)	(75)	(94)	(111)	(130)	(151)	(172)	(191)	(208)	(219)
Mean Earnings	4,093	5,443	6,986	9,216	11,413	13,811	16,456	19,316	21,961	23,477	23,856
Pct. Effect (with controls)	4.4%	3.1%	5.1%	10.2%	11.2%	10.9%	10.7%	11.1%	11.6%	12.4%	13.0%

Notes: Each cell in the first two rows reports coefficients from a separate OLS regression of earnings at a given age on test scores measured in standard deviation units, with standard errors in parentheses. We obtain data on earnings from W-2 forms and include individuals with no W-2's as observations with 0 earnings. The regressions are estimated on a constant subsample of the linked analysis sample, i.e. the subset of students for whom data on earnings are available from ages 20-30. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The first row includes no controls; the second includes the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. Means of earnings for the available estimation sample are shown in the third row. The last row divides the coefficient estimates from the specification with controls by the mean earnings to obtain a percentage impact by age.

APPENDIX TABLE 5
Heterogeneity in Cross-Sectional Correlations Across Demographic Groups

Dependent Variable:	Earnings at Age 28	College at at Age 20	College Quality Age 20	Teenage Birth
	($)	(%)	($)	(%)
	(1)	(2)	(3)	(4)
Male	2,235	5.509	1,891	n/a
	(112)	(0.069)	(18)	n/a
	[21,775]	[0.34567]	[24,268]	n/a
Female	2,819	5.828	2,142	-1.028
	(102)	(0.073)	(19)	(0.040)
	[20,889]	[0.42067]	[25,655]	[0.07809]
Non-minority	2,496	5.560	2,911	-0.550
	(172)	(0.098)	(30)	(0.039)
	[31,344]	[0.60147]	[32,288]	[0.01948]
Minority	2,583	5.663	1,624	-1.246
	(80)	(0.058)	(13)	(0.053)
	[17,285]	[0.29627]	[22,031]	[0.10038]
Low Parent Inc.	2,592	5.209	1,571	-1.210
	(108)	(0.072)	(17)	(0.072)
	[17,606]	[0.27636]	[22,011]	[0.10384]
High Parent Inc.	2,614	5.951	2,414	-0.834
	(118)	(0.072)	(19)	(0.054)
	[26,688]	[0.49882]	[28,038]	[0.05974]

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on test scores measured in standard deviation units, with standard errors reported in parentheses. Means of the dependent variable for the relevant estimation sample are shown in square brackets. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year, and we pool all subjects and grades in estimating these regressions. The dependent variable is wage earnings at age 28 in column 1, an indicator for attending college at age 20 in column 2, our earnings-based index of college quality in column 3, and an indicator for having a teenage birth (defined for females only) in column 4. All regressions include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The demographic groups are defined in exactly the same way as in Table 8. We split the sample in rows 3 and 4 based on whether a student belongs to an ethnic minority (Black or Hispanic). We split the sample in rows 5 and 6 based on whether a student's parental income is higher or lower than median in sample, which is $26,961.

APPENDIX TABLE 6
Cross-Sectional Correlations between Test Scores and Outcomes in Adulthood by Grade

Dep. Variable:	Earnings at Age 28	College at Age 20	College Quality at Age 20	Earnings at Age 28	College at Age 20	College Quality at Age 20
	No Controls			With Controls		
	($) (1)	(%) (2)	($) (3)	($) (4)	(%) (5)	($) (6)
Grade 4	7,618	18.2	5,979	3,252	6.763	2,360
	(76.7)	(0.053)	(13.8)	(157)	(0.099)	(25.4)
Grade 5	7,640	18.3	6,065	2,498	5.468	1,994
	(61.6)	(0.052)	(13.6)	(129)	(0.096)	(24.8)
Grade 6	7,395	18.0	5,917	2,103	4.987	1,778
	(63.0)	(0.057)	(14.7)	(161)	(0.118)	(29.8)
Grade 7	7,790	18.4	5,950	2,308	4.844	1,667
	(64.6)	(0.060)	(15.5)	(342)	(0.133)	(33.2)
Grade 8	7,591	18.9	6,228	2,133	5.272	1,913
	(54.7)	(0.055)	(14.1)	(196)	(0.129)	(32.3)
Mean of Dep Var.	20,867	37.17	24,678	20,867	37.17	24,678

Notes: Each cell reports coefficients from a separate OLS regression of an outcome in adulthood on end-of-grade test scores measured in standard deviation units, using data from only a single grade. Standard errors are reported in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Columns 1-3 do not include any controls. Columns 4-6 include the full vector of student- and class-level controls used to estimate the baseline value-added model, described in the notes to Table 3. The dependent variable in columns 1 and 4 is wage earnings at age 28. The dependent variable in columns 2 and 5 is an indicator for college attendance at age 20. The dependent variable in columns 3 and 6 is our earnings-based index of college quality.

APPENDIX TABLE 7
Robustness Analysis: Clustering and Control Vectors

Dependent Variable:	Score	College at Age 20	Earnings at Age 28
	(SD)	(%)	($)
	(1)	(2)	(3)
Panel A: Baseline Analysis Sample			
Baseline estimates	0.861	0.049	1815
Baseline s.e. (school-cohort)	(0.010)	(0.006)	(727)
95% CI	(0.841, 0.882)	(0.037,0.062)	(391, 3240)
95% CI using student bootstrap	(0.851, 0.871)	(0.040, 0.056)	(630, 3095)
p value using student bootstrap	<.01	<.01	<.01
Panel B: Observations with Data on Earnings at Age 28			
	1.157	0.060	1815
no clustering	(0.016)	(0.010)	(531)
school-cohort	(0.036)	(0.016)	(727)
two-way student and class	(0.029)	(0.013)	(675)
Panel C: First observation for each child, by subject			
Math	0.986	0.040	1258
no clustering	(0.009)	(0.006)	(780)
school-cohort	(0.017)	(0.007)	(862)
class	(0.016)	(0.007)	(848)
English	1.116	0.061	2544
no clustering	(0.015)	(0.010)	(1320)
school-cohort	(0.025)	(0.012)	(1576)
class	(0.024)	(0.012)	(1516)
Panel D: Additional Controls			
Baseline class controls	0.858	0.049	1696
school-cohort	(0.010)	(0.007)	(797)
Add Individual Controls	0.856	0.049	1688
school-cohort	(0.010)	(0.007)	(792)
Add School-Year Effects	0.945	0.026	1942
school-cohort	(0.009)	(0.005)	(669)

Notes: This table reports coefficient estimates, with standard errors or confidence intervals in parentheses, from OLS regressions of various outcomes on teacher value-added. The dependent variable in column 1 is score. The dependent variable in column 2 is an indicator for college attendance at age 20. The dependent variable in column 3 is wage earnings at age 28. Panel A reports the results from the baseline specifications estimated on the full linked analysis sample, along with a 95% confidence interval generated from a block-bootstrap at the student level. Panel B reports results for the subsample of observations for whom we have data on earnings at age 28. We report three sets of standard errors: no clustering, clustering by school-cohort as in our baseline analysis, and two-way clustering by student and classroom (Cameron, Gelbach, and Miller 2011). In Panel C, we eliminate repeated observations at the individual level by using only the first observation per student in each subject. We then report the same three sets of standard errors. Finally, Panel D evaluates the sensitivity of the estimates to changes in the control vector. The first and second rows of Panel D use the subsample of observations with non-missing student-level controls. The first row uses the baseline classroom-level controls used in Table 2 and other tables, while the second adds the student-level controls used to estimate our baseline value-added model (model 1 in Table 3). The third row uses the full analysis sample and includes school-year fixed effects in both the estimation of teacher VA and the outcome regressions.

APPENDIX TABLE 8
Impacts of Teacher Value-Added: Sensitivity to Trimming

	Percent Trimmed in Upper Tail						Bottom and Top 2%	Jacob and Levitt proxy
	5%	4%	3%	**2%**	1%	0%		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Test Score	0.846	0.853	0.860	0.861	0.868	0.870	0.866	0.754
	(0.011)	(0.011)	(0.011)	(0.010)	(0.010)	(0.010)	(0.011)	(0.011)
College at Age 20	5.724	5.585	5.258	4.917	4.730	4.022	4.091	6.455
	(0.693)	(0.673)	(0.662)	(0.646)	(0.622)	(0.590)	(0.668)	(0.703)
College Quality at Age 20	1,870	1,848	1,773	1,644	1,560	1,432	1,425	2,068
	(185)	(180)	(177)	(173)	(167)	(160)	(177)	(187)
Earnings at Age 28	2,058	2,080	1,831	1,815	1,581	994	1,719	1,672
	(808)	(776)	(745)	(729)	(709)	(668)	(797)	(834)

Notes: Each coefficient reports the coefficient on teacher VA from a separate OLS regression, with standard errors clustered by school-cohort in parentheses. The dependent variable is end-of-grade test score in the first row, an indicator for college attendance in the second row, our earnings-based index of college quality in the third row, and earnings at age 28 in the fourth row. The regressions in each of these rows replicate exactly the baseline cross-class specification used in Column 1 of Table 2, Columns 1 and 5 of Table 4, and Column 1 of Table 5. The baseline estimates are reported in column 4, which shows the results with trimming the top 2% of VA outliers. Columns 1-6 report results for trimming the upper tail at other cutoffs. Column 7 shows estimates when both the bottom and top 2% of VA outliers are excluded. Finally, column 8 excludes teachers who have more than one classroom that is an outlier according to Jacob and Levitt's (2003) proxy for cheating. Jacob and Levitt define an outlier classroom as one that ranks in the top 5% of a test-score change metric defined in the notes to Appendix Figure 1.

APPENDIX TABLE 9
Impacts of Teacher Value-Added on Lagged, Current, and Future Test Scores

	Dependent Variable: Test Score (SD)							
	t-4	t-3	t-2	t	t+1	t+2	t+3	t+4
	(1)	(2)	(3)	(5)	(6)	(7)	(8)	(9)
Teacher VA	-0.059	-0.041	-0.004	0.861	0.499	0.393	0.312	0.297
	(0.020)	(0.015)	(0.011)	(0.010)	(0.011)	(0.012)	(0.013)	(0.018)
Observations	1,184,397	1,906,149	2,826,978	3,721,120	2,911,042	2,247,141	1,578,551	790,173

Notes: This table reports the values plotted in Figure 2. Each column reports coefficients from an OLS regression, with standard errors clustered by school-cohort in parentheses. The regressions are estimated on the linked analysis sample described in section 3.3, which includes only students who would graduate high school in or before 2008 if progressing at a normal pace. There is one observation for each student-subject-school year. Teacher value-added is estimated using data from classes taught by a teacher in other years, following the procedure in Sections 2.2 and 4.1 and using the control vector in model 1 of Table 3. Each column replicates exactly the specification in Column 1 of Table 2, replacing the dependent variable with scores in year $t + s$ to measure the impact of teacher quality in year t, where s varies from -4 to 4. We omit the specification for $s = -1$ since we control for lagged score.

APPENDIX TABLE 10
Impacts of Teacher Value-Added on Earnings by Age

Dependent Variable:	Earnings ($)								
	20	21	22	23	24	25	26	27	28
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Panel A: Full Sample								
Value-Added	-211	-322	-211	71	449	558	1,021	1,370	1,815
	(72)	(100)	(136)	(190)	(247)	(319)	(416)	(517)	(729)
Mean Earnings	4,872	6,378	8,398	11,402	13,919	16,071	17,914	19,322	20,353
	Panel B: Schools with Low College Attendance Rates								
Value-Added	171	165	416	731	1,053	1,405	1,637	1,728	2,073
	(87)	(119)	(159)	(215)	(277)	(343)	(440)	(546)	(785)
Mean Earnings	4,747	6,183	7,785	9,752	11,486	13,064	14,319	15,249	15,967
	Panel C: Schools with High College Attendance Rates								
Value-Added	-592	-791	-870	-730	-318	-464	448	1,200	2,209
	(110)	(157)	(217)	(317)	(417)	(554)	(717)	(911)	(1,274)
Mean Earnings	5,018	6,609	9,127	13,379	16,869	19,774	22,488	24,718	26,312

Notes: Each coefficient reports the effect of teacher VA on earnings from a separate OLS regression, with standard errors clustered by school-cohort in parentheses. All regressions use the specification and sample used to estimate Column 1 of Table 5, replacing the dependent variable with earnings at the age shown in the column heading. In Panels B and C, we split the sample into two based on the average college attendance rate at each school. The mean school-average college attendance rate is 35%. Panel B considers schools with attendance rates below 35% while Panel C considers schools with attendance rates above 35%. The second row in each panel reports mean earnings for the observations in the corresponding estimation sample.

APPENDIX TABLE 11
Impacts of Teacher Quality: Instrumental Variables Specifications

Dependent Variable:	Score	College at Age 20	College Quality at Age 20	Earnings at Age 28	College at Age 20	College Quality at Age 20	Earnings at Age 28
Estimation Method:		OLS (Reduced Form)			Two-Stage Least Squares		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	(SD)	(%)	($)	($)	(%)	($)	($)
Raw Teacher Quality	0.476	2.526	837	871			
	(0.006)	(0.349)	(93)	(392)			
Score					5.29	1,753	1,513
					(0.72)	(191)	(673)
Observations	3,721,120	3,095,822	3,095,822	368,427	3,089,442	3,089,442	368,427
Mean of Dep. Variable	0.162	37.8	24,815	20,912	37.8	24,815	20,912

Notes: This table reproduces the baseline specifications in Table 2 (Col. 1), Table 4 (Cols. 1 and 4), and Table 5 (Col 1) using raw estimates of teacher quality. Raw teacher quality is the estimate ν_j obtained after Step 2 of the procedure described in Section 2.2, prior to the Empirical Bayes shrinkage correction. We define teacher quality using student score residuals from classes taught by the same teacher in all other years available in the school district dataset. Student score residuals are calculated from an OLS regression of scores on the full student- and classroom-level control vector used to estimate the baseline value-added model, defined in the notes to Table 3. Columns 1 through 4 regress the outcome on raw teacher quality with the baseline classroom-level control vector used in Table 2. Columns 5-7 report 2SLS estimates, instrumenting for mean classroom test scores with raw teacher quality. All regressions cluster standard errors at the school x cohort level and are estimated on the linked analysis sample used to estimate the baseline specifications, with one observation per student-subject-school year. For comparability to baseline estimates, observations with teacher VA in the top 2% of the distribution (estimated using the baseline model in Table 3) are excluded.

APPENDIX TABLE 12
Impacts of Value-Added on College Quality by Grade

	College Quality at Age 20				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
Panel A: Reduced-Form Coefficients					
Teacher Value-Added	2,011	832	788	2,638	970
	(296)	(314)	(363)	(472)	(398)
Panel B: Coefficients Net of Teacher Tracking					
Teacher Value-Added	1,991	802	566	2,478	970

Notes: This table reports the coefficients plotted in Figure 10. Panel A replicates Column 5 of Table 4 for each grade separately, using only cohorts who would have been in 4th grade during or after 1994. Panel B calculates the impacts of teacher VA in each grade net of tracking to better teachers in future grades. We obtain these point estimates by estimating the impact of VA on future VA (see Appendix Table 13) and then subtracting out the indirect effects using the procedure described in section 6.1.

APPENDIX TABLE 13
Tracking Coefficients

	Future Teacher Quality			
	Grade 5	Grade 6	Grade 7	Grade 8
Grade 4 Teacher VA	0.001	0.012	0.005	-0.001
Grade 5 Teacher VA		0.038	0.002	0.004
Grade 6 Teacher VA			0.067	0.058
Grade 7 Teacher VA				0.165

Notes: Each cell reports the coefficient from a separate regression of teacher value-added in a subsequent grade on teacher value-added in the current grade. All regressions include the classroom-level baseline control vector used in Table 2 and are estimated on the linked analysis sample, using all observations for which the data needed to estimate the relevant regression are available.

APPENDIX TABLE 14
Lifetime Earnings Impacts of Deselecting Teachers Below 5th Percentile

Num. of Classes Observed	Present Value at Age 12 of Earnings Gain per Class	Undiscounted Sum of Earnings Gain per Class
1	$135,228	$748,248
2	$169,865	$939,899
3	$189,247	$1,047,145
4	$201,917	$1,117,250
5	$210,923	$1,167,085
6	$217,683	$1,204,486
7	$222,955	$1,233,659
8	$227,188	$1,257,083
9	$230,665	$1,276,321
10	$233,574	$1,292,415
Max	$266,664	$1,475,511

Notes: This table shows the earnings gains from replacing a teacher whose value-added is in the bottom 5% of the distribution with a median teacher for a single class of average size (28.3 children). Column 1 reports present values of earnings gains, discounted back to age 12 at a 5% rate. Column 2 reports undiscounted sums of total earnings gains. The row labeled Max shows the gains from deselecting teachers based on their on true VA. The other rows show the gains from deselection when VA is estimated based on a given number of classes. The calculations are based on the average lifecycle income profile of individuals in the U.S. population in 2007, adjusted for a 2% annual growth rate in earnings.